Alphabetical Placement in Surveys of Persons at Institutions – A Simulation Study

Pedro J. Saavedra[1], Francine Barrington[2]
[1]ICF Macro, 11875 Beltsville Dr, Calverton, MD 21230
[2] ICF Macro, 11875 Beltsville Dr, Calverton, MD 21230

Abstract
It is often the case that a survey of persons at institutions (e.g. teachers at schools or staff at offices) requires that one or more persons be selected from each institution by an administrator at the institution. A common way of doing this is to ask that a list of eligible potential respondents be drawn and that a set of random numbers provided to the administrator be used to draw the sample. Experience shows that this task is both difficult and burdensome. An alternate approach when the number to be selected from each institution is small is to provide a list of names (first and last) and ask that the person who follows each alphabetically be selected. This method has several potential biases when the population as a whole or in some institutions differs from the one from which names were drawn. These biases are explored here through simulations.

Key Words:
Alphabetic Sampling, Bias, Simple Random Sampling, Simulation

1. Introduction

It is not uncommon that a survey calls for the sampling of persons from institution. When sampling one or more persons from institutions (e.g. teachers at schools or staff at offices), sometimes it is better to have an individual from the institution, such as an administrator or manager, select the individual. This may be for various reasons, such as there is not a list of individuals available, or the list is outdated. Typically, when asking institutions to select the sample, one common method is to ask them to enumerate a list. The contact letters provide a set of random numbers in descending order, and asks the contact person that they pick the first number that is no larger than the number of teachers. Then the teacher with that number in the enumeration is selected. For example, if a school has 45 teachers and the numbers are 106, 92, 67, 50, 32 and 6 then teacher number 32 would be selected. Another method is simply to ask the contact person to choose a person at random. However, these methods have several disadvantages. Firstly, they place a burden on the institution, which can lead to a failure to respond and decreased compliance. Second, the enumeration method can be difficult for untrained people to understand and implement, requiring calls seeking clarification or the sample being improperly selected. Third, when told to select a random person, institutions may end up selecting a non-arbitrary person, resulting in bias. This person may be a favored employee, or someone the sample selector thinks will do the best job.

Thus, the problem remains, how does one phrase instructions for persons carrying out the sampling to avoid both biased selection and confusion? The Alphabetic sampling method is an alternate approach aimed at easing the burden and the process for institutions in certain situations. However, it has several potential biases when the population as a whole or in some institutions differs from the one from which names were drawn. Section II discusses the method and its potential biases. Section III introduces the data used to conduct two simulations to explore these biases. Section IV discusses the findings, and Section V concludes.

2. Alphabetic Sampling Method

The Alphabetic Sampling method is an alternate approach when the number to be selected from each institution is small is to provide a list of names (first and last) and ask that the person who follows each alphabetically be selected. This method is easier for the selector to understand and execute. In one study of the use of technology by teachers, the first approach (enumeration and random selection) was first tried and there were a large number of calls seeking clarification of the method. In the end, it was not clear whether the method was understood or not. When the alphabetic method was used, the compliance was similar, but there were far fewer questions asked and fewer clarifications requested.

As mentioned in the introduction, this method has several potential biases when the population as a whole or in some institutions differs from the one from which names were drawn. If in a school one has two teachers with similar names – say John Johnson and Joseph Johnson – it would be less likely that the one that comes later alphabetically would be selected than that the first would be. This bias could extend itself to situations where persons of a given ethnicity who constitute a larger proportion of the institution they are associated with than of the population at large may have a lower probability of selection than others. Thus, the simulations attempt to answer whether or not the Alphabetic sampling method is biased, and whether or not it is appropriate for sampling from strata.

3. Data and Simulations

The names used have been obtained from a census file of the most common first and last names, along with the frequency. If the procedure is to be repeated 1000 times, 1000 last names are sampled with Probabilities Proportional to Size (where the size is the frequency) and the same is done for 1,000 first names. The names are then randomly matched. If 1,000 institutions had been selected, one name would be sent to each institution with the instructions that the potential respondent that follows the one sent be given the survey.

There are various methods of evaluating the sampling approach. One is simply to examine the degree to which each member of the frame has the same probability of selection (or the designated probability if a stratified sample with unequal sampling fractions). There are at least two measures which can be applied. One is to repeat the procedure twice and obtain a correlation between the number of times each member of the frame is selected under one procedure and how many times under the other. The other measure is based on calculating the formula $b=1-sr^2/sa^2$ where sr^2 is the variance of the number of times a number is selected through random sampling and sa^2 the variance of the number of times the number is selected through the alphabetic method. Note that these measures require that the number of samples be sufficiently large. It takes a large sample for the bias associated with repeated sampling of some persons and absence of others in the sample to become apparent.

The fact that a method oversamples certain persons and undersamples others is not problematic if the dependent variable were not correlated with the true probability of selection. So the question is whether using this method will result in biased estimates and whether the variance (or the root mean square error) is much greater than that of a random sample.

In order to examine these issues it was necessary to use a data base with real data and real numbers. This is ordinarily a problem since most public access data bases hide the names. It was decided that a good starting point for the study would be using a small data base and sample with replacement using the alphabetic method and using simple random sampling with replacement.

Two simulations were run: a simulation using members from the 2004 Congress, and another simulation using baseball players over teams and over seven years of Major League Play. For the Baseball Simulation, each year/team combination was selected as a stratum. The data ranged from 2001 to 2007. Initially, 2008 data was considered, however, spellings varied in the data base used. Overall, there were 210 strata, comprised of thirty teams times seven years. The Alphabetic sampling method was applied using only male names and a random selection of one per stratum was chosen. Each method was test for biased and compared for the means squared. Estimated values were also compared to the population value.

For the Congressional simulation, the database consisted of 437 members of the 2004 Congress. As a starting point, any party affiliation other than Democrat or Republican was recoded (all non-Republicans were coded Democrat) and all missing values were replaced by the mean of the party's ratings. As opposed to the Baseball study, both male and female names were used, and there were no strata; the Alphabetic method was applied without replacement. In order to compare the random sample with the alphabetic sampling method, two sets of 500 samples were drawn. One was a set of 500 simple random samples of 100 with replacement. The other was a set of 500 samples of 100 with the alphabetic method. Estimates for party affiliation, percent of votes and 15 ratings were obtained for each sample. In one case, two congressmen had the same name, so the first name of one was changed from Mike to Michael in order to avoid ambiguity. Each of the 17 estimates were tested for bias for each of the two methods, by subtracting the population mean from the estimate and testing for significant difference from zero. The raw, absolute and square deviations from the population mean each variable were tested for statistical significance for the two methods. So was the variance of the estimates. Finally, the sum of the weights was calculated for each congressman across samples, and their distribution was examined for each of the two methods.

For each simulation, we kept count of the number of times each member of the population was selected, and calculated the standard deviation for the alphabetical sample and the standard deviation for a random sample.

4. Findings

The results proved to be mixed. While the congressional simulation showed findings with obviously biased results, the baseball study, while biased, had no difference in means squared deviation.

In the case of the congressional simulation, if there were two cases close alphabetically, (e.g. the Diaz-Balart brothers), if one is selected and then removed, the other may be sampled, but the number of times the second may be selected will be greatly diminished. Some congressmen were never selected by the alphabetic method. For example, Darrell Issa was not selected at all, because a randomly selected name would have had to come between that of Steve Israel and him. In turn, Israel was only selected 19 times because he was preceded by Johnny Isakson who was very close to him alphabetically. Mike Rogers was also not selected, as he was preceded by his namesake (recoded Michael Rogers). In 500 samples, a few did not appear. The question was whether these issues would affect the estimated variables.

The answer is that they unquestionable did. All of the variables but one were significantly biased under the alphabetic method. None of them were, of course, biased under the random sample. Indeed, while the proportion of Republicans was 52.3%, and the simple random sample estimated on the average 52.6%, the alphabetic method tended to select Democrats and yielded an estimate of 48.9% Republicans. Both absolute deviations and square deviations were significantly larger for the alphabetic method than for the random method.

One sense in which the alphabetic method was no worse than the random method was variance of the estimates (as opposed to variance from the population mean). There were no differences between the two methods in the variation from the mean estimates across samples, but there were substantial differences in variation from the population mean. In addition, one variable was the exception and did not have any bias. The percentage of votes attended was the only non-partisan variable in the dataset, and was also no worse in RMS error than the stratified random sample. The bias was so prevalent, that the alphabetic sampling method yielded an average republican minority of 49 percent, when in reality the population had a republican majority of 52 percent.

Table 1: Congressional Simulations – Means

Description	*Frame Mean*	*Alpha Mean*	*Strat Rdm Mean*	*Alpha Bias T*	*Alpha Bias Prob*	*SR Bias T*	*SR Bias Prob*
Party (1=D 2=R)	1.53	1.49	1.53	-20.52	<.0001	0.56	0.5789
Percent votes '02	68.11	68.07	68.08	-0.95	0.3420	-0.61	0.5410
Democratic Action '02	43.46	46.26	43.33	18.91	<.0001	-0.75	0.4537
Democratic Action '03	48.01	50.20	47.93	15.20	<.0001	-0.46	0.6477
ACLU	43.66	45.62	43.53	16.38	<.0001	-0.97	0.3327
AFSCMR	45.60	49.00	45.45	20.95	<.0001	-0.79	0.4280
People for the American Way	42.94	45.24	42.82	16.14	<.0001	-0.70	0.4864
Conservation	46.06	47.63	45.88	12.51	<.0001	-1.22	0.2239
Public Health	49.03	52.04	48.99	19.04	<.0001	-0.22	0.8296
Family/Repro.Health '99-'02	45.84	47.06	45.64	8.05	<.0001	-1.16	0.2453
Education '03	55.76	59.09	55.68	24.09	<.0001	-0.50	0.6152
Chamber of Commerce	68.58	66.90	68.63	-18.36	<.0001	0.51	0.6119
Right to Life	55.47	53.72	55.60	-11.34	<.0001	0.75	0.4546
Conservative Union	53.58	50.94	53.72	-18.11	<.0001	0.81	0.4208
Tax Limits	49.54	46.43	49.67	-21.65	<.0001	0.79	0.4309
Christian Coalition-02	53.39	50.93	53.55	-16.68	<.0001	0.95	0.3434
Christian Coalition-03	57.50	55.47	57.56	-15.42	<.0001	0.42	0.6760

Table 2: Congressional Simulations – Real Means Squared and Absolute Deviation

Description	*Alpha RMS*	*SR RMS*	*Diff t- val*	*t-val Prob*	*Alpha Abs Dev*	*SR Abs Dev*	*Diff T-Val*	*T-Val Prob*
Party (1=D 2=R)	0.0030	0.0023	3.42	0.0006	0.0451	0.0391	3.22	0.0013
Percent votes '02	0.7400	0.8885	-2.10	0.0356	0.6878	0.7503	-1.81	0.0701
Democratic Action '02	18.7589	15.8242	2.03	0.0426	3.5092	3.2503	1.69	0.0911
Democratic Action '03	15.1999	14.1640	0.84	0.3984	3.1173	3.0978	0.14	0.8910
ACLU	10.9658	9.6223	1.50	0.1347	2.6180	2.4821	1.10	0.2701
AFSCMR	24.6285	18.5061	3.48	0.0005	4.0549	3.5188	3.17	0.0016
People for the American Way	15.5237	13.2939	1.79	0.0741	3.1273	2.9454	1.26	0.2072
Conservation	10.2927	11.0209	-0.80	0.4216	2.5968	2.6930	-0.79	0.4270
Public Health	21.6266	16.3319	3.28	0.0011	3.7395	3.2533	2.97	0.0031
Family/Repro.Health '99-'02	12.9225	14.9192	-1.59	0.1127	2.8049	3.0538	-1.70	0.0888
Education '03	20.6716	13.0068	5.70	<.0001	3.7689	2.9351	5.65	<.0001
Chamber of Commerce	6.9707	5.7371	2.35	0.0190	2.1363	1.9688	1.81	0.0705
Right to Life	14.9209	15.5793	-0.47	0.6398	3.0164	3.1404	-0.82	0.4151
Conservative Union	17.6695	14.7524	2.15	0.0316	3.3920	3.1284	1.76	0.0779
Tax Limits	19.9383	14.2035	4.09	<.0001	3.6548	3.0870	3.78	0.0002
Christian Coalition-02	16.9027	14.3275	1.92	0.0557	3.2721	3.0510	1.48	0.1405
Christian Coalition-03	12.7563	11.6122	1.08	0.2788	2.8660	2.7357	0.99	0.3229

Table 3: Congressional Simulations – 2004 Congress - # Times Sampled by Method (Selected Names)

Obs	*Last Name*	*First Name*	*Alpha Method*	*Stratified Random*
58	CANTOR	ERIC	61	111
59	CAPITO	SHELLEY MOORE	40	109
60	CAPPS	LOIS	13	130
61	CAPUANO	MICHAEL	0	116
62	CARDIN	BEN	21	112
63	CARDOZA	DENNIS	8	103
64	CARSON	BRAD	227	108
190	INSLEE	JAY	109	99
191	ISAKSON	JOHNNY	56	112
192	ISRAEL	STEVE	19	112
193	ISSA	DARRELL	0	126
194	ISTOOK	ERNEST	1	116
195	JACKSON	JESSE JR.	115	121
196	JANKLOW	BILL	234	120
294	NUNES	DEVIN	57	118
295	NUSSLE	JIM	41	126
296	OBERSTAR	JAMES	35	113
297	OBEY	DAVID	0	110
298	OLVER	JOHN	220	103
299	ORTIZ	SOLOMON	195	113
300	OSBORNE	TOM	89	101

The results of the baseball simulation were much more encouraging. While there was a definite bias with alpha, there was no difference in the means squared variance. However, as was the case with the Congressional simulation, there was an instance where a person was not sampled once in 500 samples. Jose Molina, who shared a last name with Bengie Molina on the 2001 Anaheim Angels, was not selected with the Alphabetic sampling method. When creating a random sample, it tends to rank above or below the frame value. However, when sampling through the alphabetic sampling method, the sample had a tendency to fall in a particular direction, either over or under, but just as close as the random sample. It seems that for certain variables such as baseball statistics, the alphabetic sampling method works. Over many samples the results would be underestimating or overestimating the statistic in question, but in any given sample it would be as close to the population parameter as with stratified random sample.

Table 4: Baseball Simulations – Means

Description	Frame Mean	Alpha Mean	Stratified Random (SR)	Alpha Bias T	Alpha Bias Prob	SR Bias T	SR Bias Prob
Games played	77.489	77.456	77.465	-0.21	0.8314	-0.15	0.8825
On Base Percentage+Slugging	0.649	0.647	0.650	-3.09	0.0021	1.80	0.0727
Slugging Average	0.361	0.360	0.362	-3.08	0.0022	1.73	0.0836
Batting Average	0.235	0.234	0.235	-3.86	0.0001	1.00	0.3195
On Base Percentage	0.294	0.293	0.294	-2.82	0.005	1.70	0.0898
At Bats	241.391	239.079	241.554	-3.75	0.0002	0.26	0.7976
Runs	33.424	32.993	33.472	-4.27	<.0001	0.46	0.6439
Hits	64.256	63.550	64.304	-3.92	0.0001	0.26	0.7964
Two-base hits	12.962	12.868	12.982	-2.52	0.0122	0.50	0.6173
Three Base Hits	1.340	1.299	1.341	-6.89	<.0001	0.17	0.8617
Home Runs	7.592	7.596	7.621	0.11	0.9122	0.92	0.3606
Runs Batted In	31.871	31.658	31.936	-2.14	0.0325	0.63	0.5257
Base on balls	23.080	22.849	23.136	-2.96	0.0033	0.73	0.4684
Strike outs	45.371	45.027	45.503	-3.02	0.0027	1.10	0.2713
Stolen Base	3.994	3.988	3.991	-0.28	0.7832	-0.13	0.8943
Caught Stealing	1.682	1.678	1.680	-0.48	0.6301	-0.26	0.7984

Table 5: Baseball Simulations – Real Means Squared and Absolute Deviation

Meaning	*Alpha RMS*	*SR RMS*	*DiffT -val*	*T-val Prob*	*Alpha Abs Dev*	*SR Abs Dev*	*Diff T-val*	*T-val Prob*
Party (1=D 2=R)	0.0030	0.0023	3.42	0.0006	0.0451	0.0391	3.22	0.0013
Percent votes '02	0.7400	0.8885	-2.10	0.0356	0.6878	0.7503	-1.81	0.0701
Democratic Action '02	18.7589	15.8242	2.03	0.0426	3.5092	3.2503	1.69	0.0911
Democratic Action '03	15.1999	14.1640	0.84	0.3984	3.1173	3.0978	0.14	0.8910
ACLU	10.9658	9.6223	1.50	0.1347	2.6180	2.4821	1.10	0.2701
AFSCMR	24.6285	18.5061	3.48	0.0005	4.0549	3.5188	3.17	0.0016
People for the American Way	15.5237	13.2939	1.79	0.0741	3.1273	2.9454	1.26	0.2072
Conservation	10.2927	11.0209	-0.80	0.4216	2.5968	2.6930	-0.79	0.4270
Public Health	21.6266	16.3319	3.28	0.0011	3.7395	3.2533	2.97	0.0031
Family/Repro.Health '99-'02	12.9225	14.9192	-1.59	0.1127	2.8049	3.0538	-1.70	0.0888
Education '03	20.6716	13.0068	5.70	<.0001	3.7689	2.9351	5.65	<.0001
Chamber of Commerce	6.9707	5.7371	2.35	0.0190	2.1363	1.9688	1.81	0.0705
Right to Life	14.9209	15.5793	-0.47	0.6398	3.0164	3.1404	-0.82	0.4151
Conservative Union	17.6695	14.7524	2.15	0.0316	3.3920	3.1284	1.76	0.0779
Tax Limits	19.9383	14.2035	4.09	<.0001	3.6548	3.0870	3.78	0.0002
Christian Coalition- 02	16.9027	14.3275	1.92	0.0557	3.2721	3.0510	1.48	0.1405
Christian Coalition- 03	12.7563	11.6122	1.08	0.2788	2.8660	2.7357	0.99	0.3229

Table 6: Baseball Simulations – Anaheim Angels 2001 - # Times Sampled by Method

Obs	*Last Name*	*First Name*	*Alpha Method*	*Stratified Random*
1	ANDERSON	GARRET	23	29
2	BARNES	LARRY	9	31
3	DAVANON	JEFF	94	31
4	ECKSTEIN	DAVID	21	31
5	ERSTAD	DARIN	2	34
6	FABREGAS	JORGE	4	34
7	FERNANDEZ	JOSE	7	28
8	GIL	BENJI	24	25
9	GLAUS	TROY	1	16
10	HILL	GLENALLEN	36	23
11	JOYNER	WALLY	31	27
12	KENNEDY	ADAM	6	20
13	MOLINA	BENGIE	55	22
14	MOLINA	JOSE	0	31
15	NIEVES	JOSE	9	28
16	PALMEIRO	ORLANDO	8	35
17	SALMON	TIM	57	17
18	SPIEZIO	SCOTT	39	21
19	WOOTEN	SHAWN	74	17

5. Conclusion

The results of the two simulations show that further study is warranted. The congressional simulation had a significantly larger real means squared error average of means obtained. The baseball study, while bias was present, showed no difference in accuracy in the sampling procedure. Ultimately, when sampling one to two persons per institution, for objective variables, the alphabetical method worked better. The questions that should be explored in further study should be why two different databases and simulations resulted in such different results. Is it the subject matter or the one sample per stratum that made the difference? Is the fact that politics is more related to ethnicity which is in turn related to names that is a factor?

There are other variants of the method which can be tried. One is to alternate between before and after, selected the name that precedes the name that is give with the name that follows. Another may involve skipping one name some times and not others. These may result in a more difficult set of instructions for the contact person, and a balance between these considerations will have to be established. Further research will include examining these approaches and using a larger population.