Lesson 1
Data, Samples, and Populations

Outline of the Lesson

Several years ago, as was his custom, a professor teaching a statistics course administered a short survey on the first day of class. The survey contained these twelve questions:

1. What is your gender (Male, Female)? _____
2. What is your class year? (Freshman, Sophomore, Junior, Senior, Other) _____
3. How many states have you visited (or lived in or even just driven through)?
4. Have you ever been a smoker (at least ½ pack a day)? (Yes/No) _____
5. How would you rate yourself politically? (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative)
6. What is your height (in inches)? _____
7. How many times a week (on average) do you read a daily newspaper? _____
8. What is your political affiliation? (D = Democrat, R = Republican, I = Independent) _____
9. Do you have a paying job during the school year at which you work on average at least 10 hours a week? (Yes/No) _____
10. How many minutes a day (on average) do you watch t.v.?_____
11. What is the distance (in miles) between your home and this campus? _____
12. Aside from class time, how many hours a week, on average, do you expect to spend studying and completing assignments for this course? _____

The survey provided information, in the form of *data*, about the students in the class. In that particular class, the results of the survey were tabulated as shown here:

Gender	Class Yr	States	Smoke	Politics	Height	Newspaper	Political Party	Job	TV	Commute	Course Time
Male	Sophomore	31	Yes	2	72	2	D	Yes	120	0.5	1.5
Female	Sophomore	19	No	4	71	1	R	Yes	30	0	5
Male	Junior	16	No	2	70	1	D	No	60	0.3	7.5
Female	Junior	24	No		62	1	R	Yes	3	0.5	5
Female	Sophomore	43	No	7	63	4	R	Yes	120	1.5	8
Female	Junior	22	Yes	5	65	0	R	Yes	0	24	10
Male	Sophomore	17	No	2	60	3	D	No	90	1.5	3
Male	Sophomore	10	Yes	3	57	1	I	No	90	1.5	3
Female	Freshman	11	No	3	67	0	D	Yes	90	0	8
Female	Senior	16	No	5	60	0	R	Yes	120	0.5	3
Female	Freshman	8	No	2	66	0	D	Yes	180	6	7
Male	Sophomore	14	No	4	73	1	R	No	60	0	5
Female	Sophomore	33	Yes	4	60	0	R	Yes	60	1	6
Female	Sophomore	16	Yes	4	64.5	0	R	No	120	0.2	7
Female	Sophomore	13	Yes	6	68	2	R	No	60	1	5
Male	Sophomore	26	No	6	73	0	I	Yes	30	20	3
Female	Junior	11	No	6	64	0	R	Yes	60	0.5	5
Male	Sophomore	22	No	4	69	3	D	No	60	1	6
Female	Freshman	10	No	4	61	1	D	No	180	0.2	14
Male	Sophomore	13	Yes	1	72	0	D	No	120	150	3.5
Female	Freshman	21	No	5	64	2	I	Yes	30	110	7
Female	Freshman	13	No	4	63	0		No	60	0	2
Male	Senior	11	No	5	67	3	R	Yes	120	1	5
Female	Freshman	7	Yes	4	67	1	D	No	120	0.5	5
Male	Freshman	12	No	2	69	0	D	Yes	0	30	2
Male	Junior	12	No	3	74	1	R	No	120	0.5	2.5
Male	Sophomore	18	No	4	68	3	D	No	60	0	4
Male	Freshman	11	No	4	70	1.5	R	Yes	90	160	3.5
Male	Freshman	12	No	2	69	2	D	No	120	0.5	12.5
Male	Sophomore	9	Yes	6	66	3	R	No	120	1	3

In this lesson, you will explore several important ideas about data. Some of the ideas can be illustrated briefly by examining the survey and its results. For a more detailed study of these ideas, your instructor may refer you to additional resources, perhaps including appropriate pages/sections of a statistics textbook.

1.1 – Variables

The survey contained twelve questions, and for each question the students' answers to the question varied. For example, the first two students in the listing answered "Sophomore" to the question, "What is your class year?" but the next two students answered "Junior" to that question. Because the answers can vary, we describe the situation by saying that *Class Yr* is a **variable**. There were twelve questions in the survey, which led to twelve variables, labeled this way in the table of results:

Gender
Class Yr
States
Smoke
Politics
Height
Newspaper
Political Party
Job
TV
Commute
Course Time

Categorical and numerical variables

One of the most important distinctions to make about data is to determine if the data is **numerical** or **categorical**.

In general, for categorical data there is a fixed list of possible values (answers to the survey question). For example, the *Class Yr* variable has five possible values: Freshman, Sophomore, Junior, Senior, or Other. We say that the data is categorical data, and that *Class Yr* is a categorical variable. For categorical variables, it is quite common that the possible values are words.

On the other hand, for numerical data the possible values are numbers. For this data, it makes sense to make statements such as, "The largest value was 32," or "The average value was 17.2." For example, the question, "How many states have you visited (or lived in or even just driven through)?" generates numerical data, and the corresponding variable *States* is a numerical variable.

Sometimes there are subtleties, illustrated by the *Politics* variable. The students' answers are recorded as numbers from 1 to 7, but these numbers are actually shorthand for the corresponding descriptions such as "very liberal." Therefore, *Politics* is a categorical variable. However, because the numbers correspond to a natural scale from "1 = very liberal" to "7 = very conservative" it does make sense to do certain numerical calculations. For example, if one semester the class average is 2.3 and next semester the class average is 4.7, this conveys the information that the first class is much more liberal, on average, than the second class.

Summarizing data – descriptive statistics

The table given earlier provides the raw data for the class. To aid in understanding the data, it is customary to summarize the data in a number of ways. For example, for numerical data we have already alluded to the possibility of identifying the largest piece of data, or the average (more technically, the *mean*) of the data. These pieces of summary information are referred to as **statistics** or **descriptive statistics.**

> For a particular variable, the *data* provides information about the individuals in the class. The summary data (the *statistic*) provides information about the class as a whole. For example, the first student in the list has visited 31 states; the 31 is *data*. On the other hand, the average for the entire class is 16.7, and this is a *statistic*.

We will have much more to say about summarizing numerical data in Lesson 2. For now, we will focus on summarizing categorical data.

Frequency tables and proportions

We summarize categorical data primarily by counting. For example, consider the *Class Yr* variable. In this class of 30 students, the value "Sophomore" appears 14 times. We can calculate the proportion of sophomores in the class by dividing the number of sophomores (14) by the total number of students in the class (30).

In general, we can calculate proportions using the formula:

$$proportion = \frac{count}{total}$$

In this case, 14/30 = 0.4667. We can also convert this to a percentage by multiplying by 100 (moving the decimal point two places to the right), obtaining 46.67%. For this class, 46.67% were sophomores. By doing the same calculations for the other possible responses, and listing the results in table form, we obtain a frequency table that gives a summary for the entire class:

Class Yr	Frequency	Proportion	Percent
Freshman	9	0.3000	30.00%
Sophomore	14	0.4667	46.67%
Junior	5	0.1667	16.67%
Senior	2	0.0667	6.67%
Other	0	0.0000	0.00%
Total	**30**		**100.00%**

It is even possible to do something similar for a numerical variable, by dividing the values up into ranges. As a very simple example, consider the *States* variable, divided into two ranges 0-25 and 26-50.

States	Frequency	Proportion	Percent
0-25	26	0.8667	86.67%
26-50	4	0.1333	13.33%
Total	**30**		**100.00%**

Exercise 1[1]: Create a frequency table for the ***Political Party*** variable for the class data given above. Note: One person chose not to answer the question – count that response as "Other" for purposes of this table.

The applet at the following link provides practice calculating percent values for a frequency table.

Frequency tables

Example: Suppose that a sociology class has 23 male students and 14 female students. The results of the first test showed that 18 males and 12 females passed the test. Calculate the following proportions, writing the answer as a fraction, as a decimal rounded to 4 places, and as a percent rounded to 2 places.

a. What proportion of the class passed the test?
Solution: There are 23 + 14 = 37 students in the class. Of these, 18 + 12 = 30 passed the first test. The proportion is $\frac{30}{37} = 0.8108108$ as reported by the author's calculator. Rounding to 4

[1] Solutions to the exercises may be found at the end of the lesson.

places gives 0.8108. To convert to a percent we multiply by 100 (shift the decimal place two places to the right), yielding 81.08%.

b. What proportion of the men in the class did not pass the first test?
Solution: This is a question about the men in the class, so we use only the data that pertains to the men. There are 23 men in the class, and 18 did pass the first test, so 23 – 18 = 5 did not pass the test. The proportion is $\frac{5}{23} = 0.2174 = 21.74\%$.

c. What proportion of the class are men who did not pass the first test?
Solution: At first glance this may seem to be the same question as that answered in part (b). However, this is a question about the entire class ("proportion of the class"), where part (b) is a question about only the men ("proportion of the men in the class"). To answer this question, we count the men who did not pass (5, just as in part (b)), then *divide by the size of the entire class* (37). The proportion is $\frac{5}{37} = 0.1351 = 13.51\%$.

Comment. The question in part (c) might be phrased slightly differently, as "What proportion are men who did not pass the first test?" By default, unless the question identifies that it is a question about some subset of the entire class (as in part (b)), it must be a question about the entire class.

The applet at the first of these two links provides practice calculating proportions, allowing you to enter the answer either as a fraction, or a decimal, or a percent. The second link provides some practice converting fractions to decimals and to percents, and converting back and forth from decimal to percent.

Calculating proportions

Conversions

1.2 – Connections to Probability

In this course, we will not study probability in a very technical sense, nor will we study it in great depth. However, we will make extensive use of the connection between proportions and probabilities. As a first example of that connection, consider the *States* variable as summarized in the table above. Imagine the professor that semester choosing a student at random from that class (perhaps by having the names on a thoroughly-shuffled deck of index cards, and choosing a card at random from the deck). What is the probability that the chosen student will have visited 26 or more states?

To solve this problem, we reason as follows: There are 30 students, each equally likely to be the chosen student. Only 4 of these 30 students have visited 26 or more states. So 4 out of the 30 students meet the criterion, thus the probability is "4 out of 30." Mathematically, the probability is 4/30 = 0.1333. Note that this mathematical calculation exactly matches the calculation that determined the proportion for the table above.

In general, proportions and probabilities are related. These questions have the same answer, in general:

- What proportion of the entire group has a certain characteristic?
- If a person is chosen at random from the group, what is the probability that person will have that characteristic?

Example. In the data shown above, 9 of the 30 students are freshmen. For the freshmen, 5 reported having a job and 4 reported no job.

a. What is the probability that a randomly chosen student from this class is a freshman.
Solution: This is the same as the proportion of freshmen in the class, so the answer is $\frac{9}{30} = 0.3000$. Note that we could also write this as 30%.

b. What is the probability that a randomly chosen freshman from this class has a job?
Solution: In other words, if the professor puts the cards for the freshmen in a hat and randomly chooses one, what is the probability that person has a job? This is a question about the freshmen, so there are only 9 cards in the hat – only 9 students to choose from. Of these, 5 have a job, so the answer is $\frac{5}{9} = 0.5556$. We could also write this as 55.56%.

The applet at the following link provides practice calculating probabilities.

Calculating probabilities

1.3 – Samples and Populations; Descriptive and Inferential Statistics

Described very briefly, the word *population* refers to the group of individuals being studied. For the survey described earlier, the population consisted of the 30 members of that class. We can say, for example, that 30% of this population are freshmen, and that 13.33% of this population have visited at least 26 states. In the first half of this course, we will develop a variety of ways to use the data we gather to describe the individuals from whom the data was gathered. This process is generally known as ***descriptive statistics*.**

If a population is very large, obtaining data from the entire population can be impossible, at least from a practical standpoint. For example, gathering data from the entire adult population of the state of Pennsylvania would be extremely difficult and expensive. In spite of this difficulty, in March 2011 it was reported that Governor Corbett's approval rating among Pennsylvania voters stood at 31%. Put another way, for the population consisting of all voters in the entire state of Pennsylvania, 31% approved of the job the governor was doing. How did the pollsters obtain this result? They certainly didn't pose the question to every voter in the state. Instead, they asked a ***sample***, a relatively small group of Pennsylvania voters chosen to be representative of the entire state. From the fact that 31% of this sample approved of the job the governor was doing, they concluded that 31% of the voters in the entire state likewise approved of the job the governor was doing. They used the *sample* to make an *inference* about the entire *population*. This was an example of ***inferential statistics***, the subject matter of the second half of this course. In that second half of the course we will not only address the mechanics of drawing inference from a sample to make a statement about an entire population, but we will also address the question of how it is possible to do so with some degree of confidence in the resulting statement.

Solutions to Exercises

1. Create a frequency table for the ***Political Party*** variable for the class data given above. Note: One person chose not to answer the question – count that response as "Other" for purposes of this table.

Political Party	Frequency	Proportion	Percent
R	14	0.4667	46.67%
D	12	0.4000	40.00%
I	3	0.1000	10.00%
Other	1	0.0333	3.33%
Total	**30**		**100.00%**