Analysis and Design of CMU Alumni Survey

Progress Report 2

Yi Jiang & Zhanwu Liu

March 9, 2007

1 Introduction

The client of this project is Judy Cole and Christian Krohn from Carnegie Mellon Office of Alumni Relations. Their primary interest is to ensure that alumni participation and concerns are considered through the university's advancement efforts. More specifically, in this project they are interested in designing a survey panel that is representative of the whole CMU alumni population. In this project we will analyze information from CMU alumni database and previous alumni surveys, and privide some suggestions for future panel survey design.

Based on the interest of the client, we formulated several problems that can be answered in one semester:

1. Analyze the bias of email based surveys. Nowadays, email is used widely in survey. But one problem is the sampling bias of email based survey, for example, some professionals are more likely to use email than other respondents. In our alumni database, only half of the records contain email addresses. We will do a regression or classification to analyze what is the difference between those records that have email address

and those records that do not have email address.

2. Design a panel survey. The sample should be representative of the population of all CMU alumni. We will simulate sampling from the database using different stratifying methods, and compare each sample with the full database.

3. Estimate the active alumni population. We will use the respondent data from previous surveys as well as other indicators as response, to identify the characteristics of the alumni who like to take part events related with the university. In addition, we will use capture/recapture methods to estimate the population size of all active members.

2 The Data

The office of alumni relations maintains a database with $86,240$ alumni records. The database contains the following information:

- Basic biographical information: name, address, employment information, personal relationships, degree information, affiliations, etc.
- Additional biographical information: activities/events, awards, committees, interests, philanthropic affinities, publications, student activities, sports, etc.
- Prospect information: wealth, prospect rating, etc.
- Giving summary: appeals, gift clubs, total giving, largest gift, matching claims, etc.

In addition, some data from previous surveys are also available: In 2002, a survey was conducted using mail, email, and phone; in 2004, a survey was conducted using email but supplemented with a little bit mail; and in 2006, a survey was conducted using email only.

We have signed the confidentiality agreement and obtained most of the data on Feb 28. The client has provided us the records from the alumni database, including all the variables

we have asked. But it is also possible that we may need more variable in the future. The client also provided us the 2006 survey results with the key to match the record in the database. At this point the client is not able to provide us the 2004 survey data, but Christian is working on it now.

3 Exploratory Data Analysis

The basic biographical data is the most complete set of data in the database, it contains record for all the alumni in the database, and contains most of the strata information(geographical, year/school, etc.) that can be used in stratified sampling. Thus the first thing we did is some EDA on basic biographical data.

3.1 Gender

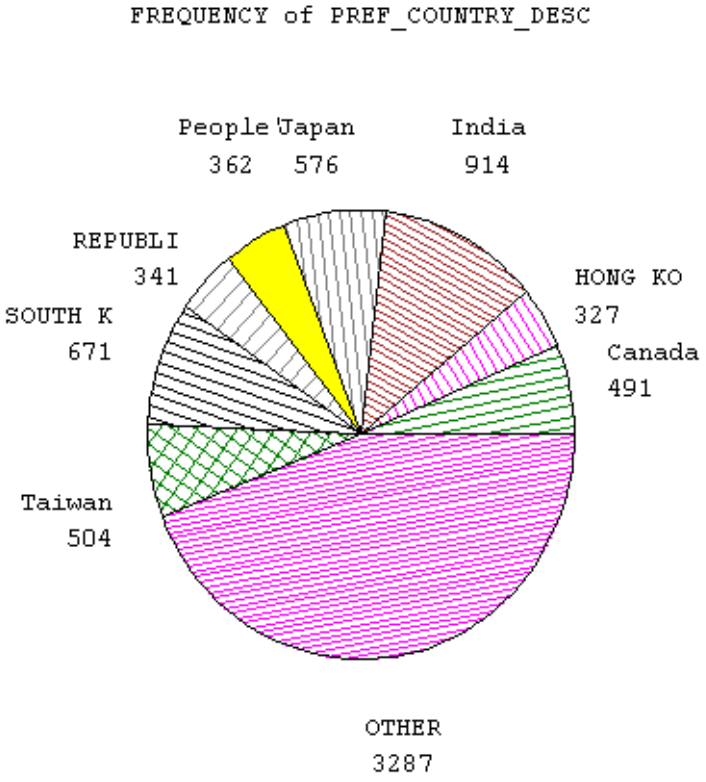
In the alumni database, there are $86,240$ records, including $25,492(29.56\%)$ female and $60,747(70.44\%)$ male. $7,473$ records have international addresses. This information is missing in one record.

3.2 Geographical distribution of the alumni

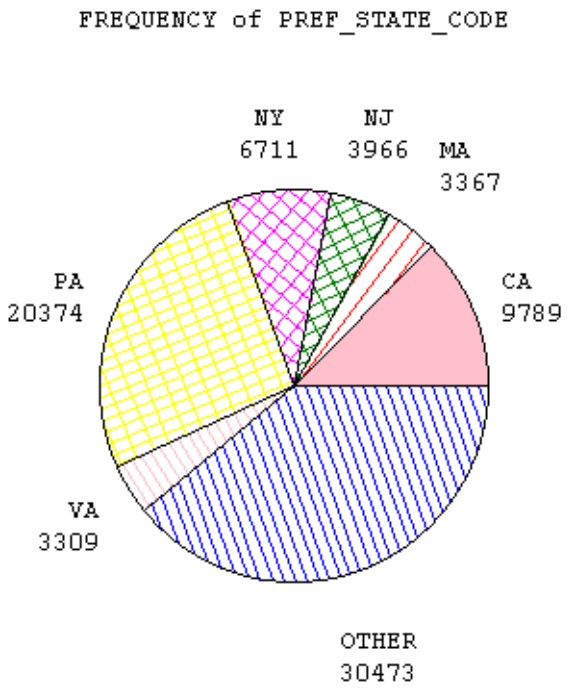
Out of all record, $7,473$ records have international addresses, and $77,986$ records have US addresses (including Puerto Rico). Address information is missing for 981 records (no zip code, and no country name).

The distribution of international students are plotted below. Note that because the country/region name is incomplete in the data we received(only 8 characters for the country name), we expect some possible change in the future after we obtained accurate country name from the client. But from the pie chart, it is still very clear that a large portion of

the international alumni are in Asia countries/regions. One should note that this number is not same as the number of international-origin alumni because the alumni may work in the different country than he came from.

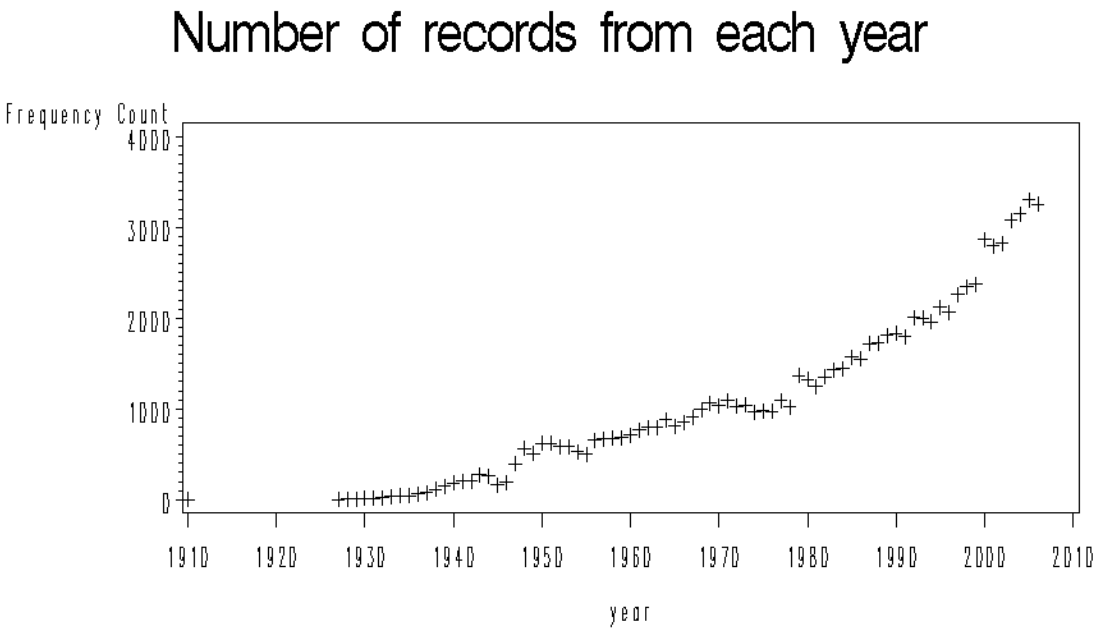


For the alumni in the US, the following pie chart shows the distribution of alumni in the states. It is natural to expect that a great fraction of the alumni will be in PA, and indeed this is the case. CA comes second, followed by east coast states (NY, NJ, MA and VA).

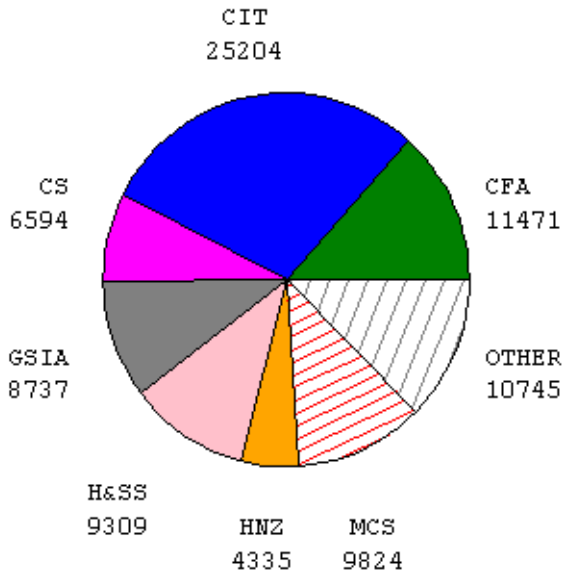


3.3 School and graduation year

The number of alumni graduated in each year is plotted below. There is a clear increasing tendency over the years. We assume that the increase in number is due to the increase in number of students recruited. However, we also need to study the history of the database and estimate the underestimate of alumni number in earlier years when all the records were input manually.

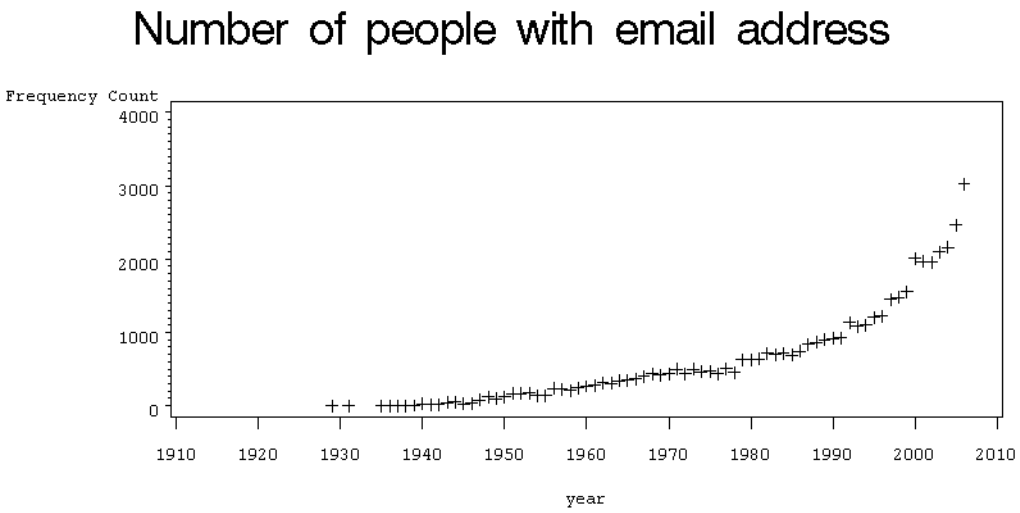


The number of records from each school is also plotted. Note that since there is some change in the school names, for example GSIA changed to TSB in 2004, the pie chart can be improved after we find out all school name changes in the past.

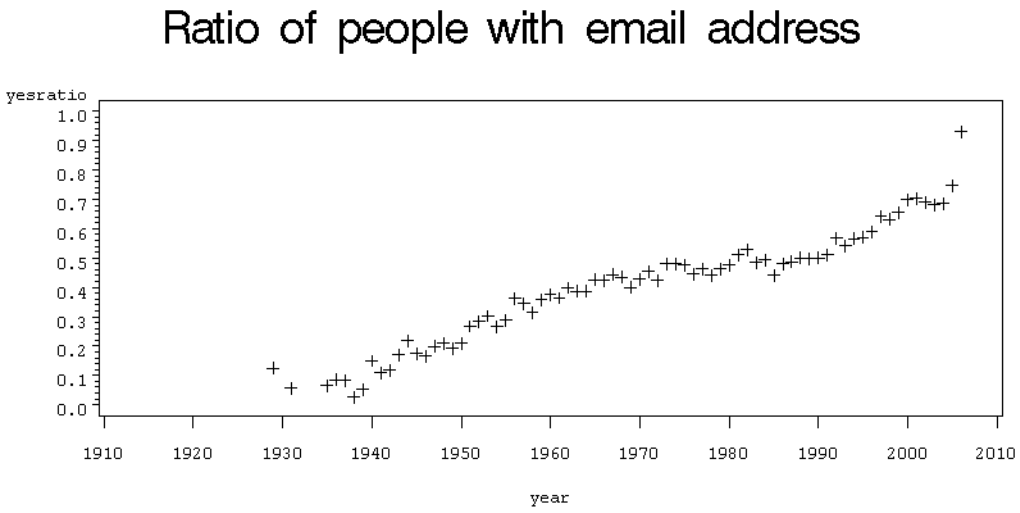


3.4 Email address information

More than half of the records $(45,971(53.31\%))$ contain email information. The number of people with email address is plotted below.



The ratio of people with email address:



It is natural to expect that the number of people with email increase with year, because the overall number of alumni increase over year. It should be noted that the ratio of records with email also increase over years, this tendency must be accounted for when doing email based surveys.

4 Discussion

4.1 Difficulties

There are several difficulties facing the project. First, it is hard to quantify the involvement of the alumni in university advancement. Monetary donations given to the university, if used as a response variable, is easy to quantify. But there is a drawback with the use of donation amount. Because the amount given is related to many factors, including wealth, giving habit, personal life, etc, we could not and should not relate the amount directly with the attitude towards university advancement. Of course, other indicators are no better than the giving amount, and even more difficult to quantify, such as involvement in recruiting new students to CMU, event participation, etc. Thus, another possibility is that we just classify the alumni as active and non-active, but not try to quantify "how active" they are.

Second, to assess which sampling method can generate a sample that is representative of the population, we will need to learn about survey sampling and panel survey. Currently we are reading the related books and papers.

Besides, some technical difficulties exist. For example, to simulate sampling, I need to process and extract alumni information from the $86,240$ records, that require some specialized software, we choose to use SAS, but are not familiar with the program. Another technical difficulty comes from unfamiliar with the theory behind the methods. Dealing with those difficulties would be a challenging but fruitful process.

5 Suggestions for the Design of the Panel Survey(literature review)

Panel surveys are a kind of survey that measures the same sample at different points in time. They offer attractive features that cross-sectional surveys do not have. To better design the panel survey, its good to first review the purposes of a survey for different kinds of information needs.

1. understand the characteristics, behavior, hobbies, and attitudes
2. understand the external factors that are associated with or influencing the characteristics, behavior, and hobbies
3. estimate the changes of these characteristics over time
4. estimate the relationships among characteristics
5. estimate the frequency of occurrence for specific kinds of events
6. estimate the size of a specific kind of group of people
7. estimate the influences of surveys on the characteristics
8. estimate the causal effects of a specific characteristic
9. estimate the group characteristics by the individual characteristics

In designing a survey to understand and then improve the relationship between the Carnegie Mellon alumni and Carnegie Mellon, estimates of the changes of the alumnis relationship with the school with time, and the changes due to the external factors, such as the school activities for the alumni, are certainly of great importance to the development of the strategies for the improving the relationship. In summary, purposes 2), 3), 6), 8), 9) from the above

list are to be investigated in this project. And for these purposes, panel surveys are the right choice, where we can look at how the exogenous and endogenous factors interrelate with each other over a short or long time.

However, there are difficulties in conducting panel surveys which need a great deal of care, since they could affect the survey design, such as the basic sample structure, administration of the survey, and the database structure, estimation, and analysis of the panel data. [1]

According to [1], these difficulties mainly arise from three features in conducting panel surveys. They are 1) interview spacing, 2) the mode of the interview, and 3) respondent selection. In the sections below, we describe these features and further infer their influences and analyze possible solutions, particularly in our project.

5.1 Interview Spacing

Interview spacing is the most important factor that affects the design and the response characteristics. Proper selection of the interview spacing should take the following four facts into account:

1. If the interview spacing is too short, meaningful changes may not have taken place and its not meaningful to take the next survey.

2. If the interview spacing is too long, there will possibly be telescoping and omissions of events in the panel data, since the respondents are not likely to remember those that happened too long ago. For example, under some sociological topics such as crime victimization rates survey, short recall periods are shown to have significant higher rates than long recall periods.

3. There are panel effects that will affect the survey results. Panel effects are the results of behavioral effects which is a common problem in many types of evaluation stud-

ies where peoples behavior might be changed after participating in the survey. Take the election for example, people who took the survey may pay more attention to the election after taking the survey and this will certainly affect their final decision on the vote.

4. This is the budget of the survey itself to find the optimal length of time between interviews.

There are implications of these facts on the interview spacing in our project.

Since a majority of the panel survey is expected to be done by emails, we want to be able to track the person through the same email address as before. Therefore, interview spacing for certain topics or events should not be longer than the expected time that people are going to change their email addresses. For example, people who have graduated for a short period of time have the tendency to change their email addresses more frequently. Surely, the difficulty on choosing the interview spacing when considering this fact can be eased by sending surveys to more reliable email addresses if we can get them.

5.2 Mode of the Interview

The mode of the interview has been shown to be related to the response rate on many cross-sectional surveys (Dillman, 1978; Oksenberg et all, 1986). Particularly, the mode of the first interview in the panel survey will affect both the response rate of itself and that of the subsequent interview waves. Some studies showed that in-person surveys are good modes for the interview although it might be costly. However, partly for this reason, in-person surveys are conducted in the initial interview of the panel survey to obtain the best tradeoff between high response rates and low costs.

Particularly, this is important in our project because surveying by email can actually be better than surveys by mail or in person if its advantages are taken and disadvantages are

avoided. Specifically, its advantages over mail or in person delivery methods include faster delivery, lower cost, wider coverage, and easier maintenance. Its disadvantages include the reduced deliverability due to spam filtering, rarely used email addresses, and the limited number of techniques to express the information. Thus, its worth attractive design and wording of the email title and the first few paragraphs. Particularly, highlights of the benefits of the survey for the participants in those locations are of great help in increasing the response rate in both the first interview and the subsequent interviews.

Another advantage in using emails in our project that is worth mentioning is that panel surveys by email make it easier to track the respondent if the interview spacing is properly chosen. Thus, its the combined consideration of the first two features we just discussed.

5.3 Respondent Selection

Certain types of panel survey may use a variety of types of respondents. These different types of respondents have different effects on the response quality. For example, bias can be introduced if the respondents have different levels of exposure to the survey questions, particularly when some respondents are trained by their exposure to the survey and they do well in the later interviews. This is also important in our project since we need to rotate a fraction of people out of the panel after surveyed a number of times. Here we need to make sure the new selected people can also represent the whole population, without a significantly higher or lower level for the panel.

6 To-dos

Following is a list of to-dos in the second half of the semester.

6.1 Data analysis

1. Characterize the difference between those with and without emails.
2. Find a method to assess how good a sample represents the population.
3. Based on previous method, find a good stratifing strategy.
4. If we can have the 2004 survey data, use both 2004 and 2006 survey data as well as other event participation data to estimate the number of active population.

6.2 Ask Christian for the following information

1. We still don't have the 2004 survey data.
2. Country name for internationals are incomplete in the biographical information.
3. The history of the alumni database.

References

[1] Kasprzyk, D., G.J. Duncan, G. Kalton, M.P. Singh (1989), Panel Surveys, John Wiley & Sons, New York.

[2] Dillman, D.A. (1978), Mail and Telephone Surveys: The Total Design Method, John Wiley & Sons, New York.

[3] Oksenberg, L., Coleman, L., & Cannell, C.F. (1986), "Interviewer Voices and Refusal Rates in Telephone Surveys," Public Opinion Quarterly, 50, 97-111.