Some Statistical Sampling and Data Collection Activities

Andrew Gelma
Columbia University
Deborah Nolan
University of California, Berkeley

Abstract

This article presents several student participation activities combining (i) the basics of random sampling, (ii) practical complications (e.g., how do survey takers deal with biases from selection, nonresponse, and question wording), and (iii) theoretical ideas (e.g., sampling with unequal probabilities).

One way students learn about sampling is actually to collect some data. It gives them a feel for the practical struggles and small decisions needed in real data gathering, and it illustrates many key ideas in sampling. Another way is to discuss surveys reported in the popular press. It's fun to read and critique unusual news stories, and real survey findings bring home the importance of the statistical idea being illustrated. We have developed several demonstrations and examples of survey sampling to use in the classroom, which we use regularly in our introductory statistics courses for college students who have completed high school algebra. We have also had success introducing them to eighth graders in New York City public schools, and as these activities are in the spirit of the NCTM standards for statistics for grades 9-12, they would be particularly useful in the high school setting where statistics is widely taught.

First digits and Benford's law

It is fun and instructive for students to learn about random sampling by conducting simple classroom activities in which they use random numbers to sample from actual populations. We don't recommend assignments with personal interviewing—this is too hard, and the difficulty of contacting people, although providing an important lesson in itself, is a distraction from the mathematical concepts of sampling.

Here's a demonstration of how to take a random sample that involves the whole class. Also, this demo has the added bonus of illustrating Benford's Law (the digit 1 shows up as a leading digit for about 30% of the numbers). To introduce the demonstration, we raise the following questions: *What proportion of phone numbers have a 1 for the first digit of the four digit part of the number (e.g. 555-X212)?* and *What proportion of street numbers have a 1 for the leading digit?* Students (like most of us) expect that the digits, whether from a phone number or street address, should have roughly the same rate of occurrence. That is, we expect to see the digit 1 about 1/10 of the time for phone numbers and 1/9 of the time for street numbers (you can't have a leading digit of 0 for a street address). But they are surprised when they find out that this is not the case for street addresses.

We show the students a local telephone book, from which we rip out several pages that we have chosen at random. We divide the students into pairs and pass out one page from the telephone book to each pair and one twenty-sided die. The dice can be bought in a game store for about 50 cents apiece. For generating random numbers, we prefer using dice rather than computers (can be awkward to use in class) or random number tables (which are awkward).

We then explain our plan: we would like to sample several entries from the phone book. Each pair of students will sample ten listings at random from their page. Our phone book has five columns of 118 entries on each side of the page, so students need to figure out how to use rolls of the die to select an entry in a column on a side of the page. We want each of the 118*5*2 = 1180 entries to be equally likely. Figure 1 shows one way to sample the entries. Picking a column and a side of the page is easy: the twenty-sided die has the digits 0 to 9 written twice so we can simply roll the die and assign the numbers 0 through 9 to the 10 columns on the two-sided page.

Choosing a random number between 1 and 118 with equal probability is harder. The simplest method is to roll the die three times, where the first roll determines the digit in the hundreds place, the second digit is for the tens place, and the third digit gives the ones place. We would discard the number if it is not between 001 and 118, and repeat the procedure; i.e., roll the die three times to determine the hundreds, tens, and ones places, respectively. (Of course in practice, if the first roll is 2 or higher then we know our number will be too big, and we can discard the number before completing the final two rolls.) This procedure begins by generating numbers between 000 and 999 with the probability 1/1000, but because we ignore 000 and numbers 119-999 (and roll again), the probability that any number from 001 to 118 is chosen is 1/118.

FIGURE 1 GOES HERE

Figure 1 In this diagram of a page of the phone book, if the first of four rolls is 5 then an entry from the first column on the back of the page will be sampled. If the next three rolls are 0-1-8 then the 018 or 18th row of that column is sampled. If the three rolls land 0-0-0, 1-1-9, 1-2-0, ... 9-9-9 then discard these three rolls and roll three times again until a number between 001 and 118 is obtained.

There are many ways to select an entry at random. The students spend a few minutes coming up with a formal sampling plan. We have them work in groups to come up with ideas, and then we lead the class in a discussion on how to do it. (One procedure that is more efficient that the one described above, but still samples according to the equally likely principle, is to: roll the die once to determine the hundreds place (0 if even, 1 if odd); roll the die twice to pick the tens and units place; if the resulting number is not between 001 and 118 then discard it and begin again.) Once we agree on a protocol, the students start collecting data. They write the information for each entry on a sheet that we provide (Figure 2). We have them record both the phone number and street address.

Sample	Side	Column	Entry	Address	Telephone #
1	Back	5	009	2217 Ortis	xxx-0332
2	Front	1	101	1059 Peralta	xxx-5750
3	Back	2	037	2600 Havenscourt	xxx-6427
4	Back	4	068	2742 Prince	xxx-7626
5	Back	1	023	527 Woodmont	xxx-4986
6	Front	1	021	412 Kitty Hawk	xxx-1879

7	Front	3	085	1539 8th	xxx-9685
8	Front	2	053	1865 …	xxx-4xxx
9	Front	1	091	1151 …	xxx-5xxx
10	Back	4	099	705 …	xxx-9xxx

Figure 2 Example of a filled-out form for a pair of students sampling ten listings at random from a page (front and back) of the telephone book.

When the students are nearly finished collecting data, we ask if any difficulties arose. As is typical in real-life settings, the act of data collection reveals problems not anticipated in the planning stage. We raise questions such as, *What did you do if the entry was blank?* or *What if the listing took up two lines and you selected the line without the phone number?* For example, Sireesha Katragadda takes up lines 4 and 5 on the page displayed in Figure 3. If we record Katragadda's information whether line 4 or 5 is selected, then people with two-line listings would be twice as likely to be sampled. Instead, we should treat line 4 as a blank entry. When we get a blank entry, we must sample a new page, column, and entry in order that each listing has the same chance of being selected. (That probability will be 1/*the number of listings*, not 1/*the number of entries*.)

Questions of duplicate listings and unequal sampling probabilities lead naturally into a discussion of real-life difficulties in survey sampling. We ask the students to come up with other possible sources of bias in telephone sampling, such as households with multiple telephone numbers, unlisted numbers, nonresponse, and multiple people living in a household. (In Figure 3, Moshe Katvan has listings for three different phone numbers at the same address.)

FIGURE 3 GOES HERE

Figure 3 Section from the white pages of a telephone book. Blank lines and multi-line listings create potential for biases if the sampling design is not chosen carefully.

The data the students have collected from sampling the telephone book provide a fascinating example of Benford's Law (Browne, 1998, Hill, 1998 & 1999, Matthews, 1999). We collect the students' data (for example, we pass around a marker and transparency sheet on which space is allocated for tallying the first digits of the addresses and telephone numbers). Figure 4 shows a typical set of data: the first digits of the phone numbers are uniformly distributed, but the addresses are much more likely to begin with low digits.

According to Benford's law, the digit 1 occurs as a leading digit about 30% of the time, and the other digits are successively less frequent (17.8%, 12.5%, 9.7%, 7.9%, 6.7%, 5.8%, 5.1%, 4.8% for 2 through 9, respectively). These percentages can be explained by a uniform distribution on the logarithm of the number. If the numbers are equally likely on the $\log_{10}$ scale then the proportion of numbers with 1 as a leading digit is approximately $\log_{10}(2) - \log_{10}(1) = 0.301$. In general, the probability that the leading digit is i, for any i between 1 and 9, is $\log_{10}(i+1) - \log_{10}(i)$. This result connects nicely with the study of the log function.

How does this result apply to street addresses? If a street is only numbered up to 400, say, then the leading digit is most likely to be a 1, 2, or 3. More generally, there will be some streets numbered only up to 100 (in which case each digit will roughly be equally likely), some numbered up to 200 (and then about half of the addresses begin with a 1), some numbered up to 300, and so forth. Considering all of these, it makes sense that 1 is the most common leading digit, followed by 2, then 3, etc. Further, the Benford probabilities arise when we consider numbers (for example, prices) that grow by, say, 5% per year. It takes $\log_{10} (2) / \log_{10} (1.05) = 14$ years to reach the value 200 from a starting point of 100, and 47 years to reach the value of 1000. So the proportion of these values with first digits of 1, if selected at random, should be about 14/47 = 0.30.

FIGURE 4 GOES HERE

Figure 4 The counts for leading digits of street addresses (left) and phone numbers (right) for 80 randomly selected entries from a telephone book. The lines on the graphs show expected frequencies under Benford and uniform distributions, respectively.

Benford's Law is named for Dr. Frank Benford, a physicist at General Electric Company. In 1938, he noticed that in books of logarithms the pages of logarithms corresponding to numbers with a leading digit of 1 were more worn than other pages. (In fact, this phenomenon was first noticed by Simon Newcomb in 1881.) Benford concluded that the digit 1 is more likely to occur in scientific constants than the digits 2 through 9, and he found this also for many other types of data: areas of rivers, street addresses of people listed in the *American Men of Science*, county populations, and baseball statistics. Recently, Benford's law has been in the news as a method for checking fraud in financial data.

Wacky surveys

Ask the students in your class to raise their hands if they love statistics. The number who raise their hands, divided by the number of students in the room, is an estimate of the proportion of people with this opinion in the general population. Now give the students a minute to work in pairs and come up with as many sources of bias as they can think of for this example. These can be listed on the board and divided into categories: differences between sampled and target populations (students in a math or statistics course are not typical of all students, let alone all people, and, in addition, taking a statistics course may increase your love for the subject!), nonresponse bias (most students hesitate to raise their hands), and response bias (students may want to please the teacher and say Yes, or conversely they may be embarrassed and say No even if they do love the subject).

This mini-example can lead to a class discussion of biased survey questions: when do the news media or other organizations get misleading impressions because of question wording? How much of a difference can question wording actually make? We try to answer these questions and others with examples from the press.

We provide the template shown in Figure 5 to help students understand the sampling scheme and identify sources of bias. In particular, the template distinguishes between the target population (the group we want to study), the sampled population (the group from which the sample was taken), and the sampling frame (the list used to contact the sampled population, e.g., a list of phone numbers). The following news stories give examples of selection bias, measurement bias, questions bias, and nonresponse bias, and for each survey students fill in the template.

FIGURE 5 GOES HERE

Figure 5 Students fill in this template by identifying the target population, sampled population, and sampling frame. See Figure 6 for an example of a filled-out template. (This template is adopted from Lohr (1999).)

"1 in 4 youths abused, survey finds," Oct 4, 1994, *San Francisco Examiner*

A telephone survey of 2,000 children ages 10-16 found that 25% were slapped, punched, kicked, hit, or threatened with an object in the past year by an adult, sibling, or another child. Here the target population is all children ages 10 to 16 in the US; the sampling frame is all phone numbers in the US; and the sampled population is children living in a home (as opposed to an institution) with at least one phone number who would be at home the time of the call and whose guardian would consent to the interview.

Measurement bias is key to this survey. The definition of child abuse departs markedly from the more common definition used by the National Crime Survey. The authors of the survey chose a broader definition of abuse with the goal of raising public awareness about the violence to which children are exposed.

"Poll finds 1 out of 3 Americans open to doubt there was a holocaust," Apr 20, 1993, *Los Angeles Times*

The results from this Roper poll created a big stir. But a closer look at the question asked reveals a potential problem with question bias, *Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?*

The compound structure of the sentence and the use of the double negative makes the question confusing. Twenty two percent reported that it was possible that the holocaust didn't happen, and 12% didn't know. A year later, Roper repeated the survey, keeping all other questions the same and rewording this one, *Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?* This time only 1% reported that it was possible that it didn't happen, and 8% didn't know.

"11 million Internet addicts," Aug 13, 1999, *Associated Press*

The Associated press reported an online survey conducted by ABC News estimated there to be 11 million Americans suffering from "some form of addiction to the World Wide Web." This estimate is based on data collected from 17,251 responses to an Internet use questionnaire distributed and returned through the Web site ABCNEWS.com. Here the target population consists of persons who use the Web, and the sampling frame consists of visits to the ABC Web site. The sampled population are those Web users who visited the ABC News Web site during the time that the survey was on the net and chose to complete it. Note that there are problems with selection probabilities here because the more time someone is on the Web, the more likely he or she is to hit this particular site, and surfers who like to read news on the Web are more likely to find this site. We also do not know if someone can complete the survey multiple times.

FIGURE 6 GOES HERE

Figure 6 Example of a filled-out template for the survey of Web users described in "11 million Internet addicts," Aug 13, 1999, *Associated Press*.

Election exit poll

Election polling is a popular example for teaching sample surveys and provides many interesting examples of successes and failures in sampling. For exit polls, the target population is voters in the election, the sampling frame is the polling places, and the sampled population are those voters in the frame who voted in person (not absentee) and are willing to participate in the survey. One example of an exit poll we personally encountered (but, alas, too late to fix) was an exit poll in New York City, in which pollsters selected a sample of polling places and, at each, intercepted every fifth exiting voter. Those willing to participate were asked whom they voted for, demographics (sex, ethnicity, marital status, income, etc.), and various other questions (including political ideology, previous voting, and opinions about the news media). But a key piece of information was *not* gathered by the pollsters … (At this point, we divide the students into pairs and give them a minute to try to think of the missing information.) If none of the students figure it out, that's ok—neither did the political scientist who conducted the poll: the pollsters should have been instructed to gather information on the *nonrespondents*, to record the sex, ethnicity, and approximate age of the people who refused to cooperate. This would allow some judgment as to the scale of nonresponse bias.

"Kids choose reading over computer use, study finds," Nov 18, 1999, *San Francisco Chronicle*

And then there is just plain miscommunication of the facts. According to this news report, "Parents who worry that technology will turn their kids into a generation of illiterates can rest easy" because a study by the Kaiser Foundation found that children read more than they use computers. We put the statistics from the article on the board (Figure 7), and ask the students what's wrong. It seems to us that a more appropriate comparison would be to add the three computer related activities (49%) and that the comparison of time spent reading to time on the computer should be made for just those children who have computers in their homes.

Activity	Time
Watch TV	2:46
Listen to music	1:27
Read	0:44
Watch video	0:39
Computer email/fun	0:21
Play video games	0:20
Internet	0:08

Figure 7 Time spent (hours:minutes) on various activities by children ages 2 to 18 (Nov 18, 1999, *San Francisco Chronicle)*. The time spent reading is less than the total time spent on the computer or playing video games.

How large is your family?

After discussing sampling bias, with examples such as given above, we drive the point home with a demonstration. Each student is asked to tell how many children are in his or her family ("How any brothers and sisters are in your family, including yourself?"). We write the results on the blackboard as a frequency table and a histogram, and then compute the mean, which is typically around 3 (see Figure 8 for an example).

Family size (# of siblings including self)	Count
1	3
2	3
3	5
4	5
5	3
6 or more	0

—

BAR CHART GOES HERE

Figure 8 Data on number of children in families (displayed as frequency table and histogram) from students in a typical class. The average is 3.1, which is at first a surprise, given that the average family has about 2 children.

We tell the students that the average number of children in families that were having children 20 years ago (about the age of the students in the class) was about 2.0. Why is the number for this class so high? Students give various suggestions such as, perhaps larger families are more likely to send children to this school, or to take this math class. After some discussion, a student notes that if the family had zero children, they certainly did not send any to school. The 2.0 figure is the average number of children when sampling *by family*; 3.0 is the average number of children when sampling *by child*. When sampling by child, a family with 4 children is 4 times more likely to be sampled than a family with 1 child. This illustrates the general point that it is not enough to say you sampled at random; you must also know the method of sampling. It can also be considered as an example of *sampling bias*.

At this point, we can get the students further involved by asking the question, how can data be gathered to estimate the average number of children per family? In discussing the problem, students can consider the relative difficulties of correcting the data on students for sampling bias, compared to the direct approach of sampling families. The former approach is tricky because it still requires some estimate of the proportion of families with zero children. The students also have to realize that, for either approach, a careful definition of "family" is required. We sometimes ask the students how many children their oldest aunt or uncle has. Leaving out those students who have no aunt or uncle, the average number of children is much smaller (in one class survey we found it to be about 2.2).

Related issues arise in telephone sampling, when you call a telephone number at random and then pick a person at random in that household to interview: Is a person with many phone lines more or less likely to be sampled than a person with one line? Is a person with many roommates more or less likely to be sampled than a person living alone?

Summary

Students enjoy discovering the basics of sampling, the practical complications, and the theoretical concepts through these activities.

It is always fun to illustrate lecture points with examples of studies reported in the popular press, but we have found that in class discussion, it's important to be critical without being dismissive of the survey findings. We lay out possible sources of bias while at the same time recognizing that these results might be useful as part of a comprehensive perspective on the problem. The template is a great aid for figuring out complications with a survey.

The students find intriguing the unexpected twists in our activities (Benford's law and family sizes). These activities push the students to think more deeply about the problem, where for example they discover that it is not enough to say you sampled at random; you must also know the sampling units. On the practical side, students find creating random numbers using dice more compelling than computers or graphing calculators.

Our demonstrations and activities are in the spirit of the NCTM standards for students in grades 9-12, where an "understanding of statistics and probability could provide them with ways to think about a wide range of issues that have important social implications…" In particular, the random sampling demonstration helps students "know characteristics of well-defined studies, including the role of randomization in surveys," and our discussions of wacky surveys helps students learn how to "evaluate published reports that are based on data by examining the design of the study, the appropriateness of the data analysis, and the validity of conclusions."

References

Browne, Malcolm W. (1998) Following Benford's Law, or Looking Out for No. 1*, New York Times*, Aug 4.

Hill, Ted P. (1998) The First-Digit Phenomenon, *American Scientist* **86**(4):358-363.

Hill, Ted P. (1999) The difficulty of faking data, *Chance* **12**(3):27-31.

Lohr, Sharon (1999) *Sampling: Design and Analysis*. Duxbury Press: Pacific Grove, CA.

Matthews, Robert (1999) The power of one, *New Scientist*, 10 July, 26-30.