Survey Methodology, June 1990
Vol. 16, No. 1, pp. 145–151
Statistics Canada

An Example of the Use of Randomization Tests in Testing the Census Questionnaire

YVES BÉLAND and ALAIN THÉBERGE

ABSTRACT

Modular Test 2 was a survey conducted by Statistics Canada that used two different questionnaires. Its purpose was to assist in the making of the 1991 census questionnaire. The sample used for the survey was not a probability sample. This article briefly describes the survey methodology, and the use of randomization tests to compare the two questionnaires.

KEY WORDS: Randomization tests; Non-probability sample; Experimental design.

1. INTRODUCTION

Statistical tests could be classified into two groups, randomization tests and classical tests. A classical test, is based on a comparison of the observed value of a statistic with the distribution, under the null hypothesis, of the values of this statistic for the set of samples that could have been selected. To conduct this kind of test, the probability of selecting any given sample must be known; therefore probability sampling using a known design is required. A randomization test is based on a comparison of the observed value of a statistic with the distribution, under the null hypothesis, of the values of this statistic for all possible permutations of the data. This was the method used by Fisher to compare two seed samples (1935), and Edgington (1987) also discusses various aspects of this method. "Treatments" are required to define the permutations in a randomization test, and the probability of obtaining a given permutation must also be known. Which unit will be given which treatment must be decided randomly; that is, the experimental design must incorporate randomization.

In an organization like Statistics Canada, classical tests are generally used because most of the sample surveys done by Statistics Canada use probability sampling, and also because there are no treatments in these surveys. This article describes how randomization tests were used in a survey that was an exception to the rule.

In Section 2, the methodology used in the modular tests is described briefly. Section 3 describes using simple examples the procedure used in a randomization test. Section 4 describes how randomization tests were applied to Modular Test 2.

2. MODULAR TESTS

As part of the planning for the 1991 census, two modular tests were carried out to test questions likely to be asked in the census. The purpose of these surveys was to ensure that each question whether new or just reformulated was easy to understand. We refer to the tests as "modular" because they were independent surveys that tested different sections of the census questionnaire.

[1] Yves Béland, Social Survey Methods Division; Alain Théberge, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6

Modular Test 1 was carried out in November, 1987 in order to revise newly-formulated questions dealing with population coverage, marital status, fertility, volunteer work, and nuptiality. This first survey used neither classical nor randomization tests.

Modular Test 2, carried out in January, 1988, was designed principally to measure the reaction of ethnic groups to questions on language, ethnic origin, religion, citizenship, and mobility. In Modular Test 2, a two-stage sampling plan was used to select about 3,500 households taken from within the metropolitan areas of Halifax, Québec, Montréal, Toronto, Winnipeg, and Vancouver. To reduce costs and to make data collection easier, and to get a sample that contained people of diverse ethnic origins, a non-probability method was used to select the sample. The questionnaire used in Modular Test 2 came in two versions. The differences are described in Section 4. The households in the sample were given either version 1 or version 2 on a random basis.

Randomization tests were used to allow us to statistically test hypotheses pertaining to Modular Test 2. Randomizations tests can be used to compare two treatments applied to units in samples which may not be probability samples.

3. RANDOMIZATION TESTS

The procedure for doing a randomization test will now be described. First, the value of a statistic is calculated for the observed data. Next, the value of the same statistic is calculated for the other permutations of the data that are possible with the experimental design used. H_0 is rejected if the value of the statistic for the observed data is extreme in relation to the values obtained under H_0 for the set of permutations.

For example, suppose there are four households. Household 1 has three persons, households 2 and 3 have two, and household 4 has one. These households may have been chosen arbitrarily, but a household whose members will receive treatment Y is chosen at random. Members of the three other households will receive treatment X. Suppose that household 4 is selected for treatment Y. For household 1, the treatment succeeds for two of the three members, for households 2 and 3, for one of two members, and for household 4, it fails for the sole member. Our null hypothesis states that the results are independent of the treatment used. To measure the impact of treatment X compared to treatment Y, the statistic S, giving the average number of successes for treatment X minus the average number of successes for treatment Y is calculated. Here $S = (2 + 1 + 1)/(3 + 2 + 2) - 0/1 = 4/7$. To find out whether this value is significant, the values for S obtained by permuting the observations are given in Table 1. Each observation in Table 1 shows the number of members in the household after the vertical bar, and the number of successes before the vertical bar. If a right-tailed test is used, H_0 is rejected when $\alpha \geq 3/12 = .25$, because three of the twelve permutations yield an S value greater than or equal to 4/7, the observed value.

Rather than permuting the observations, we could have permuted the treatments. Table 2 gives the results when this is done. Because only one of the four permutations yields a value for S greater than or equal to 4/7 for a right- tailed test, we again reject H_0 *if* $\alpha \geq 1/4 = .25$. It is not a coincidence if the results are the same. Note n_{ki}, the number of units that receive treatment k ($k = 1, \ldots, K$) and for which the result r_i ($i = 1, \ldots, I$) is observed; $n_{k.} = \sum_i n_{ki}$ the number of units that receive treatment k, $n_{.i} = \sum_k n_{ki}$, the number of units for which the result r_i is observed; and $n_{..} = \sum_k \sum_i n_{ki}$, the total number of units. The number, N_t, of permutations of the treatments is given by

$$N_t = n_{..}! \Big/ \prod_k (n_{k.}!). \tag{1}$$

Table 1
Values of the Statistics S for each Permutation of the Observations

Treatment	Permutations										
X	2 | 3	1 | 2	1 | 2	2 | 3	2 | 3	1 | 2	1 | 2	0 | 1	0 | 1	1 | 2	1 | 2	0 | 1
X	1 | 2	2 | 3	1 | 2	1 | 2	0 | 1	2 | 3	0 | 1	1 | 2	2 | 3	1 | 2	0 | 1	1 | 2
X	1 | 2	1 | 2	2 | 3	0 | 1	1 | 2	0 | 1	2 | 3	2 | 3	1 | 2	0 | 1	1 | 2	1 | 2
Y	0 | 1	0 | 1	0 | 1	1 | 2	1 | 2	1 | 2	1 | 2	1 | 2	1 | 2	2 | 3	2 | 3	2 | 3
S	4/7	4/7	4/7	0	0	0	0	0	0	−4/15	−4/15	−4/15

Table 2
Values of the Statistics S for each Permutation of the Treatments

Observation	Permutations			
2 | 3	X	X	X	Y
1 | 2	X	X	Y	X
1 | 2	X	Y	X	X
0 | 1	Y	X	X	X
S	4/7	0	0	−4/15

Of these N_t permutations, there are N_t^* for which n_{ki} units are associated with treatment k and the result r_i ($k = 1, 2, \ldots, K$; $i = 1, 2, \ldots, I$), where

$$N_t^* = \prod_i \left(n_{.i}! \Big/ \prod_k (n_{ki}!) \right). \tag{2}$$

In addition, there are N_o permutations of the observations where

$$N_o = n_{..}! \Big/ \prod_i (n_{.i}!). \tag{3}$$

Of these N_o permutations, there are N_o^* for which n_{ki} units are associated with treatment k and the result r_i ($k = 1, 2, \ldots, K$; $i = 1, 2, \ldots, I$), where

$$N_o^* = \prod_k \left(n_{k.}! \Big/ \prod_i (n_{ki}!) \right). \tag{4}$$

Because $N_o^*/N_o = N_t^*/N_t$, the tests are equivalent. To reduce the number of calculations, it is preferable to permute the treatments if $N_t < N_o$, and to permute the observations if $N_t > N_o$. Dwass (1957) suggests that when there are a large number of permutations, a sample of permutations can be taken, and the observed value of the statistic can be compared to the set of values for the sample. If all of the permutations are not considered, the level of the test is not affected, only its power is.

If the permutations are sampled, the rule given above can still be applied, not to reduce the number of calculations, but to minimize the loss of power due to sampling. For example, Dwass shows that for a one-tailed test at the 0.05 level, the loss of power for a sample of 999 permutations is no more than 5.5%. Bradley (1968) notes that when the power of randomization and classical tests are compared, the results depend on to what extent the requirements of the classical tests have been met.

Because of the way in which randomization tests are constructed, the inference applies only to the effect of treatment on units in the sample, and not to the entire population. Classical tests, however, are based on a random sample drawn from a population that rarely matches the population of interest. In the present case for example, the population of interest is the Canadian population on Census Day, June 4, 1991. So for both types of tests, non-statistical arguments must be used to generalize inferences to the population of interest.

4. THE USE OF RANDOMIZATION TESTS IN MODULAR TEST 2

As mentioned above, there are two questionnaire versions for Modular Test 2, versions X and Y. Questions on ethnic identity and ethnic origin differ in the two versions. "CANADIAN" is a response category in version X that the respondent can select to answer the questions on ethnic identity and origin. In version Y, those who want to respond "CANADIAN" must write it out in full after selecting the category, "OTHER."

We wanted to know whether questions on ethnic identity and origin in version X of the test questionnaire got more or got less multiple responses than these questions in version Y. By a multiple response we mean any response in which more than one category has been chosen. We also wanted to find out what bearing the type of questionnaire had on multiplicity (number of response categories selected by the respondent), and on the selection of certain response categories (such as "FRENCH") for these questions. The types of questionnaire constitute the treatments. Because the sample for each region had its peculiarities, the randomization tests were done separately for each of the metropolitan areas from which the sample was taken.

First of all, we generated at random a sample of 999 permutations of the questionnaire versions. A permutation is generated as follows: For any given region, let N_x and N_y represent the number of X and Y questionnaires respectively. Using Bebbington's algorithm (1975), from the $N_x + N_y$ households take a simple random sample of N_x households. Household members in this sample are then assigned version X of the questionnaire. This process is repeated 999 times. Next, calculate for a given question the proportion of respondents who gave a multiple response for version X and for version Y. These proportions are denoted P_x and P_y.

Next, for each of the 999 permutations of the questionnaire versions, as well as for the initial observed sample, we calculated the statistic $S = P_x - P_y$. In this way we obtained 1,000 values for S, which we ranked in increasing order. If more than one statistic had the same value, we generated a random number between 0 and 1 and used it to determine the order of statistics of the same value. We used the variable RANKP_{x-y} to represent the rank of an observed S statistic.

Let μ_x and μ_y represent the expected proportion of respondents who gave a multiple response for version X and version Y respectively. For all regions excluding Halifax we tested:

$$H_0: \mu_x = \mu_y$$

versus

$$H_1: \mu_x > \mu_y.$$

For Halifax, the counter-hypothesis H_1: $\mu_x < \mu_y$ was used because more multiple responses were expected for version Y of the questionnaire. Because "CANADIAN" was not an available response category on version Y of the questionnaire and because the majority of households selected in this region were made up of people of British origin (that is, English, Scottish, or Irish), members of households that received version Y marked one or more of these categories. Members of households that received version X had the option of marking only the "CANADIAN" category.

The critical level, $\hat{\alpha}$, is calculated as follows: for the Halifax region, given that H_0 is rejected if the proportion of respondents who gave a multiple response in version X is significantly lower than the proportion observed for Y, the critical level is $\text{RANKP}_{x-y}/1000$. For all the other regions, given that H_0 is rejected if the proportion for X is significantly higher than the proportion observed for Y, the critical level is $(1001 - \text{RANKP}_{x-y})/1000$. The results are shown in Table 3.

Randomization tests were also used to test multiplicity (the number of response categories selected by the respondent) for questions on ethnic identity and origin in each of the regions, but this time ratios (R_x, R_y) are used, instead of proportions (P_x, P_y). Ratio R_x is the average number of response categories selected by respondent for a question in version X of the questionnaire, and ratio R_y is the average number of response categories selected in version Y. The rest of the method is the same except that instead of RANKP_{x-y}, RANKR_{x-y} is used, and the statistic S is defined as $R_x - R_y$. However, because there is greater variability for the values of the statistic S in the tests for multiplicity, a sample of 1,999 permutations was generated instead of 999.

Let F and G represent the distribution functions of the number of response categories selected in version X and version Y respectively. For all the regions excluding Halifax, we test the hypothesis

$$H_0: F = G$$

versus

$$H_1: F(z) \leq G(z) \text{ for all } z \text{ and } F \neq G.$$

If H_0 is rejected, the number of response categories selected for an X questionnaire is said to be stochastically larger than the number of response categories selected for a Y questionnaire. For Halifax, the counter-hypothesis used is H_1: $F(z) \geq G(z)$, for all z and $F \neq G$. The results are shown in Table 3. In the Québec region, the value of R_y is less than 1 for each question. This is because most respondents in this region chose only one response category, and some respondents did not answer one or other of the questions.

Finally, versions X and Y for Modular Test 2 were compared for some regions as to the number of respondents who identified themselves as being of French, Italian, or British origin. By "BRITISH", we mean that at least one of the categories "IRISH," "SCOTTISH," or "ENGLISH" was chosen. For example, if a test was done on the proportion of people selecting "FRENCH", μ_x and μ_y were defined as the expected proportion of questionnaires where the response "FRENCH" would be chosen in versions X and Y of the questionnaire. In all regions, we tested

$$H_0: \quad \mu_x = \mu_y$$

versus

$$H_1: \quad \mu_x < \mu_y.$$

The randomization tests were done using 999 permutations. The results are shown in Table 4.

Table 3
Critical Levels for the Rate of Multiple Responses and Multiplicity

		Multiple Response			Multiplicity		
Question	Region	P_x	P_y	$\hat{\alpha}$	R_x	R_y	$\hat{\alpha}$
ORIGIN	HALIFAX	0.435	0.536	0.087	1.617	1.914	0.062
ORIGIN	QUÉBEC	0.154	0.043	0.001	1.143	0.986	0.001
ORIGIN	MONTRÉAL	0.185	0.194	0.612	1.141	1.152	0.585
ORIGIN	TORONTO	0.127	0.122	0.393	1.124	1.125	0.495
ORIGIN	WINNIPEG	0.293	0.307	0.622	1.439	1.398	0.345
ORIGIN	VANCOUVER	0.285	0.296	0.621	1.440	1.392	0.280
IDENTITY	HALIFAX	0.220	0.335	0.035	1.244	1.502	0.029
IDENTITY	QUÉBEC	0.140	0.016	0.001	1.131	0.959	0.001
IDENTITY	MONTRÉAL	0.159	0.125	0.063	1.075	1.044	0.186
IDENTITY	TORONTO	0.186	0.120	0.001	1.154	1.075	0.005
IDENTITY	WINNIPEG	0.224	0.195	0.248	1.253	1.208	0.298
IDENTITY	VANCOUVER	0.186	0.183	0.457	1.182	1.137	0.202

Table 4
Critical Levels for Selected Variables

Question	Variable	Region	P_x	P_y	$\hat{\alpha}$
ORIGIN	FRENCH	QUÉBEC	0.127	0.897	0.001
ORIGIN	FRENCH	MONTRÉAL	0.038	0.210	0.001
ORIGIN	BRITISH	HALIFAX	0.321	0.837	0.001
ORIGIN	BRITISH	MONTRÉAL	0.034	0.092	0.002
ORIGIN	BRITISH	TORONTO	0.085	0.135	0.003
ORIGIN	BRITISH	WINNIPEG	0.167	0.234	0.054
ORIGIN	BRITISH	VANCOUVER	0.267	0.325	0.065
IDENTITY	FRENCH	QUÉBEC	0.138	0.899	0.001
IDENTITY	BRITISH	HALIFAX	0.153	0.828	0.001
IDENTITY	BRITISH	MONTRÉAL	0.022	0.117	0.001
IDENTITY	BRITISH	TORONTO	0.050	0.215	0.001
IDENTITY	BRITISH	WINNIPEG	0.074	0.276	0.001
IDENTITY	BRITISH	VANCOUVER	0.104	0.325	0.001
IDENTITY	ITALIAN	TORONTO	0.412	0.463	0.060

5. CONCLUSION

The results for tests on the rate of multiple responses are similar to those on multiplicity, which is not surprising. When you compare the critical levels for the question on ethnic origin to the critical levels for the question on ethnic identity, it is seen that the differences between the two versions of the questionnaire affect the responses to the question on ethnic identity the most.

Our main reason for using randomization tests was that the sample for Modular Test 2 was a non-probability sample. However, there are also other cases where randomization tests are appropriate. For example, to do a "Student's" t test for means equality the hypothesis of normality is required, and it must also be assumed that the variances are equal. These assumptions are not needed for a randomization test. It should be kept in mind that the results of a randomization test apply to the sample, and not necessarily to the entire population, unless a simple random sample is used.

REFERENCES

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

BRADLEY, J.V. (1968). *Distribution-free Statistical Tests*. Englewood Cliffs: Prentice-Hall.

DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28, 181-187.

EDGINGTON, E.S. (1987). *Randomization Tests*, (2nd ed.). New York: Marcel Dekker.

FISHER, R.A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.