K: Sampling Scheme 15/30 Total: 85/100

L: Questionnaire 50/50

M: Sample Size 20/20

Assignment II.4. Sampling Scheme & Question Design
Spatial and Analytical Study of Student Housing at Carnegie Mellon

Ariel Liu, Sam Lavery, Alejandra Munoz, Terra Mack, Shannon Lauricella

A. TOPIC

Carnegie Mellon is an urban university with many students living off-campus. Finding housing off-campus is generally left up to individual students, who take into account many variables when choosing a house or apartment. Many students list their off-campus addresses in the C-Book directory published by Alpha Phi Omega (APhiO). We are interested in investigating the possibility of a correlation between where students choose to live and what they choose to study. The results of the survey will be a valuable tool that would be useful to the university for the planning of shuttle routes, campus police coverage, and future housing projects. Students would also be able to use the survey results to find neighborhoods in the city that are popular with other students like themselves. We are seeking to answer questions about the dynamics of student housing at CMU. An example of this is: Is there a correlation between address (either on-campus building or off-campus neighborhood) and major? Do students in certain majors cluster together?

K. SAMPLING SCHEME

We were successfully able to attain off-campus housing records from the University registrar. The records have 891 undergraduate records and 4,036 graduate records. According to the CMU Factbook, (found at http://www.cmu.edu/ira/factbook/pdf/facts2012/11_campus-space-section-final.pdf), there are 2,252 undergraduates living off-campus and 5,769 graduates living off-campus. Clearly, the ratio of undergraduate records to graduate records is not the same as the population ratio, but there could be response errors that affect undergraduates more than graduate students. Most undergraduates start their CMU careers living on-campus so changing their address to an off-campus location will probably be less likely reported to the registrar (especially if they still use their SMC mailboxes to get mail from the university).

elaborate - what's your concern here? I can't tell what you're referring to.

If we were to use random sampling on the records we have, we are worried about the possibility of creating a coverage error. So, we think that including all of the records we have (after cleaning the data, there will definitely be fewer records for graduate students) will still be appropriate for our analysis.

these are in conflict -- what do you mean?

Our sampling frame is clear and we will implement a stratified SRS without replacement. We think that the decision where to live for students vary based on their needs and expectations. While graduate students might search for quiet places, close to groceries stores, larger departments/houses and not so expensive rental if they are accompanied by family, Undergraduate students might search for more active places, close to restaurants and smaller department/houses because a higher percent of them are singles. Taking into account the differences on needs among graduate and undergraduates student, we will divide the population (stratification) between undergraduate and graduate students.

L. QUESTIONNAIRE OR OBSERVATIONAL PROTOCOL:

(1) is this info from the registrar or your conjecture? is it the only reason you do not have records for all students?
(2) is this set of 4927 records all that the registrar has?
(3) what other sources of bias might there be, in this collection of records (what else did the registrar say?

Undergraduate/Graduate

Is the person an undergraduate student?

Is the person a~~n~~ graduate (Master) student?

Is the person a~~n under~~graduate (PhD) student?

College

Is the person a member of Marianna Brown Dietrich College of Humanities and Social Sciences (DC) (ex -HSS)?

Which department?

Is the person a member of Carnegie Institute of Technology (CIT)?

Which department?

Is the person a member of David A. Tepper School of Business (TSB)?

Which department?

Is the person a member of School of Computer Science (SCS)?

Which department?

Is the person a member of College of Fine Arts (CFA)?

Which department?

Is the person a member of H. John Heinz III College at Carnegie Mellon University (HC)?

Which department?

Is the person a member of Mellon College of Science (MCS)?

Which department?

Housing

Residence address

Type of Building - a house? An apartment? Number of stories *

Neighborhood

City

Distances

Distance to Campus

Distance to shuttle/bus stops

Distance to Restaurants

Distance to Grocery stores/Supermarkets

Distance to Pharmacies

these questions seem to be an ok start.

please do some pre-testing of the with a wide variety of student records, in order to catch anything that will be difficult to extract from the records, difficult to interpret, etc.

you may also discover other questions that you should add to this list.

Are students from some colleges or majors more likely to live off campus?

Do undergrads/grads, colleges or majors cluster together? If so, where?
We plan to further develop this question in looking at each college (HSS, Tepper, SCS, etc) and then majors within each college. This question will ultimately have many different results.

* This information is obtained from Department of city Planning of Pittsburgh - GIS database

M. SAMPLE SIZE:

To calculate the sample size we select the following question: Is this person a member of CIT?. Then, We used the source of the factbook from February 2012 which provided a head count of students in each college in the Fall Semester 2011 (only for Pittsburgh, PA campus) to calculate a value for p.

The total head count of students: 10,957

The head count for CIT students: 3,217

The proportion of CIT students out of total students: p : 3,217/10,957 = .293 = 29.3%
Total population size of students living off-campus: 2,252+5,769 = 8,021
p = .293
n = 8021
$z_{\alpha/2}$= 1.96
SD = sqrt (p(1-p)) = sqrt (.293(1-.293)) = .4551
The first that we considered was 0.05
ME = $z_{\alpha/2}\frac{SD}{\sqrt{n}}$ = 0.05
With this value we calculate n for a SRS with replacement.
1.96 (.4551/$\sqrt{n}$) = 0.05 ---> n = 318

Because the sample size is small we tried smaller MOE values.
Second ME = 0.001
1.96 (.4551/$\sqrt{n}$) = 0.01 ---> n = 7,957
Third ME= 0.012
1.96 (.4551/$\sqrt{n}$) = 0.012 ---> n = 5,525
Fourth ME = 0.011
1.96 (.4596/$\sqrt{n}$) = 0.011 ---> n = 7,800

We estimated the sample size using a MOE of 0.05 as our first reference point. Then we decided that our sample of data available to us was much larger (in our database given by the registrar) and we could use a lower MOE. Thus, we estimated the sample size again to match our sample size available to us using an MOE of 0.01. However this sample was larger and closer to the sampling frame size we have available. We also took into account that we must perform data-cleaning and this may leave us a lower number of records. Therefore, we estimated the sample size using a slightly larger MOE of 0.0012 to obtain a sample size of 5,525 which seems more reasonable for the data available to us. Table 1 contains the different MOE values used and the n values obtained.

Table 1. MOE selected and n values obtained for defining a sample size

MOE	sqrt(n)=(1.96*SD)/MOE	n
0.010	90.08	7957
0.011	89.19	7800
0.012	75.07	5525
0.015	60.05	3607

0.050	18.02	325

Then we calculated the adjustment needed for sizing a sample without replacement:

Adjustment calculation for SRS without replacement:
$n>= (N*n_0) / (N+n_0) = (10,957)(7,955) / (10,957+7,955) = 3,672.9$

Seeing the size of the data obtained, we reconsidered the MOE and n used for the SRS without replacement. In this case we used a n value of 7957.

$n>= (N*n_0) / (N+n_0) = (10,957)(7,955) / (10,957+7,955) = 4,609$

This results gave us a larger size of the sample. Because we have a larger sampling frame, we think a n size of 4,609 is reasonable. However, after cleaning the data we could use a more accurate MOE value and sample size number..

Reference:
Carnegie Mellon University Factbook Volume 26. Headcount Enrollment by Location of Study, Home College, Level, and Status. Fall Semester 2011. Office of Institutional Research and Analysis. http://www.cmu.edu/ira/factbook/pdf/facts2012/entire-fb-for-web-as-of-3-1-121.pdf

what sort of cleaning do you have in mind?

I'm assuming that the data is in an excel file or data base which makes looking at this many records feasible.

I want you to look at the records also -- you will inevitably find coding errors, extra codes, missing values, etc. etc that will mess up any "automatic" tabulation of results. Fixing these things is a part of the data cleaning you will need to do.

Please include the following information in I.6:

(1) whether the 4927 records you have are *all* the records the registrar has, or just a subset

(2) a *complete* discussion of
- what determines whether the registrar has a housing record for a student (using information you get from the registrar and/or other sources)
- how the registrar selected records for you (if you have only a sample of the registrar's records).

(3) one or two sample records (so that I can understand what you are working with)

(4) a complete plan for what things you will do for data cleaning. if possible, also examples of records that need to be fixed in the data cleaning, and what you plan to do with those records.