Sampling and Data: Homework*

Susan Dean
Barbara Illowsky, Ph.D.

This work is produced by The Connexions Project and licensed under the
Creative Commons Attribution License †

Abstract

This module presents students with a number of problems related to statistical sampling and data. In particular, students are asked to demonstrate understanding of concepts such as frequency, relative frequency, and cumulative relative frequency, random samples, quantitative vs. qualitative data, continuous vs. discrete data, and other key terms related to sampling and data.

Exercise 1 ***(Solution on p. 10.)***

For each item below:

i. Identify the type of data (quantitative - discrete, quantitative - continuous, or qualitative) that would be used to describe a response.
ii. Give an example of the data.

a. Number of tickets sold to a concert
b. Amount of body fat
c. Favorite baseball team
d. Time in line to buy groceries
e. Number of students enrolled at Evergreen Valley College
f. Most–watched television show
g. Brand of toothpaste
h. Distance to the closest movie theatre
i. Age of executives in Fortune 500 companies
j. Number of competing computer spreadsheet software packages

Exercise 2

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

Part-time Student Course Loads

*Version 1.19: Jun 30, 2012 10:36 am -0500
†http://creativecommons.org/licenses/by/3.0/

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

Table 1

a. Fill in the blanks in the table above.
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

Exercise 3 ***(Solution on p. 10.)***
Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnoses. The (incomplete) results are shown below:

Flossing Frequency for Adults with Gum Disease

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

Table 2

a. Fill in the blanks in the table above.
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

Exercise 4
A fitness center is interested in the mean amount of time a client exercises in the center each week. Define the following in terms of the study. Give examples where appropriate.

a. Population
b. Sample
c. Parameter
d. Statistic
e. Variable
f. Data

Exercise 5 ***(Solution on p. 10.)***
Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to optimally plan their ski classes. Define the following in terms of the study. Give examples where appropriate.

a. Population
b. Sample
c. Parameter
d. Statistic
e. Variable
f. Data

Exercise 6
A cardiologist is interested in the mean recovery period for her patients who have had heart attacks. Define the following in terms of the study. Give examples where appropriate.

a. Population
b. Sample
c. Parameter
d. Statistic
e. Variable
f. Data

Exercise 7 ***(Solution on p. 10.)***
Insurance companies are interested in the mean health costs each year for their clients, so that they can determine the costs of health insurance. Define the following in terms of the study. Give examples where appropriate.

a. Population
b. Sample
c. Parameter
d. Statistic
e. Variable
f. Data

Exercise 8
A politician is interested in the proportion of voters in his district that think he is doing a good job. Define the following in terms of the study. Give examples where appropriate.

a. Population
b. Sample
c. Parameter
d. Statistic
e. Variable
f. Data

Exercise 9 ***(Solution on p. 10.)***
A marriage counselor is interested in the proportion the clients she counsels that stay married. Define the following in terms of the study. Give examples where appropriate.

a. Population
b. Sample
c. Parameter
d. Statistic
e. Variable
f. Data

Exercise 10

Political pollsters may be interested in the proportion of people that will vote for a particular cause. Define the following in terms of the study. Give examples where appropriate.

a. Population
b. Sample
c. Parameter
d. Statistic
e. Variable
f. Data

Exercise 11 ***(Solution on p. 10.)***

A marketing company is interested in the proportion of people that will buy a particular product. Define the following in terms of the study. Give examples where appropriate.

a. Population
b. Sample
c. Parameter
d. Statistic
e. Variable
f. Data

Exercise 12

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys 6 flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

a. Using complete sentences, list three things wrong with the way the survey was conducted.
b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

Exercise 13

Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in $3 - 5$ complete sentences. Make the description detailed.

Exercise 14

Suppose you want to determine the mean number of cans of soda drunk each month by persons in their twenties. Describe a possible sampling method in 3 - 5 complete sentences. Make the description detailed.

Exercise 15 ***(Solution on p. 11.)***

771 distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed in the table below. (Source: http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus).

LBCC Distance Learning Survey Results

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

Table 3

a. What percent of the students surveyed do not have a computer at home?
b. About how many students in the survey live at least 16 miles from campus?
c. If the same survey was done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

Exercise 16

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows:

2; 5; 7; 2; 2; 10; 20; 15; 0; 7; 0; 20; 5; 12; 15; 12; 4; 5; 10

The following table was produced:

Frequency of Immigrant Survey Responses

Data	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0	2	$\frac{2}{19}$	0.1053
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

Table 4

a. Fix the errors on the table. Also, explain how someone might have arrived at the incorrect number(s).
b. Explain what is wrong with this statement: "47 percent of the people surveyed have lived in the U.S. for 5 years."
c. Fix the statement above to make it correct.

d. What fraction of the people surveyed have lived in the U.S. 5 or 7 years?
e. What fraction of the people surveyed have lived in the U.S. at most 12 years?
f. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
g. What fraction of the people surveyed have lived in the U.S. from 5 to 20 years, inclusive?

Exercise 17
A "random survey" was conducted of 3274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users. (*Source: San Jose Mercury News*)

a. Do you consider the sample size large enough for a study of this type? Why or why not?
b. Based on your "gut feeling," do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?

Additional information: The survey was reported by Intel Corporation of individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."

c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

Exercise 18

a. List some practical difficulties involved in getting accurate results from a telephone survey.
b. List some practical difficulties involved in getting accurate results from a mailed survey.
c. With your classmates, brainstorm some ways to overcome these problems if you needed to conduct a phone or mail survey.

1 Try these multiple choice questions

The next four questions refer to the following: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

Exercise 19 ***(Solution on p. 11.)***
What is the population she is interested in?

A. All Lake Tahoe Community College students
B. All Lake Tahoe Community College English students
C. All Lake Tahoe Community College students in her classes
D. All Lake Tahoe Community College math students

Exercise 20 ***(Solution on p. 11.)***
Consider the following:
X = **number of days a Lake Tahoe Community College math student is absent**
In this case, X is an example of a:

A. Variable
B. Population
C. Statistic
D. Data

Exercise 21 *(Solution on p. 11.)*
The instructor takes her sample by gathering data on 5 randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

A. Cluster sampling
B. Stratified sampling
C. Simple random sampling
D. Convenience sampling

Exercise 22 *(Solution on p. 11.)*
The instructor's sample produces an mean number of days absent of 3.5 days. This value is an example of a

A. Parameter
B. Data
C. Statistic
D. Variable

The next two questions refer to the following relative frequency table on hurricanes that have made direct hits on the U.S between 1851 and 2004. Hurricanes are given a strength category rating based on the minimum wind speed generated by the storm. (*http://www.nhc.noaa.gov/gifs/table5.gif*[1])

Frequency of Hurricane Direct Hits

Category	**Number of Direct Hits**	**Relative Frequency**	**Cumulative Frequency**
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Table 5

Exercise 23 *(Solution on p. 11.)*
What is the relative frequency of direct hits that were category 4 hurricanes?

A. 0.0768
B. 0.0659
C. 0.2601
D. Not enough information to calculate

[1] http://www.nhc.noaa.gov/gifs/table5.gif

Exercise 24 *(Solution on p. 11.)*

What is the relative frequency of direct hits that were AT MOST a category 3 storm?

A. 0.3480
B. 0.9231
C. 0.2601
D. 0.3370

The next three questions refer to the following: A study was done to determine the age, number of times per week and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

Exercise 25 *(Solution on p. 11.)*

"'Number of times per week"' is what type of data?

A. qualitative
B. quantitative - discrete
C. quantitative - continuous

Exercise 26 *(Solution on p. 11.)*

The sampling method was:

A. simple random
B. systematic
C. stratified
D. cluster

Exercise 27 *(Solution on p. 11.)*

"'Duration (amount of time)"' is what type of data?

A. qualitative
B. quantitative - discrete
C. quantitative - continuous

Exercises 28 and 29 are not multiple choice exercises.

Exercise 28 *(Solution on p. 11.)*

Name the sampling method used in each of the following situations:

A. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers sitting near gates and who are not taking naps while they wait.
B. A teacher wants to know if her students are doing homework so she randomly selects rows 2 and 5, and then calls on all students in row 2 and all students in row 5 to present the solution to homework problems to the class.
C. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out which asks for information about age, as well as about other variables of interest.
D. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether the books are checked out by an adult or a child. She records this data for every 4th patron who checks out books.

E. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked who he/she intends to vote for and whether the debate changed his/her opinion of the candidates.

** Contributed by Roberta Bloom

Exercise 29 ***(Solution on p. 11.)***

Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following 7 subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these 7 textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

** Contributed by Roberta Bloom

Solutions to Exercises in this Module

Solution to Exercise (p. 1)

a. quantitative - discrete
b. quantitative - continuous
c. qualitative
d. quantitative - continuous
e. quantitative - discrete
f. qualitative
g. qualitative
h. quantitative - continuous
i. quantitative - continuous
j. quantitative - discrete

Solution to Exercise (p. 2)

a. Cum. Rel. Freq. for 0 is 0.4500
Rel. Freq. for 1 is 0.3000 and Cum. Rel. Freq. for 1 or less is 0.7500
Freq. for 3 is 11 and Rel. Freq. is 0.1833
Cum. Rel. Freq. for 6 or less is 0.9833
Cum. Rel. Freq. for 7 or less is 1
b. 5.00%
c. 93.33%

Solution to Exercise (p. 2)

a. Children who take ski or snowboard lessons
b. A group of these children
c. The population mean
d. The sample mean
e. X = the age of one child who takes the first ski or snowboard lesson
f. Values for X, such as 3, 7, etc.

Solution to Exercise (p. 3)

a. The clients of the insurance companies
b. A group of the clients
c. The mean health costs of the clients
d. The mean health costs of the sample
e. X = the health costs of one client
f. Values for X, such as 34, 9, 82, etc.

Solution to Exercise (p. 3)

a. All the clients of the counselor
b. A group of the clients
c. The proportion of all her clients who stay married
d. The proportion of the sample who stay married
e. X = the number of couples who stay married
f. yes, no

Solution to Exercise (p. 4)

a. All people (maybe in a certain geographic area, such as the United States)

b. A group of the people
c. The proportion of all people who will buy the product
d. The proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

Solution to Exercise (p. 4)

a: 4%
b: 100

Solution to Exercise (p. 6)
D
Solution to Exercise (p. 6)
A
Solution to Exercise (p. 7)
B
Solution to Exercise (p. 7)
C
Solution to Exercise (p. 7)
B
Solution to Exercise (p. 8)
B
Solution to Exercise (p. 8)
B
Solution to Exercise (p. 8)
B
Solution to Exercise (p. 8)
C
Solution to Exercise (p. 8)

A. Convenience
B. Cluster
C. Stratified
D. Systematic
E. Simple Random

Solution to Exercise (p. 9)
The answer below contains some of the issues that students might discuss for this problem. Individual student's answers may also identify other issues that pertain to this problem that are not included in the answer below.

The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled 7 subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The 7 subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and many other subject areas, (for example: literature, art, history, psychology, sociology, business) that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.

He also only looked at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- the most popular textbooks may be more readily available online, because more new copies are printed and more students nationwide selling back their used copies OR
- the most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks.

He could improve this study by

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the popular and less popular textbooks.