A COMPARISON OF THREE STATISTICAL TECHNIQUES WITH CUSTOMER SATISFACTION RATING SCALES

Julian Luke, Pedro Saavedra, Macro International Inc., Marcia S. Scott, Corporation for National Service
Julian Luke, Macro International Inc., 11785 Beltsville Drive, Calverton, MD 20705

Key Words: Gamma, Permutation Tests, Simulations, Non-parametric, Customer Satisfaction

Introduction

The Corporation for National Service (the Corporation) oversees national service activities in communities across the United States through AmeriCorps programs. This oversight occurs through the administration of grants to states through State Commissions for Community Service and directly to national nonprofit organizations (National Directs). The State Commissions and National Directs (grantees) in turn, issue grants to community-based organizations (subgrantees) to conduct programs in which AmeriCorps members provide services in the four areas of education, public safety, environment, and other unmet human needs.

The Customer Satisfaction Survey is conducted annually among AmeriCorps program grantees and subgrantees to determine their perspectives on the administrative functions of the Corporation. The purpose of the Customer Satisfaction Survey is to determine the efficacy of the Corporation's operations. An examination of the ratings of the Corporation's customers provides opportunities to develop improved program quality and service delivery for its customers.

Research Issues

Social scientists often have several possible statistical techniques available in the analysis of survey questionnaires. When conducting a study, one is often confronted with a need to decide among various tests of significance. But deciding on a particular technique presents the analyst with several considerations:

- How are responses distributed?
- What type of scale is being used to measure satisfaction?
- Is the mean the proper key statistic to be tested, or is there a better measure?
- Considering the sample design, which significance testing technique is more appropriate?

In an effort to address these underlying considerations to determine the best possible method of conducting significance testing, satisfaction ratings from two groups of respondents who were given several four point rating scales were compared using a parametric test, Goodman and Kruskal's Gamma, and permutation tests. The data used in this study were collected by Macro International Inc. for the AmeriCorps Customer Satisfaction Survey.

In an evaluation of Corporation grantees and subgrantees, 81 Likert scale variables were part of a questionnaire administered to 218 subgrantees. Respondents were asked to rate their satisfaction with the Corporations service in several key areas, for example:

- Did the Corporation provide a vision of National Service?
- Did the Corporation provide consistent information?
- Did the Corporation provide feedback on program performance?

One issue that the evaluation had to address was whether subgrantees of State Commissions and those of national non-profit organizations (labeled State and National Directs, respectively) had different opinions as reflected by the 81 variables. Given that subgrantees had been selected from a stratified sample and that responses to some of the Likert scales were somewhat skewed, the issue was raised as to whether t-tests were appropriate to make the comparison. Searching for an appropriate statistic to conduct the analysis, the authors decided upon Goodman and Kruskal's Gamma (1954). This statistic (in the present context) measures, for pairs of units with different ratings, the degree to which the one with the higher rating tends to fall more often in one group than in another. Thus in a 2 X k table comparing two groups, Gamma=(A-B)/(A+B) where:

- A is the numbers of pairs of individuals, one from each group, in which the member of the pair from the first group provided the higher response;

- B is the number of pairs in which the member of the pair from the second group provided the higher response.

This statistic ignores tied pairs, so it is possible to obtain a value of 1.00 with a far from perfect relationship (if there are many tied scores). However, one can easily calculate an asymptotic standard error for Gamma, and therefore use it as a significance test.

Survey Design

Macro International conducted the Customer Satisfaction Survey for the Corporation for National Service in 1998. Files of 1997-98 grantees and programs that were renewed for 1998-99 were obtained from the Corporation. The sampling frame consisted of all grantees and a sample of subgrantees.

The sample for the grantee survey consisted of a census of the 92 grantees in the original frame. After analyzing the file for duplications, we had 80 unique grantees. They were categorized for analytic purposes into State grantees (49) and National Directs (31). As a result of including all grantees in the sample, each grantee was assigned a weight of 1.0.

The original file of subgrantees contained 708 programs (some of which might have been the same respondent under different grant numbers). In sampling from this frame, we sought to include a program for each grantee. To this end, we selected "principal programs" among the subgrantees. They are defined as:

- The largest program a grantee had, or
- a program with at least 10 members.

This approach allowed us to sample the large programs and operating sites for each grantee, plus insured that *at least one* program from each grantee is represented.

The objectives of this sampling design were threefold. First, it should permit population estimates for percentages with a 95% confidence interval of plus or minus five percent. Second, it should provide sufficient power to conduct a test of significance comparing National Direct and State grantees, and be 90% certain of detecting a medium size effect (defined by Cohen as equal to one-half a standard deviation) with a two-tailed test at the .05 level. Third, it should permit the sampling of at least one program/operating site per grantee. This paper considers only the comparisons done for the subgrantee population.

For the precision desired, we drew a total sample size of 299 subgrantees. In initial data cleaning, we discovered 6 duplicates, leaving a final unduplicated sample of 293 records. A total of 216 subgrantees were interviewed during the period yielding a response rate of 73.7 percent. Additionally, there were 16 problem numbers through which contact could not be made and 4 refusals. A self-weighting subsample of subgrantees is actually used in this paper.

Study Methodology

Initially a t-test was conducted to test for significant differences between two sets of Likert scale meanss (for State Commission subgrantees and National Direct subgrantees). However, examinations of the distributions of scores (each variable had four possible answers that formed at least an ordinal scale) indicated that there were highly skewed distributions for many of these variables (i.e., many times in the 3-4 range). The t-test is known to be inappropriate in extreme cases where the distribution is highly skewed. As a result the Goodman and Kruskal Gamma non-parametric statistic was used in its place.

Gamma can be easily described. Consider all pairs of scores such that:

1) The two scores in a pair are not tied, and

2) Each answer was given by a different type of grantee (State Commission or National Direct).

Let P be the number of pairs where the State Commission grantee has the higher score and Q be the number of pairs where the National Direct grantee has the higher score. Then the test statistic Gamma=(P-Q)/(P+Q). Using a complex formula for the standard error of Gamma, one can calculate the significance of this statistic, and use it in place of a t-test or a correlation. If for a given variable Gamma divided by its asymptotic standard error is greater than 1.96, for instance, there is a significant difference at the .05 level on a two-tail test for that variable.

Comparing The Tests

The sampling methodology involved two strata and required that at least one subgrantee be sampled for each grantee in the program, but for the comparisons presented here it was reduced to a self-weighting sample of 216 subgrantees. The number of respondents for individual items, however, varied considerably from 216 to under a fourth of that figure.

Comparing the results from the two statistics (t and Gamma), and using a two-tailed test at the .05 level for each, each approach separately identified 11 variables as exhibiting significant differences, but only eight variables were identified as significant by both approaches. Thus each test identified three variables as being significantly different between the two groups which the other test failed to identify. This is not an unusual situation, but reporting the conflicting results is problematic.

At this point, the issue arose of whether one of the tests might be inappropriate because its assumptions were not met by the data, whether the tests measured different things, or whether the comparisons needed to be formulated in a different manner. This question suggests asking whether the reason for the differences was that the Gamma statistic ignored the Likert scales, or that the t-tests were affected by the distribution. Thus it would be useful to apply a test which in fact tested the difference between the means, but was not as dependent on assumptions of normality. It seemed clear that a permutation test for differences between the means would fulfill these two conditions.

The Permutation Tests

Ordinarily one carries out a permutation test using only respondents to the item in question. In this case we decided to simulate the classification, allowing for nonresponse to be one of the possible choices. A total of 7500 Monte Carlo simulations were carried out. In each simulation 158 subgrantees were classified as into group 1 and 58 as into group 2 (the same numbers as in the State Commission classification and the National Direct classification). The mean of the two randomly assigned groups was calculated for each of the two variables, and their differences calculated. Then the difference between the State Commission subgrantees and the National Direct subgrantees was calculated and compared with the 7500 simulated values. Let q be the number of values smaller than the real difference of the means. Then 1- 2*abs(.5-(q+.5)/7501) provides the two-tailed probability for the rejection of the null hypothesis that the two means are equal.

One would expect the permutation test to be closer to the t-test. Even though t makes no distributional assumptions, it is a measure of the difference between means, whereas the scale does not enter into the gamma statistic. We made the comparison with the equal variance t, the unequal variance t, and the optimal t (selecting the one appropriate to the F test of differences between variances). The mean probabilities were very similar, and the correlation across the 81 variables (which are, of course, not independent) was .99. The correlations for gamma were somewhat lower (.93), and so were the actual probabilities.

However, when looking at the results of the permutation test for purposes of rejecting the null hypotheses, the test proved more conservative at the .05 level than either of the other two. The permutation tests rejected only five of the 81 null hypotheses, and these were five of the eight variables found to be significant by both of the other tests (See Table 1).

We then tried the same procedure using a one-tailed test, or rather a two-tailed test at the .10 level. The results for the .10 level were that each test rejected 17 null hypotheses, with the t-test and the permutation test agreeing on 16 of them and the gamma agreeing on 15 with each of them (See Table 2.).

Table 1. Results of Two-Tailed Test at the .05 Level of Significance

T-Test	Gamma	Permutation	Frequency
p > .05	p > .05	p > .05	67
p > .05	p < .05	p > .05	3
p < .05	p > .05	p > .05	3
p < .05	p < .05	p > .05	3
p < .05	p < .05	p < .05	5

Table 2. Results of Two-Tailed Test at the .10 Level of Significance

T-Test	Gamma	Permutation	Frequency
p > .10	p > .10	p > .10	62
p > .10	p < .10	p > .10	1
p > .10	p < .10	p < .10	1
p < .10	p > .10	p < .10	2
p < .10	p < .10	p > .10	1
p < .10	p < .10	p < .10	14

Thus, in these particular examples, the permutation test is slightly more conservative (or less sensitive) at a lower level of significance, but seems to yield similar results at higher levels.

Two Examples to Think About

Analysis of table 3 shows that a t-test reveals unequal variances, and yields a significant difference between the means at $p<.01$. One suspects that a permutation test would yield the same results. However, a quick computation of Gamma reveals that it is equal to 0. The means are very different, but if one selects a pair from each group, the probability that the higher one is from Group 1 is the same as the probability that it is from Group 2.

Table 3. Explanatory Example Distribution 1

Group	Scale Value			
	1	2	3	4
1	60	0	0	60
2	0	5	90	0

In table 4 it is Gamma that is significant at the .01 level, but t is exactly zero. The mean for both groups is 3.5, but if one selects a random pair, one from each group, where respondents did not give the same response, one finds 2,400 such pairs where the second group is higher, and 6,000 where the first group is higher, so that Gamma=-3,600/8,400 =-.429 with an asymptotic standard error of 0.107.

Table 4. Explanatory Example Distribution 2

Group	Scale Value			
	1	2	3	4
1	20	0	0	100
2	0	0	60	60

Conclusions

1) The three tests tend to give similar, but not identical results, with the permutation test being the most conservative.

2) The t-test and gamma are measuring different hypotheses. These different hypotheses tend to hold for the same variables, but this need not be the case.

3) The t-test may yield inaccurate assessments of what it tries to measure, not because it is necessarily inappropriate to treat the data as an interval scale (which the permutation test does) but because of the distributions of the variables.

4) Conceptual hypotheses may be so ambiguous that significant results may be largely dependent on how one translates them into operational tests.

The question is: When we say that one group has a higher score than another, which group do we mean? Or do we know beforehand?

References

Good, P. 1993, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* New York: Springer Verlag

Goodman, L.A. and Kruskal, W.H. , 1954, "Measures of Association for Cross-Classifications" *Journal of the American Statistical Association* 49:732-64.

Some of the research referred to in this paper was conducted by Helene Jennings. The authors wish to thank Ellen Marks and Hoke Wilson for acting as peer reviewers.