MATH 3070 Introduction to Probability and Statistics
Lecture notes
Sampling theory

Objectives:

1. Learn the basic concepts of sampling theory
2. Learn what bias is and why it is bad
3. Learn the difference between
   - voluntary response,
   - convenience,
   - systematic, and
   - simple random samples
4. Create a simple random sample with a random number table
5. Issues with sampling (undercoverage, nonresponse)

Why do we sample?

To do statistics one needs data. We cant use the population because usually its too large. We cant get our hands on it all, at least not cost effectively. So we use a sample, a smaller subset of the population. But we cant use just any sample, we must choose correctly.

Choosing correctly means avoiding bias. Since we are making decisions about the population from our sample, if the sample is biased, or unrepresentative, then the conclusions based upon it are flawed.

Sometimes we can collect our own data. This is called **primary data**. We have ultimate control over it. But sometimes we cant afford the time or money to collect data so we rely on someone else, like the government. This is called **secondary data**. We have to trust that the data were collected in an unbiased manner.

It would be nice to have access to all the data we need to make a decision or a prediction. The hard fact is, though, we don't always have all the data. Sometimes we can't hope to have all the data if the data set is quite large (like the US population) or constantly arriving (like a manufacturing process). In that case we must use a smaller portion of the whole data and generalize back to the population at large. This smaller portion is called a **sample**.

There are several ways to select samples. The easiest are sometimes the most prone to bias. Creating a sample is relatively easy, but there are rules to be followed if we want a useful sample and a useful conclusion.

What are we doing?

The data we are sampling can either be from an **observational study** or from an **experiment**. An observational study does not attempt to influence the phenomenon being observed. It can be a survey which only collects opinions and data. An experiment attempts to influence the outcome of the phenomenon being studied, usually to determine which of several approaches is the best solution (if any).

Various sample designs

What we want to avoid is **bias**. Biased samples produced biased conclusions, and this does not serve the purpose of sampling and generalizing of conclusions. It can, but then the outcome was already known and the data collected were just being used to support a predefined conclusion. We want a truly representative sample of the population so we know what is true in general.

Samples must be designed and are not actually arbitrary. Have you noticed that the elections are often predicted with a rather small (1,500) number of registered voters? This isn't just asking the next 1,500 people that walk in the door.

That is, however, a method of sampling. **Voluntary reponse** is a method whereby the sample "self selects". People choose to respond or participate based on a general appeal. This is what happens on call-in radio and television programs. This is a bad method, if cheap and easy, because the participants usually are those people with strong opinions and very often negative opinions to the question at hand. Why do you think this is?

Another cheap and easy sampling method is the **convenience sample**. It's name tells you about it. In this method, the sample is composed of those people easiest to reach. This is probably not a very good representation of the population in question, unless the pollster has a very large group of friends and colleagues.

This can be useful, though, in a test situation. When developing a new survey tool it must be actually taken to identify errors and confusing statements. It would be best if the group used to identify these items is not the real group being surveyed. Surveys cost money to deploy and collect, so you want to make sure that data being returned is good. A convenience sample can help in the "beta testing" of a survey tool. College students (and co-workers) make excellent survey testers.

The easiest (good) sample design is a **simple random sample**. This is a sample where every member of the population had an equal chance at being selected. This is true no matter how many samples are selected from the population. Repeated samples should each be equally representative of the population.

Using random numbers to generate a sample

How do we generate a simple random sample? Randomly, of course. We user random numbers to select who might be asked to participate in the survey or study. Random numbers come from a variety of sources such as a roll of a die, a flip of a coin, or a random number generator on a computer. A nontechnological method is to use a table of random numbers, one of which appears in every statistics textbook.

So how do we use a table of random numbers? First we assign each observation in the dataset a number. This number can start at zero or at one, or anywhere else for that matter. The number of digits in the number should be padded with leading zeros so that all the assigned numbers have the same length. For example, if there are ten to ninety-nine observations the numbers assigned should be $01, 02, 03, 04, \ldots, 98, 99$. If there are one-hundred to nine-hundred ninety-nine then the assigned numbers should be three digits. And so forth.

In this example, the total number of observations would be 50 which would include one observation for each state. The assigned numbers should then be two digits in length.

01	7263	16347	Alabama
02	12441	28395	Alaska
03	9185	19017	Arizona
04	6271	14641	Arkansas
05	10729	21537	California
06	9552	21279	Colorado
⋮	⋮	⋮	⋮

Now, choose a row in the table of random numbers. Since any row should give as good a sample as any other, the choice of row does not matter. Starting at the left end of the row, choose numbers of the same size as your assigned labels. In the above example, if our table row was

$$07511889154126716853845697936732337033 16$$

then our two digit random numbers would be

$$07, 51, 18, 89, 15, 41, 26, 71, 68, 53, \ldots$$

Notice that not all the numbers selected from the row are usable. We don't have 89 states so there isn't an observation with 89 as it's assigned number. We discard that number from the row and continue. The same action is taken when we encounter a number we've already used. When we get to the end of the row we continue on the next row until we have enough observations in our sample.

How many is enough? It depends on the size of the population you are trying to generalize about. There are calculations depending upon the statistical test you want to run that will help you decide the right size for your sample. It isn't by chance that we use a sample of 1,500 registered voters to predict the national elections.

So if we were to want a sample of five states from the fifty states and District of Columbia (a population of 51) we could use the SRS methodology as follows:

1. Assign each observation (state) a two digit number beginning at 01 and ending with 51. (See above for an example.)
2. Choose a row in the random number table.

00	15544	80712	97742	21500	97081	42451	50623	56071	28882	28739
01	01011	21285	04729	39986	73150	31548	30168	76189	56996	19210

3. Beginning at the first digit on line 00, begin selecting two digit numbers. When a gap intervenes, consider the two digits on either side as one two digit number. The numbers from the first row are

15	54	48	07	12	97	74	22	15	00
97	08	14	24	51	50	62	35	60	71
28	88	22	87	39					

4. Only those numbers which fall within the range 01 to 51, inclusive, are useful in selecting our sample. Those numbers are

15	54	**48**	**07**	**12**	97	74	**22**	**15**	00
97	**08**	**14**	**24**	**51**	**50**	62	35	60	71
28	88	**22**	87	**39**					

This line of the table appears to be rich in values that meet our needs so we don't need to use any further rows. The values that we would select for our random sample of five would be

$$15, 48, 07, 12, 22$$

Note that the next number in the sequence was another 15. Since we already have that number in our sample we would discard the duplicate and continue on if we still needed numbers. So our sample consists of

15	Indiana	20535
48	Washington	20018
07	Connecticut	23149
12	Hawaii	22750
22	Massachusetts	21166

Other sampling methods

There are two other sampling methods that are routinely used. **Systematic sampling** selects every n^{th} observation, such as "every third car" or "every fifth student." Sometimes the choice of n is based upon the population size divided by the desired sample size ($\frac{N}{n}$). Rounding this value to the nearest whole number, most likely the largest integer not greater than, or $floor$, will give you the choice of which observations to select.

The other method is **stratified random sampling**. Sometimes the population naturally separates into groups, called **strata**. A simple example of this would be gender. Any population of people can be divided by gender into two relatively homogeneous group. If there are discernible groups within the population it is advisable that samples be taken from each of the subgroups to insure that the whole sample is representative of the population.

Another example of this would be voters by county. Especially in Georgia, a majority of the population resides in the urban areas. In this case, predominantly in the Atlanta MSA (roughly 4,768,685 out of 9,072,576 people in Georgia, 52.5%). If you take a random sample of voters in Georgia the majority of them will naturally reside in the Atlanta MSA. This may not give a truly representative sample and may not give an accurate response to your question. It would be better to group the voters by county and then take simple random samples from each of the groups.