USE OF MULTIPLE COMPARISON PROCEDURES IN THE DESIGN AND ANALYSIS OF YEAR 2000 CENSUS RESEARCH STUDIES*

Henry F. Woltman, Bureau of the Census
Washington, D. C. 20233

KEY WORDS: Census Mail Response, Pairwise Comparisons, Treatments vs. a Control.

I. Background

The rate at which questionnaires were returned in the 1990 Census via the mailback methodology was about 65 percent. This represents about a ten percentage point decline as compared to the 1980 Census. In addition, 1990 Census evaluation studies indicate that the quality of data, particularly in terms of coverage, is somewhat better for mail return questionnaires versus those questionnaires not returned by mail and subsequently completed by enumerators during a follow-up operation (Griffin and Moriarity 1992).

Based on these factors, a major part of the overall research and design program for the Year 2000 Census involves a series of relatively small scale national experiments designed to investigate methods for increasing the mail response rate. The general nature of these experiments involves pairwise comparisons of mail response rates for experimental questionnaires/methods with the 1990 Census short form questionnaire and comparisons among the experimental questionnaires/methods.[1]

Thus, the analysis of the experiments can be characterized as involving all possible pairwise comparisons between a number of treatments.

This paper discusses the application of multiple comparison procedures (MCP's) to the first response test, the simplified questionnaire test (SQT). In addition to illustrating the use of MCP's for the "all pairwise comparison" analysis, use of another analytical MCP method is illustrated. This is the method of comparison with a control. All methods are discussed in Hochberg and Tamhane (1987). Finally, determining sample size for experiments involving the use of a pairwise MCP and experimental designs like the SQT is discussed.

II. Multiple Comparison Procedures

A. Background

Hochberg and Tamhane (1987) provide the basis for the basic notions and philosophy of multiple comparison procedures. As they note, in comparative studies such as those discussed in this paper, assessing each comparison separately (a per-comparison approach) by a suitable procedure, e.g., a t-test, does not account for the multiplicity or selection effect (Tukey 1977). This has been put as follows: "If enough statistics are computed, some of them will be sure to show structure" (Diaconis 1985). As a result, use of multiple t-tests will result in spurious overall and pairwise significant results more frequently than indicated by the per comparison alpha level.

Statistical procedures that are designed to take into account and properly control for the multiplicity effect through some combined or joint measure of erroneous inferences are called multiple comparison procedures (MCP's).

B. Families

Any collection of inferences for which it is meaningful to take into account some combined measure of error is called a family.

For the purposes of experiments cited in this paper, the family is defined as the collection of all possible pairwise comparisons. Hochberg and Tamhane (1987) indicate that such a family may be considered as finite in nature.

C. Error Rates

Hochberg and Tamhane define three possible error rates. Their formal definition of these follows.

- Let
  F denote a family of inferences
  P denote an MCP for this family; we assume that every inference P makes (or can potentially make) is either right or wrong
  M (F,P) be the random number of wrong inferences
- Familywise Error Rate (FWE)
  (also called experimentalwise error rate)
  FWE (F, P) = Pr { M (F,P) >0 }
- Per-Family Error Rate (PFE)
  (also called per-experiment error rate)
  PFE (F,P) = E { M (F,P) }, where E denotes expectation
- Per Comparison Error Rate (PCE)
  $PCE(F,P) = \frac{E(M(F,P))}{N(F)}$, where N(F) denotes the number of inferences in the family

For the analysis of experiment cited in this paper, it was decided to control the familywise error

rate (FWE) since it was deemed essential for decision purposes that all inferences be simultaneously correct. Analysis of the test results for other response rate experiments have also controlled this error rate.

III. Simplified Questionnaire Test

A. TEST DESIGN

The Simplified Questionnaire Test (SQT) was developed to evaluate alternative short form questionnaire designs and methods for increasing the number of contacts with households in an effort to improve response rates in future censuses (Sinclair and West, 1992).

Specifically, the SQT was designed to determine the effects on response rates from:

- ASKING FEWER QUESTIONS -The objective is to determine whether asking fewer questions than the 1990 short form can increase response rates.
- USING RESPONDENT FRIENDLY FORMS - Comments about the 1990 questionnaire suggest that some of the nonresponse may have resulted from difficulty following the instructions on the form. It might also be speculated that the design and appearance of the questionnaire package itself have contributed to the nonresponse. Thus, one of the secondary goals of the SQT is to determine whether a form that is less complex and easier to use can increase response rates.
- REQUESTING SOCIAL SECURITY NUMBERS - If fewer questions are asked in the future on the census short form, supplementing the reduced data with data from administrative records might be considered. One way to link census data for an individual to information from these records would involve asking for social security number (SSN) on the census questionnaire. While there clearly are privacy, coverage and other concerns about large-scale use of administrative record systems, there also may be significant opportunities to reduce public burden, costs, and staffing requirements for the census. Thus, the SQT is also designed to measure the effect on the mail response rate of requesting SSN on the form.
- USING AN IMPLEMENTATION STRATEGY THAT RELIES UPON MULTIPLE CONTACTS BY MAIL - Survey research literature clearly demonstrates that higher response rates can be obtained through multiple contacts such as pre-notification, reminder cards, and replacement forms.

Such features are incorporated in the implementation of the SQT. Since the census implemented an early alert card for only a small portion of the country and a reminder card nationwide, there is interest to observe effects of a pre-notice letter and a replacement questionnaire in the SQT.

The study was conducted using the 1990 Census short form as the control and four alternative versions of the 1990 Census short form questionnaire. The forms are denoted as: 1) 1990 Short Form, 2) Booklet, 3) Micro, 4) Micro/SSN and 5) Roster Form.

The Booklet form contained the same items as the 1990 Census short form, but in a "user friendly" format. The Micro form contained the population data items from the 1990 Census short form, but no housing items. The form was in a one page front/back format. The SSN was added to this form for the Micro/SSN form. The Roster form was limited to obtaining the household roster and each person's date of birth in a half page card stock format.

The SQT evaluation had three components: a mail response analysis, a respondent/ nonrespondent telephone debriefing, and an item nonresponse evaluation. This paper focuses on the mail response analysis component only.

A stratified nationally representative sampling design was implemented for this study consisting of a total sample of 17,000 housing units. The first stratum was defined to encompass District offices (DO's) having low mail response rates in the 1990 Census (1990 LRA).[2] The second stratum included the remaining Do's denoted as High Response Areas (HRA). The LRA stratum contained 10,354,310 housing units in 1990; the HRA stratum contained 78,529,760 housing units in 1990. Each of the four objectives of this test was evaluated at the national and the stratum level. Differences between the two strata also were examined.

The sample was allocated equally to the five questionnaire variants by stratum. Thus, each questionnaire/stratum variant had a sample size of 1700 housing units. In an effort to increase response rates in a non-census environment, each household received a pre-notice letter and a reminder postcard. Initial mailout of the questionnaire was conducted on March 23, 1992. All nonrespondents and Post Master Returns (PMRs) as of April 17, 1992 received a second questionnaire.[3]

B. ANALYSIS USING MULTIPLE COMPARISON PROCEDURES

1. All Pairwise Comparisons

The experimental design of the SQT involved random assignment of the five questionnaires to clusters of five adjacent housing units. The purpose

of this was to reduce the sampling error of the pairwise comparisons by virtue of the correlation that is likely to exist among similar households with respect to their propensity to complete and mail back a questionnaire. In this case, the overall design can be characterized as a balanced equicorrelated design. Hochberg and Tamhane (1987) refer to this type of design as pairwise balanced. These designs are defined by Hochberg (1974) as designs for which all pairwise differences, $(\hat{\theta}_i - \hat{\theta}_j)$, have the same variance. That is, Var $(\hat{\theta}_i - \hat{\theta}_j) = \sigma^2(V_{ii}+V_{ij}-2V_{jj}) = 2\sigma^2 V$.

$$\text{therefore } [Var\hat{\theta}_i - \hat{\theta}_j]^{\frac{1}{2}}/\sqrt{2} = \sigma\sqrt{V}$$

In this case the exact $(1-\alpha)$-level simultaneous confidence intervals for all contrasts is given by

$$\sum_{i=1}^{k} c_i\theta_i \in \left[\sum_{i=1}^{k} c_i\hat{\theta}_i \pm Q(k,\nu,\alpha)\, s\sqrt{V} \sum_{i=1}^{k} \frac{|c_i|}{2} \right]$$ where $Q(k,\nu,\alpha)$ is the upper α point of the studentized range distribution for k treatments, ν degrees of freedom.

For pairwise comparisons, say for treatments i and j, $C_i=1$, $C_j=-1$, other $C^s=0$ and

$$s\sqrt{V} = [V\hat{A}R(\hat{\theta}_i-\hat{\theta}_j)]^{\frac{1}{2}}/\sqrt{2}$$

$$(\theta_i-\theta_j) \in \left[\hat{\theta}_i-\hat{\theta}_j \pm Q(k,\nu,\alpha)\ [(V\hat{A}R\ (\hat{\theta}_i-\hat{\theta}_j)]^{\frac{1}{2}}/\sqrt{2} \right]$$

Table 1 provides the mail response rates by stratum (LRA and HRA) by questionnaire. Also included is the standard error of the estimates calculated using a jackknife variance estimation procedure. Table 2 provides all pairwise comparisons by stratum along with the estimated sampling error (also calculated by the jackknife procedure) on each difference. Based on the results in Tables 1 and 2 we can infer that the design is essentially pairwise balanced within stratum. Simple calculations indicate the correlation coefficient for any two treatments (ρ_{ij}) is about 0.12.

Table 3 provides the pairwise comparisons of mail response rates by stratum and combined over stratum for four experimental objectives: "comparison to the 1990 census short form", "asking fewer questions", "user friendliness" and "asking social security number" (SSN). The 90% simultaneous confidence intervals on the differences are also shown using $Q(5,\infty,.10) = 3.48$. The national estimates are a weighted average of the stratum estimates using (.1164) for the LRA stratum and (.8836) for the HRA stratum. These weights are based on the 1990 census housing units counts cited in section III. A.

Thus, with 90% confidence in each category (National, LRA, HRA) separately the intervals all contain their respective true difference (i.e., simultaneously). The underlined intervals indicate significant differences. Sinclair and West (1992) provide detailed analysis of these and other results.

2. Comparison to the Control

Had the SQT objective been simply to compare the mail response rates of the experimental forms with the 1990 census short form then the analysis would use a "comparison with control" procedure. To illustrate the results of this analysis I use the exact procedure for balanced designs described by Hochberg and Tamhane (1987).

Specifically, the $(1-\alpha)$ level simultaneous one-sided C.I.'s for $(\theta_i - \theta_k)$; k denotes the control, are

$$(\theta_i-\theta_k) \geq (\hat{\theta}_i-\hat{\theta}_k - T(\alpha,k-1,\nu,\rho_{ij})\ s\sqrt{V_{ii}+V_{jj}-2V_{ij}})$$

where ρ_{ij} = CORR $[(\hat{\theta}_i-\hat{\theta}_k),\ (\hat{\theta}_j-\hat{\theta}_k)]$

Where $T(.,.,.,.,.)$ is the upper α point of the k- variate t distribution with ν degrees of freedom and associated correlation matrix $R = (\rho_{ij})$.

If $\rho_{ij} = \rho$ for all $i \neq j$, the design is balanced with respect to treatments. In general, for $n_i = n$ for all i and $\rho(\hat{\theta}_i, \hat{\theta}_j)$ constant for all $i \neq j$, $\rho = \rho_{ij} = 1/2$. These conditions apply for the SQT.

The results of this analysis are shown in Table 4 for the combined stratum (i.e., national) level mail response rates.

Here

- $T(.10,4,\infty,1/2) = 1.84$
- $\hat{\sigma}(\hat{\theta}_i-\hat{\theta}_j) \doteq 1.4$ for all $i \neq j$ (Table 2)
- all treatments equicorrelated
  Corr $(\hat{\theta}_i, \hat{\theta}_j) = 0.12$ for all $i \neq j$

The conclusion based on this analysis is that all treatments are "better" than the control; with 90% confidence, the following are simultaneously correct:

1. The mail response rate for the booklet form exceeds that for 90SF by at least 0.8 percentage points.

2. The mail response rate for the micro/SSN form exceeds that for the 90SF by at least 2 percentage points,

.
.
.
etc.

IV. Designing Experiments for Multiple Comparisons

The approach first suggested by Tukey (1953) provides a relatively simple method to determine

the sample size in an experiment where all pairwise comparisons are to be estimated. The philosophy of this procedure is "We seek for a chosen degree of confidence (that is, a chosen probability) that the allowance (comparative) that we seek will be no larger than a chosen bound". That is, we desire for all $i \neq j, (\theta_i - \theta_j) \in [\hat{\theta}_i - \hat{\theta}_j \pm A]$ with confidence $(1-\alpha)$ simultaneously.

Hochberg and Tamhane (1987) show that, assuming independence for all pairwise comparisons, if one specifies a fixed common allowance $A > 0$, and a familywise desired confidence coefficient of $(1-\alpha)$, the sample size n is given by $n = \left(\frac{Q(k, \infty, \alpha)\sigma}{A}\right)^2$ where it is assumed the common error variance σ^2 for each treatment is known. If σ^2 is unknown, then they suggest using a conservative upper bound for design purposes. For response rate experiments this means $\sigma = 0.5$. Note, for pairwise balanced designs, $\sigma(1-\rho)^{1/2}$ can be used in place of σ, if an approximate value of ρ is known.

This approach provides a simple method to determine the sample size requirements. For the SQT, the allowance specified was about .05, and the confidence coefficient was 90%. Thus, the true difference between mail response rates for all pairwise comparisons should be contained within the estimated difference plus or minus five percentage points. Using $Q(5, \infty, 0.10) = 3.48$ *and* $\sigma = 0.5$ *and* $A = .05$, the sample size (HU's) is 1211. The sampling frame for the SQT was the 1990 census address file and as a result the sample size was adjusted upward to a conservative value of 1700 to account for the fact that some proportion (20-30%) of the mailout sample was expected to be vacant housing units and addresses that were otherwise undeliverable by the USPS (e.g. nonexistent).

The realized allowances for the SQT were about ± 4.0 percentage points in the LRA stratum and ± 3.8 percentage points in the HRA stratum. This was the case since only about 10% of the mailout sample cases were returned by USPS as PMRs.

The sample size requirements for several other response rate experiments conducted after the SQT also used this approach.

Reading (1975) provides another method for classifying all pairs of means out of k as close or distant. In this method one specifies the parameter δ_1, the largest amount by which the means of two populations can differ and still be considered practically the same; and δ_2, the smallest amount by which the two means have to differ to be considered definitely different. This type of procedure uses the indifference zone (Gibbons, Olkin, and Sobel, 1977) formulation along the lines of a ranking and selection procedure. We have not used this procedure for designing the response rate experiments, but it is mentioned as it may have potential for use in the design of other experiments conducted as part of the Year 2000 research and development program.

References

1. Griffin, Deborah and Moriarity, Christopher, (1992), "Characteristics of Census Errors", American Statistical Association, Proceedings of the Survey Research Sections, Boston, Mass.
2. Hochberg, Yosef and Tamhane, Ajit, (1987), "Multiple Comparison Procedures", John Wiley & Sons.
3. Tukey, J. W., (1977), "Some Thoughts on Clinical Trials, Especially Problems of Multiplicity," Science, 198, 679-684.
4. Diaconis, P., (1985), "Theories of Data Analysis From Magical Thinking Through Classical Statistics," in Exploring Data Tables, Trends and Shapes, New York: Wiley, 1-36.
5. West, Kirsten and Sinclair, Michael, (1992), "Simplified Questionnaire Test (SQT) Mail Response Evaluations," internal Census Bureau research report.
6. Gibbons, J., Olkin, I. and Sobel, M., (1977) "Selecting and Ordering Populations."
7. Reading, J. C., (1975), "A Multiple Comparison Procedure for Classifying All Pairs of Means as Close or Distant," J. Amer. Stat. Assoc., 70, 832-838.
8. Hochberg, Yosef, (1974), "The Distribution of the Range in General Unbalanced Models, Amer. Stat., 28, 137-138.
9. Tukey, J. W., (1973), The Problem of Multiple Comparisons", mimeographed document.

[1] The "short form" questionnaire includes only the data items asked of all persons and housing units enumerated in the Census (i.e., the complete count data).

[2] There were 449 District offices created for the 1990 Census. They averaged about 250,000 units.

[3] A PMR occurs when the USPS cannot deliver a questionnaire to an address. Reasons include the unit is vacant, no longer exists, or the mailing piece is otherwise undeliverable. The USPS returned these mailing pieces to the Census Bureau.

* This paper reports the results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

Table 1. Mail Response Rates by Form Type Stratum

Form Type	Mail Response Rates	Standard Error
	STR=LRA	
1. 1990 Census SF	45.1	1.24
2. Booklet SF	52.7	1.24
3. Micro SF	55.1	1.24
4. Micro W/SSN	48.8	1.25
5. Roster	54.6	1.24
	STR=HRA	
1. 1990 Census SF	65.8	1.18
2. Booklet SF	68.7	1.16
3. Micro SF	73.5	1.10
4. Micro W/SSN	70.5	1.14
5. Roster	73.1	1.10
	NATIONAL	
1. 1990 Census SF	63.4	1.05
2. Booklet SF	66.8	1.03
3. Micro SF	71.4	0.98
4. Micro W/SSN	68.0	1.01
5. Roster	70.8	0.99

Table 2. Mail Response Rate, Pairwise Comparisons by Form Type

Comparison of Form Types Comparison	Difference in Response Rates	Standard Error of Difference
	STR=LRA	
2 - 1	7.6	1.67
3 - 1	10.0	1.61
4 - 1	3.7	1.69
5 - 1	9.5	1.63
3 - 2	2.4	1.65
4 - 2	- 3.9	1.64
5 - 2	1.9	1.66
4 - 3	- 6.2	1.60
5 - 3	- 0.5	1.65
5 - 4	5.8	1.68
	STR=HRA	
2 - 1	2.9	1.54
3 - 1	7.7	1.55
4 - 1	4.7	1.55
5 - 1	7.3	1.54
3 - 2	4.8	1.52
4 - 2	1.8	1.57
5 - 2	4.4	1.51
4 - 3	- 2.9	1.52
5 - 3	- 0.4	1.48
5 - 4	2.6	1.49
	National	
2 - 1	3.4	1.38
3 - 1	8.0	1.38
4 - 1	4.6	1.38
5 - 1	7.5	1.38
3 - 2	4.6	1.35
4 - 2	1.2	1.40
5 - 2	4.1	1.36
4 - 3	- 3.4	1.32
5 - 3	- 0.4	1.33
5 - 4	2.9	1.33

Table 3. Comparison of Mail Response Rates by SQT Form Type

Experimental Comparisons	Response Rates (%) and 90% Confidence Intervals (C.I.)					
	National		1990 Low Response Areas (LRA)		High Response Areas (HRA)	
	Difference	90% C.I.	Difference	90% C.I.	Difference	90% C.I.
Comparison to the 1990 Short Form						
Booklet - 1990 SF	3.4	0.0 to 6.8	7.6	3.5 to 11.7	2.9	-0.9 to 6.7
Micro - 1990 SF	8.0	4.6 to 11.4	10.0	6.0 to 13.9	7.7	3.9 to 11.5
Mfcro/SSN - 1990 SF	4.6	1.2 to 8.0	3.7	-0.4 to 7.8	4.7	0.9 to 8.5
Roster - 1990 SF	7.5	4.2 to 10.9	9.5	5.5 to 13.5	7.3	3.5 to 11.1
Number of Questions Asked						
Micro - Booklet	4.6	1.2 to 7.8	2.4	-1.7 to 6.4	4.8	1.1 to 8.5
Roster - Booklet	4.1	0.8 to 7.4	1.9	-2.2 to 6.0	4.4	0.7 to 8.1
Roster - Micro	-0.4	-3.7 to 2.8	-0.5	-4.5 to 3.6	-0.4	-4.0 to 3.2
User Friendliness						
Booklet - 1990 SF	3.4	0.0 to 6.8	7.6	3.5 to 11.7	2.9	-0.9 to 6.7
Asking SSN						
Micro-Micro/SSN	-3.4	-6.7 to-0.1	-6.2	-10.3 to-2.3	-3.0	-6.7 to 0.8

Table 4. SQT Results: Comparison With a "Control"

Form Type	National Mail Response Rate	Difference (treatment minus control)	One-Sided Lower Simultaneous 90% C.I.'s
1. 1990 Census SF (control)	63.4	NA	NA
2. Booklet	66.8	3.4	0.8
3. Micro/SSN	68.0	4.6	2.0
4. Roster	70.0	7.4	4.9
5. Micro	71.4	8.0	5.4