STA 215: Sampling Methods

Al Nosedal.
University of Toronto.

Summer 2019

My momma always said: "Life was like a box of chocolates. You never know what you're gonna get."

Forrest Gump.

Population, Sample, Sampling Design

The **population** in a statistical study is the entire group of individuals about which we want information.
A **sample** is a part of the population from which we actually collect information. We use a sample to draw conclusions about the entire population.
A **sampling design** describes exactly how to choose a sample from the population.
The first step in planning a **sample survey** is to say exactly what population we want to describe. The second step is to say exactly what we want to measure, that is, to give exact definitions of our variables.

Customer satisfaction

A department store mails a customer satisfaction survey to people who make credit card purchases at the store. This month, 45,000 people made credit card purchases. Surveys are mailed to 1000 of these people, chosen at random, and 137 people return the survey form.

a) What is the population of interest for this survey?

b) What is the sample? From what group is information actually obtained?

Solutions

a) The population is all 45,000 people who made credit card purchases.
b) The sample is the 137 people who returned the survey form.

How to sample badly

The final step in planning a sample survey is the sampling design. A sampling design is a specific method for choosing a sample from the population. The easiest- but not the best - design just chooses individuals close at hand. A sample selected by taking the members of the population that are easiest to reach is called a **convenience sample.** Convenience samples often produce unrepresentative data.

Bias

The design of a statistical study is **biased** if it systematically favors certain outcomes.

Voluntary response sample

A **voluntary response sample** consists of people who choose themselves by responding to a broad appeal. Voluntary response samples are biased because people with strong opinions are most likely to respond.

Sampling on campus

You see a woman student standing in front of the student center, now and then stopping other students to ask them questions. She says that she is collecting student opinions for a class assignment. Explain why this sampling method is almost certainly biased.

Solution

It is a convenience sample; she is only getting opinions from students who are at the student center at a certain time of day. This might underrepresent some group: commuters, graduate students, or nontraditional students, for example.

Simple Random Sampling

A **simple random sample (SRS)** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

Random digits

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these two properties:
1. Each entry in the table is equally likely to be any of the 10 digits 0 through 9.
2. The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

Using Table B to choose an SRS

Label: Give each member of the population a numerical label of the same length.
Table: To choose a SRS, read from Table B successive groups of digits of the length you used as labels. Your sample contains the individuals whose labels you find in the table.

Apartment living

You are planning a report on apartment living in a college town. You decide to select three apartment complexes at random for in-depth interviews with residents. Use software or Table B to select a simple random sample of 4 of the following apartment complexes. If you use Table B, start at line 122.

Ashley Oaks	Country View	Mayfair Village
Bay Pointe	Country Villa	Nobb Hill
Beau Jardin	Crestview	Pemberly Courts
Bluffs	Del-Lynn	Peppermill
Brandon Place	Fairington	Pheasant Run
Briarwood	Fairway Knolls	River Walk
Brownstone	Fowler	Sagamore Ridge
Burberry Place	Franklin Park	Salem Courthouse
Cambridge	Georgetown	Village Square
Chauncey Village	Greenacres	Waterford Court

Solution

01	Ashley Oaks	11	Country View	21	Mayfair Village
02	Bay Pointe	12	Country Villa	22	Nobb Hill
03	Beau Jardin	13	Crestview	23	Pemberly Courts
04	Bluffs	14	Del-Lynn	24	Peppermill
05	Brandon Place	15	Fairington	25	Pheasant Run
06	Briarwood	16	Fairway Knolls	26	River Walk
07	Brownstone	17	Fowler	27	Sagamore Ridge
08	Burberry Place	18	Franklin Park	28	Salem Courthouse
09	Cambridge	19	Georgetown	29	Village Square
10	Chauncey Village	20	Greenacres	30	Waterford Court

Solution (cont.)

With Table B, enter at line 122 and choose 13 = Crestview, 15 = Fairington, 05 = Brandon Place, and 29 = Village Square.

R Code

```
set.seed(2016);
# Use to reproduce the sample below;

sample(1:30,4);
# 2nd number represents sample size;
```

R Code

```
## [1]  6  5 24  4
```

Inference about the Population

The purpose of a sample is to give us information about a larger population. The process of drawing conclusions about a population on the basis of sample data is called **inference** because we infer information about the population from what we know about the sample.

Inference from convenience samples or voluntary response samples would be misleading because these methods of choosing a sample are biased. We are almost certain that the sample does not fairly represent the population. **The first reason to rely on random sampling is to eliminate bias in selecting samples from the list of available individuals.**

Sampling frame

The list of individuals from which a sample is actually selected is called the **sampling frame**. Ideally, the frame should list every individual in the population, but in practice this is often difficult. A frame that leaves out part of the population is a common source of undercoverage.

Suppose that a sample of households in a community is selected at random from the telephone directory. What households are omitted from this frame? What types of people do you think are likely to live in these households? These people will probably be underrepresented in the sample.

Solution

This design would omit households without telephones or with unlisted numbers. Such households would likely be made up of poor individuals (who cannot afford a phone), those who choose not to have phones, and those who do not wish to have their phone number published.

Cautions about sample surveys

Random selection eliminates bias in the choice of a sample from a list of the population. When the population consists of human beings, however, accurate information from a sample requires more than a good sampling design.

To begin, we need an accurate and complete list of the population. Because such a list is rarely available, most samples suffer from some degree of **undercoverage**. A sample survey of households, for example, will miss not only homeless people but prison inmates and students in dormitories. An opinion poll conducted by calling landline telephone numbers will miss households that have only cell phones as well as households without a phone. The results of national sample surveys therefore have some bias if the people not covered differ from the rest of the population.

A more serious source of bias in most sample surveys is **nonresponse**, which occurs when a selected individual cannot be contacted or refuses to cooperate.

Nonresponse

Academic sample surveys, unlike commercial polls, often discuss nonresponse. A survey of drivers began by randomly sampling all listed residential telephone numbers in the United States.
Of 45,956 calls to these numbers, 5029 were completed. What was the rate of nonresponse for this sample? (Only one call was made to each number. Nonresponse would be lower if more calls were made.)

Solution

The response rate was $\frac{5029}{45956} = 0.1094$, so the nonresponse rate was $1 - 0.1094 = 0.8906$

Undercoverage and nonresponse

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.
Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to participate.

Stratified Random Sample

To select a **stratified random sample**, first divide the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Toy Example

Suppose we have a population of size 5. We measure a variable for each of these 6 individuals, the result of our measurements follows: $50, 55, 60, 70$, 80, and 87. Now, we compute the population mean, which we denote by μ, $\mu = 67$.

Toy example

Let us see what happens if we use a simple random sample of size 2 to compute the population mean, μ.

Sample	Measurements	$\bar{x}_i$	Sample	Measurements	$\bar{x}_i$
1	(50, 55)	52.5	9	(55, 87)	71
2	(50, 60)	55	10	(60, 70)	65
3	(50, 70)	60	11	(60, 80)	70
4	(50, 80)	65	12	(60, 87)	73.5
5	(50, 87)	68.5	13	(70, 80)	75
6	(55, 60)	57.5	14	(70, 87)	78.5
7	(55, 70)	62.5	15	(80, 87)	83.5
8	(55, 80)	67.5			

Toy example (errors)

	Measurements	$\bar{x}_i - \mu$		Measurements	$\bar{x}_i - \mu$
1	(50, 55)	-14.5	9	(55, 87)	4
2	(50, 60)	-12	10	(60, 70)	-2
3	(50, 70)	-7	11	(60, 80)	3
4	(50, 80)	-2	12	(60, 87)	6.5
5	(50, 87)	1.5	13	(70, 80)	8
6	(55, 60)	-9.5	14	(70, 87)	11.5
7	(55, 70)	-4.5	15	(80, 87)	16.5
8	(55, 80)	0.5			

Note. The average error is zero!

Toy example

Let us see what happens if we use a stratified random sample of size 2 to compute the population mean, μ. Assume that the stratum 1 is formed by:$\{50, 55, 60\}$ and stratum 2 by: $\{70, 80, 87\}$.

Sample	Measurements	$\bar{x}_i$
1	(50, 70)	60
2	(50, 80)	65
3	(50, 87)	68.5
4	(55, 70)	62.5
5	(55, 80)	67.5
6	(55, 87)	71
7	(60, 70)	65
8	(60, 80)	70
9	(60, 87)	73.5

Toy example (errors)

Sample	Measurements	$\bar{x}_i - \mu$
1	(50, 70)	-7
2	(50, 80)	-2
3	(50, 87)	1.5
4	(55, 70)	-4.5
5	(55, 80)	0.5
6	(55, 87)	4
7	(60, 70)	-2
8	(60, 80)	3
9	(60, 87)	6.5

Note. The average error is zero!