36-303: Sampling Surveys and Society
Some Sampling Details

Brian Junker
132E Baker Hall
brian@stat.cmu.edu
March 3, 2012

Contents

1 Selecting a Sample from C-Book

1.1 My sample size calculation says I need 100 respondents. So I just sample 100 from C-Book, right?

We have seen that response rates for the most convenient survey methods are not very good.

- The 2007 Pew survey on religion in America was a very well run national telephone survey by a modern, professionsl survey organization. The response rate was around 35%.
- The Spring 2007 Dietrich College of Humanities and Social Sciences (HSS) survey of students' satisfaction with HSS advising was conducted by inviting *all* HSS students *by email* to go to surveymonkey.com to fill out a survey. There was no nonresponse followup of any kind, except for generic spam from HSS that the surveymonkey survey was "still open". As an incentive a raffle was held for a prize (an ipod, I believe) to be given to one of the students who completed the survey. The response rate was around 23%.

1.1.1 How many people to contact

Your survey's initial response rate is going to be somewhere between 10% and 30%. Let's be optimistic and say 25%. And let's say your sample size calculation leads you to want 100 respondents. Let x be the number of people you contact. Then you want

$$x \cdot 0.25 = 100$$

or

$$x = 100/0.25 = 400$$

so, to get a sample size of 100, you need to contact about 400 people. That is to say, you should sample 400 people from C-Book.

For that list of 400 people, go through it one-by-one[1]. Keep track of how many names on the list you've contacted as you begin to accumulate respondents. Let's say you had to contact 265 people on the list to get to your sample size of $n = 100$. Then your actual response rate was

$$r = 100/265 = 38\%,$$

pretty good! A lot better than 25% anyway!

[1] Do the names in the order they are on your list! Don't do the ones you can easily find on Facebook first, and then the email ones, and then the other last. This will distort your sample by selection effects (what's different about Facebook people? What's differnt about Facebook people that you have friended?) and you won't have a valid SRS anymore.

1.1.2 What to do with a low response rate

You will have to make a lot of effort with nonresponders to get the response rate up into the 40's, let alone above 50%.

But it may be worth it. Remember, the higher your response rate, the less you have to worry about arguing that your sample really is representative of your target population.

If you do have a low response rate, then the demographic data you collect will come in handy. You want to compare proportions by gender, age, occupation/major, ethnic or national background, etc. with your target population, to argue that your sample is representative at least on these demographic variables. And if your sample demographics don't match up with the target population demographics, we can talk about how to compute sample weights to adjust the information in your sample so that it is more like a representative sample.

1.1.3 My sample size calculation says 100 but I've contacted 500 people already and all I have is 50 respondents. I'm sick of this!

At some point you will have to give up, take what respondents you have, and write up what you can. It may not be as bad as you think:

- If your sample size was calculated to represent a proportion, and you plugged in 0.5 to "be conservative" in making the sample size calculation, you may be pleasantly surprised: if $\hat{p}$ in your sample is very far from 0.5, a much smaller sample size will lead to reasonable confidence intervals for p.
- Even if your confidence intervals are larger than you wanted, you probably still can say interesting things about the data and about the quantities you estimated. Be honest about what you got, but also be on the lookout for interesting things to say *anyway*.

1.2 Making sure everyone is equally likely

C-Book has some interesting characteristics for people who are trying to make a simple random sample. The specific details below come from C-Book in 2007, but the ideas for the current C-Book are the same:

- The student listing in C-Book 2007–2008 run from page 10 to page 108, so 99 pages.
- Each page has a different number of students. Sometimes this is very obvious, for example p. 18 has fewer listings on it than p. 19, and p. 108 has fewer still.
- All students are mixed together on a page. Graduate students are in among undergraduates. On p. 19, there are even two campus police listing stuck in among the student listings. So even on two pages with the same number of listings, it is unlikely that the same number of undergraduates are listed on the two pages.

Let's assume you just want a sample of 400 undergraduates (because you need a sample of 100 for your survey). In the following two subsections are two acceptable ways to generate a sample.

1.2.1 Random page, random location

Figure out what the maximum number of listings per page is. I'll estimate that it's 110, but don't take my word for it. Check it out yourself. Guess too low, and you'll miss some members of your target population, increasing your coverage error. Guess too high, and you'll make extra work for yourself.

Now, generate 400 or so pairs of numbers:

$$(page_1, loc_1), (page_2, loc_2), \ldots, (page_{400}, loc_{400})$$

where each "page" in the pair is an independent random draw from the numbers 10 to 108. Each "loc" should be an independent random draw from the numbers 1 to 110 (or whatever your guess about the max number of listings is). *It doesn't matter if you sometimes generate the same page number and/or location. We will knock out repeats below.*

For each pair in the list:

- Go to that page & location in C-Book. If there is the name of an undergraduate there, and he or she isn't crossed off (because he/she is already in your sample) then add him/her to your sample, and cross him/her out in C-Book. Then start over with the next pair in the list.
- If you land on an ineligible location (already chose that person, or that person is a graduate student, or there are only 60 people on this page and your location is 83, etc.), *throw away that whole pair* and start over with the next pair in the list.

repeat this until you have 400 good names. *Note: you will probably need somewhat more than 400 pairs to get 400 good names.*

If I am listed in the C-Book the probability that you land on me with any given pair is:

$$P(\text{“page” is my page}) \cdot P(\text{“loc” is my location on page}) = \frac{1}{99} \cdot \frac{1}{110}.$$

This probability clearly doesn't depend on anything else, and it is the same for every person in the C-Book. So this generates a SRS of equally-likely potential respondents for your survey.

1.2.2 Random start, random skip

Random start, fixed skip. There are approximately 10,000 listings in C-Book, pp 10–108. Say you need 500 names (you really need 400 but you are building in some slop for landing on grad students, etc.).

$$10,000/500 = 20,$$

so you need every 20^{th} name. One way to do this is to pick a random number between 1 and 20. Mark that listing on p. 10, then every 20th listing in the book after that (going from one page or column to the next whenever you need to): if it is a valid undergrad listing, add it to your sample, and put a check on it so you know you've been there. If not cross it out in C-Book.

This will not generate a true SRS but it will generate a kind of random sample where almost everyone has an equally likely chance to get in the survey. You can treat your sample as an SRS anyway, it will behave very much like one.

Random start, random skip. A second method that is very similar, proceeds as follows:

- pick a random number between 1 and, say, 100, call it T. T is your skip number.
- now pick a random number between 1 and T, call it S. That is your start number.

Now start at the S^{th} listing, counting forward from the first listing on p. 10 of the C-Book, and mark that name. After that, mark every T^{th} name, just as in the "random start, fixed skip" method above.

If you didn't get enough names on one pass through the book, pick a new S and T (randomly!) and go through the book with those numbers, adding more names to your list.

Like the "random start, fixed skip" method, this doesn't quite generate a SRS , but the sample it does generate behaves a lot like an SRS, and you can treat it as an SRS.

1.2.3 Stratified samplers beware!

It is also possible to do stratified samples from C-Book, though it will probably take more than one pass through the book.

- If your strata are fr/so/jr/sr, one possibility would be to just do 4 SRS's from C-book, the first time treating only freshmen as eligible, the next time only sophomores, etc. Remember to oversample to account for low response rates (if you need 20 freshmen and you think your response rate is 25%, sample 100 freshmen!).
- Another possibility would be to use one of the previous methods to generate an SRS-like sample from C-book. Then see how well the SRS fills out the samples that you need in each stratum. Do additional more focused SRS's (freshmen, sophomores, jrs, seniors as in the first bullet) to fill out and strata for which you haven't met yout sample size quota yet.

If your stratifying variables are not listed in C-Book, then you will probably need a different sampling frame than C-Book to carry out a stratified sample.

2 Contacting Respondents

2.1 OK, I've got my sample. Now I just send everybody email asking them to go to surveymonkey[‡], right?

Sending email is a good first pass. Remember, though, that the response rates for email/web surveys are typically in the low 20% range, or worse. You will have to think of other ways to prod respondents as well.

Generally speaking, people do respond to other people, and they don't respond to machines. A 36-303 project from a couple of years ago found that people generally don't respond to "generic institutional" email (CMU spam), nor from senders they don't recognize, so you will have to work hard to get their attention. Here are my suggestions, based on my own experience as a writer and reader of email:

- The subject line matters. Too formal or too casual, and it won't get read. Brief subjects, with medium-to-low-frequency words that are relevant to your project, or to your appeal to the respondent, may work best.
- Email with impersonal greetings ("Dear student", "Hi there", "Dear CMU undergraduate", etc.) or no greeting at all, tends to get ignored.
- Email that is addressed to the person by first name ("Dear Vicki", "Dear Ming Mei", "Dear Sanjay", etc.) gets noticed more, and response rates are better. Even if you do no more than this (by hand or with an email-merge script in MS Outlook, or Perl or something) it will matter a lot.
- Email with a personalized body gets responded to better than email with a generic body.
- In my experience, a graceful personal touch matters even more for reminders and nags, than for first contacts.
- The second nag gets responded to better than the 4^{th} nag. I don't know whether the response rate on the 6^{th} nag is better or worse than that on the 4^{th} nag, but I suspect it's worse.

[‡]Surveymonkey is just an example in this document; you do not have to use Surveymonkey. There are other web services that provide the same features as surveymonkey, and Google Docs is able to do all this for free, if you are willing to work at it a little. *Whatever method of data collection you choose, you must be able to record and work with each individual respondent's answers to each individual question on your survey. If you only have summaries of the responses, you may not be able to do the analyses you need to get a good grade in this class.*

2.2 Well, how else should I contact them? I said in my proposal it was an email/surveymonkey survey…

You want to be thinking about alternative ways to contact people, on the first pass, as well as on later passes when you are trying to catch nonrespondents.

- If your primary contact method is telephone, think about email or one of the other methods below if you can't seem to get through on phone at first (invite them to call you, or ask whether there's a number at which you can call them, etc.)
- You may have promised me an email/surveymonkey survey, but the only part about that that I really care about is that everyone in your SRS responds on surveymonkey. It's fine with me if you find another way to contact them! Here are some ideas:
  - *Social Networking Sites.* A lot of people are on Facebook, and you probably have a substantial number of CMU students among your Facebook friends. And many people are sick of email and just respond better in social networking environments like Facebook or Google+. Once you have your SRS from C-Book, make a pass through Facebook or whatever else to see if some of your sample is active there. If they are, a personal appeal on Facebook may generate more positive responses than an email blast (Remember: People respond to other people [even on Facebook] not machines [spam email]).
  - *Telephone??* You have their numbers from C-Book. If they're not responding by email, maybe a phone call might do the trick. Call them up, ask them to please go to surveymonkey (or do the survey with them on the phone, get it out of the way for both of you!). People respond to people.
  - *Walking around campus??* Do you know someone in your SRS? Do they always pass the fence at 10:20 on Tuesdays? Catch them there, ask them to do your survey.
  - *Etc.* Be creative. How else can you contact someone in your SRS? But don't be so intrusive that you violate someone's privacy rights; and do back off if they say they're not interested. "No" means no.

3 Nonresponse Followup

3.1 Surveymonkey tools

Many of you are using a web service to present your survey questionnaire and collect responses. I have only looked at "surveymonkey.com" carefully so I will only speak about that here. But I imagine the same sorts of services are similarly priced elsewhere (and Google Docs is free, if you're willing to work to design an adequate survey form).

Surveymonkey offers several levels of user, particularly “Basic” and “Monthly Pro”.

- The “Basic” service is free and is enough to poke around and see what they’ve got. But it is limited to 100 respondents and 10 questions per survey. Pretty useless for your class project.
- The “Monthly Pro” service costs around \$20/mo and I think it can be renewed. If you time it right (to start soon after Spring Break I think) you will be able subscribe to it for just one month and get all of your data collection done on one \$20 fee, split among the members of your group. The monthly pro service offers you
    1. More help in writing questions and formatting your survey;
    2. 1000 respondents and unlimited questionnaire length;
    3. The ability to download raw survey response data (responses of <u>each individual</u> respondent to <u>each individual</u> survey question) in Excel or similar file format;
    4. The option to record respondent’s email address with every response;
    5. the option to redirect respondent to a website of your choosing after completing the survey.

Item 1, 2, and 3 are basic neccesities for conducting a survey for this class. You do not want to limit your questionnnaire or number of respondents unnecessarily. Also, you want to be able to look at the raw data in Excel, R, Minitab, etc. to produce the analyses you want to see, not the ones that Surveymonkey thinks are marketable.

Items 4 and 5 are key to nonresponse followup with Surveymonkey.

- If you are not asking any sensitive questions, I suggest you simply record each respondent’s email address with their response, and then you have a nice simple way of telling who has responded and who needs a reminder in email, Facebook, or whatever.
- If you *are* asking sensitive questions, then I suggest you make use of item 5. After the survey is done, direct the respondent to your own website. There, ask the respondent to identify him/herself by name, email address, or whatever you think is appropriate (e.g., by asking them to register for a small prize drawing, if you are using a prize increase response rates). Both in the informed consent paragraph before they begin on the surveymonkey survey, and here on your own website, explain that you have turned off respondent identification on surveymonkey, but that the respondent is being directed to this independent website so that you can keep track of who needs to be reminded to do the survey. It should be possible to explain all this in a way that doesn’t drive respondents away.

3.2 Following up with Nonrespondents

Please read Groves Ch 6 about nonresponse. There is good information there about unit nonresponse (missing a whole respondent) and item nonresponse (missing one or more questions for a respondent). Section 6.7 also talks about strategies for reducing unit nonresponse. My own guidelines here may be a bit thin.

There are at least two kinds of nonrespondents and they need to be handled differently:

1. People for whom you haven't had a confirmed contact. (e.g. they never answered the phone number you have for them, or you have the wrong email address or phone number for them)

2. People for whom you do have a confirmed contact, and either

   (a) they've refused to take the survey, or
   (b) they just haven't gotten around to it yet

You should exercise all due diligence in trying to get good contact information for, and contacting, people in category 1. But don't become obsessive. Some of these will be dead ends that you'll have to drop and move on to the next person in your list. Others you will eventually get through to. I suggest you keep notes on who you have or haven't been able to get through to, perhaps on the master list of your SRS, so that you can keep track of who you have and haven't tried to get contact info for (that way you are not skipping people you could be contacting, and you are not duplicating effort that you don't remember because you didn't write it down).

People in category 2(a) are done and out of your survey. "No" means no. You will have to be a little socially aware, since some people's version of "no" is just putting it off and putting it off, and others are too polite to come right out and say "forget about it." But with a little care, it should be fairly easy to distinguish the procrastinators from the ones who are too polite in their refusals.

People in category 2(b) are willing to take the survey and just haven't gotten around to it yet. It is OK to contact them to remind them to take the survey. Your contact (whether it is email, phone, Facebook or whatever) should

- be polite and friendly
- be as personalized as possible to increase the chances of a response (people respond to other people, not machines)
- offer them an opportunity to refuse, each time you contact them (perhaps as part of your "informed consent paragraph" if it is well written)
- not be so frequent as to be considered a constant nag

Your survey will probably be active for less than a month, so this will not go on forever anyway. Plan now for a schedule of reminders for people in category 2(b), over the active life of your survey.

It can be hard to distinguish category 2(a) people from 2(b) people, especially if your main line of contact is cold email that someone may or may not bother to read. Probably you will find some 2(a)'s in irate responses to any reminders you send out. Remember to be polite and friendly yourself. You are representing your team, this class, me as your instructor, and this University, which sanctions these student projects.

3.3 Nonresponders and Late Responders

As you are following up on the various nonresponders above, I suggest that you keep records of how and when each of your attempted contacts takes place, and what response you get.

One reason to do this is that you will learn more about how to interact with nonresponders, and likely you will learn how to turn some of them into responders. This is worth writing down and sharing (with other members of your team, or with the class in your oral progress reports), and/or saving for your next survey project.

Another reason to do this is that in writing up your results, you may wish to say something about what the nonresponders were like, or especially what they might have said on the survey if they had responded. For this, you may wish to take the data from a late responder, who only responded after several contacts, and think of it as similar to what a nonresponder who "looks like" that late responder would have said. By "looks like" I mean, similar demographics: gender, age, class, occupation/major, etc.