Types of Data

Statistical data varies. Consequently, when we speak of data, we use the term variable[1].

Example 1

How big are the families of college students? To learn more, data was collected: Each student was asked "How many siblings are in your family (include yourself)?"

While you and a friend may have the same number of siblings (no variation), you certainly have other friends who have a different number of siblings. In a situation such as this, the object that is measured is a *unit of observation*. Any attribute of the unit (of observation) that is recorded is a *variable*. The following sentence can serve as a template for determining whether you have properly identified the variable and unit in a particular situation. Suppose you identify the variable as "X."

X varies among the units or X varies from unit to unit

In our example a unit is a "student" (one like yourself, who reads this). The variable is "how many siblings." (This could be phrased in a number of equivalent ways: "number of siblings," "count of siblings," "sibling count," "total number of sibling," etc.) To check that these are right, substitute them into the template sentence:

How many siblings varies among students.

or

How many siblings varies from student to student.

Be careful to thoroughly understand how a variable is measured. Is this variable measured by asking students "How many brothers and sisters do you have?" or by asking "How many children in your family?' Of course one follows from the other (they differ by 1), but it's still important to be clear on this issue. In this example we will take it for granted that the sibling count includes the student being asked.

Example 2

In a study of employee efficiency, calls to a service center are timed. Consider the calls fielded by a particular customer service representative. Here the variable is "call length." The units are the "calls." Try this out in our template: *Call length varies among calls.* Or *Call length varies from call to call.* That sounds right.

Example 3

At the same service center, the area code of each incoming call is recorded. The units are still the calls. Now the variable is area code. *Area code varies among calls.* Or *Area code varies from call to call.* Yes.

[1] "Variable" is different in statistics than it is in traditional math. In math you might have the variable x and the function $x^2 + 2$. The value of x varies – but it does so under the control of the person working with the function. In statistics we might measure the variable x to be the weight loss of a person in a diet program. x varies, but not in the same sense as the algebraic "x." You have no control over x = weight loss – you simply measure and record it. As you will see in just a bit, a statistical variable does not even have to be a number.

Example 4

The service center handles customer issues about three different products: Laptop computers, Desktop computers, and PDAs. Now the variable is "which product" *Which product varies among calls.*

Examples 2 – 4 illustrate that a number of variables might be investigated during the course of a single study.

A unit is generally an object or thing; a variable is an attribute of the object. When one examines (studies) different units of observation, different values for a variable arise.

Example 5

In a study of a weight loss aid, the participants are weighed initially and at the conclusion of a weight loss program. Each participant is also asked to respond to a survey question asking: "Overall, how satisfied are you with this program." The participants answer from the choices {"Very" "A good deal" "Somewhat" "Not much" "Not at all". The participants in the program are the units. (When units are humans they are often referred to as subjects.) There are three variables: Initial weight, ending weight, and response to the satisfaction question. The first two variables are measured; the third is queried – a question is posed about the units which, in this case (because they are human) the units themselves answer.

Examples 2 – 4 revisited

The length of the call is certainly measured (probably digitally by the software controlling call routing). We don't really measure area code (nor which product) – but we can think of area code (or which product) as the answer to a question. Of course the units don't "answer" this question – most likely software that handles the phone calls does it automatically.

Data Tables

While we are at it, we may as well discuss how data are arranged. You might be thinking there is flexibility here – if so, you are right. However, there is a standard for organizing data for entry into spreadsheets and statistical software programs. The standard is this: Each row corresponds to a unit of observation; each column corresponds to a variable. (A special row – at the top of our table – lists variable names. A special column often exists to identify the units[2].)

Consider Examples 2 – 4. The units are the calls, which we could label "Call 1," "Call 2," etc. Let's suppose the first call was handled in 5.33 minutes, came from area code 315, and was about a PDA. The data table below is arranged in the standard format:

Call	Length	Area Code	Product
Call 1	5.33	315	PDA
Call 2	7.56	414	Desktop
Call 3	4.89	212	Desktop

[2] This column is sometimes omitted: 1) Sometimes unit identifiers are not recorded or even ignored. You may survey people (= units) and record their responses, but never get to placing their names in the data table. In this case, a spreadsheet can be configured to simply number the people (1, 2, …). Statistical software often has a special column with nothing but 1, 2, … in it – identifying the units. 2) In some cases the natural numbers 1, 2, … are the best way to identify the units. For calls into a service center, you might simply identify them as calls 1, 2, 3….

:	:	:	:
Call *n*	12.03	608	Laptop

You might label the units by the time at which the call was received, rather than "Call 1," "Call 2," …. The letter *n* stands for the number of calls that are monitored.

The call lengths are given in minutes. Minutes are the *measurement units* for the Length variable.

It certainly is *possible* to record some of this data more efficiently. For instance, one could simply tally for the "Which Product" variable:

PDA: |
Desktop : | | [based on what is shown…]
Laptop : |

These tallies form a *summary* of the data. A formal statistical data table will always be more particular and more informative than a summary: It will assign a row to each unit, and list, within each row, the values of all recorded variables for that particular unit.

Quantitative variables vs. Categorical variables

A *quantitative variable* is a variable that is naturally measured as a number on which at least some of the usual arithmetical operations (addition, subtraction, multiplication, division) are meaningful.

> A quantitative variable often comes with a *measurement unit*. The measurement unit specifies the scale on which data are measured. Do not confuse this with the *unit of observation*. When statisticians say "unit" they mean "unit of observation." If statisticians refer to measurement units, they say "measurement units."

A *categorical variable* is a variable that is not quantitative. (The word *qualitative* is also used.)

The value of a variable for a particular unit is called just that – a *value*. Sometimes the term *level* is used instead of *value* – especially for categorical data: "There are three levels of product: Laptop, PDA and Desktop."

Examples

1: Number of siblings is quantitative. If you are from a 4 sibling family, and your friend is from a 2 sibling family, and the two families got together, there'd be 6 total children. It makes sense to add. The measurement unit here is people (which is pretty clear – there's only one sensible way to tally siblings).

2: Length of a service call is certainly quantitative. In fact, the employer wants to see a total length that is small. The measurement unit here is some measure of time (minutes? seconds? hours? – we don't know because this is not detailed in the description of the problem).

3: Area code is categorical. While there are given as numbers, these numbers are merely symbols. It makes no sense to operate on these numbers. (Yes: One *could* add two area codes together. But the result would correspond to nothing.)

4: "Which product" a customer calls a service center about is categorical.

5: "Weight" (both initial and ending) are quantitative. (The measurement units are not specified. In the U.S. probably pounds; in the rest of the world weight would be measured in kilograms.) "Satisfaction" is a categorical variable.

Exercises

For each of 1 – 5, first identify the units of observation. Then identify all variables. For each variable, say aloud the sentence that summarizes the unit/variable relation. It should make sense. Also, determine, for each variable, whether it is quantitative or categorical. For quantitative variable, what are the measurement units (if this is not completely specified, make an appropriate suggestion)?

1. A company manufactures ball bearings that should measure between 9.99 and 10.01 millimeters in diameter. To ensure high quality, each week some bearing are sampled, and the diameter of each is recorded.
2. A fair coin is tossed 100 times.

   Suppose you have an offer to play the following game: You pay $100 (assume you have the money), and then a fair coin is tossed 100 times. The "threshold amount" is 75. If there are 75 or fewer heads, you win $1 ($101 is returned to you). If there are more than 75 heads, you lose the whole $100. Assume that you may play it over and over very quickly so that results "average out."

   a) Would you play this game for $100?

   b) What is the most money you would pay to play this game?
3. A fair die is tossed 100 times.
4. A marketing firm conducts a web survey of 500 people who recently bought a large Napanosic LCD television. Each is asked to complete a questionnaire asking for their annual income, household size, make of the former primary TV in their home, and whether or not they are satisfied to date with their new TV.
5. At a telephone customer service center, a computer automatically records information on each call from a customer, including the area code the call comes from, the length of the call, and how the call was resolved (customer is satisfied, customer hangs up dissatisfied, customer is transferred to speak to a manager).
6. In the situations A - C below, underline the description of the units. Put a circle around the description of any quantitative variable, put a box around the description of any categorical variable. Your marks should not only identify these items, but do so as briefly as is possible. Correctly done, your marks will not overlap. (Note: Exam questions are often formatted this way!)

   A: An experiment was conducted to test the performance of four brands of batteries in three different environments (room temperature; hot and humid; cold). For each combination of brand and environment, batteries were put into a flashlight. The flashlight was then turned on and allowed to run until the light went out. The amount of time until the flashlight stopped shining (in minutes) was recorded. Do brand and environment play a role in the lifetime of these batteries?

Here's the answer to A

A: An experiment was conducted to test compare four brands of batteries in three different environments (room temperature; hot and humid; cold). For each combination of brand and environment, batteries were put into a flashlight. The flashlight was turned on and allowed to run until the light went out. The amount of time until the flashlight stopped shining (in minutes) was recorded. Do brand and environment play a role in the lifetime of these batteries?

Given the statement of this setting, the two circled expressions are equivalent. Either is sufficient for an answer.

B: At the end of May 2010, on-campus student residents at SUNY Oswego were asked to report their class standing (FR, SO, JR, SR) and the number of times they visited home during the Spring of 2010. The number of visits ranged from 0 to 40.

C: An investment analyst collected data to investigate whether or not there is a relation between a company's stock % gain in the last year and the golf handicap of its chief executive officer. The analyst also incorporates the type of industry a company is involved in (for example: Information systems; heavy industry; internet services; etc.) as part of the analysis.

D: (First some background.) The paragraph below is taken from Vanessa Woods' book *Bonobo Handshake* (page 107.)

> There is a left and right side of your brain. When you are negatively aroused, which includes emotions like fear, anger, and stress, the right side of your brain becomes active. As blood rushes in to feed the right hemisphere, it generates enough heat to increase the temperature of your right ear canal. Your hands get cold because in response to stress, all the blood rushes from your outer limbs to your heart. If you are positively aroused, which includes emotions like joy and contentment, it is the left side that is active and the left ear canal that heats up.

Ms. Woods goes on to describe experiments conducted on primates living at sanctuaries (Oleti and Lola ya Bonobo) in Africa . Two species of primate are studied: chimps and bonobos. Each of the apes involved in the study listens to an audio recording of an unfamiliar ape of the same species. Researchers then measure the ape's hand temperature, left ear temperature, and right ear temperatures.

E: My iTunes music collection contains thousands of songs. For each song, I can display the artist, genre, playing time, number of times played, and the star rating (* = horrible; ***** = awesome) I have assigned the song.

F: 55 year old men are recruited into a study about heart attacks. The heart rate of each man is recorded. Each is tracked for a one-year period, and whether or not he has a heart attack is determined.

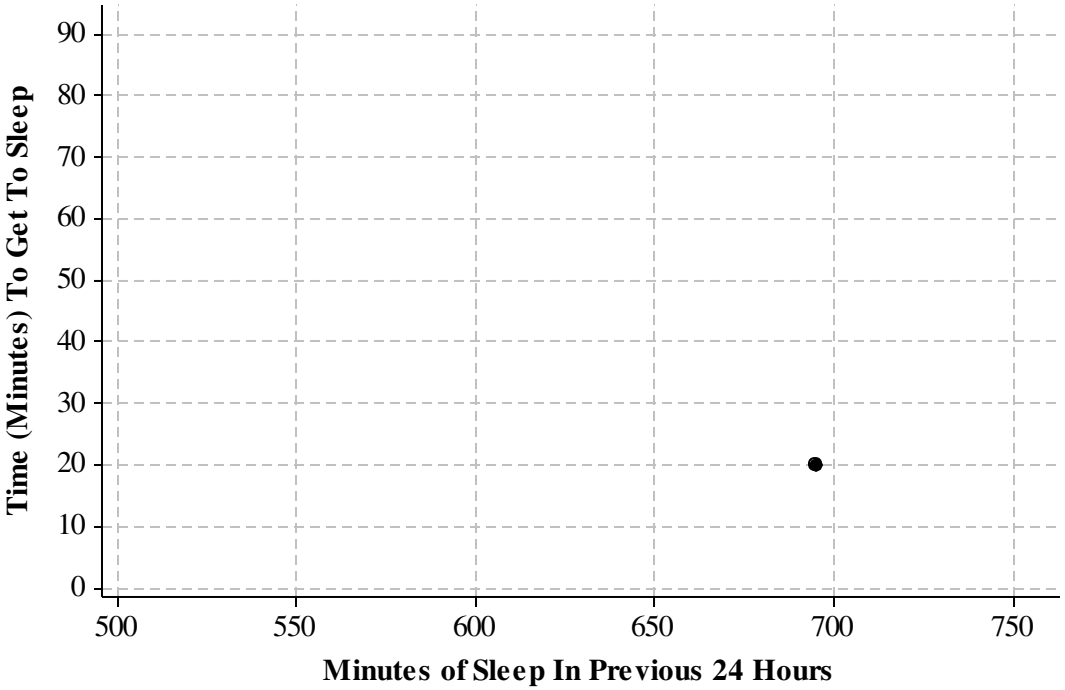
G: A student runs an experiment to study the effect of tire pressure on gas mileage. He devises a system so that his car uses gasoline from a one-liter container. Each time the container is filled, he randomly selects a tire pressure between 20 and 35 psi, then drives the car at 60 mph on a divided highway. When he runs out of gas, he records the distance driven on that fill. Does tire pressure impact the distance driven?

7. For exercises 1 – 5, show how the first 2 – 3 rows of the data table might look (you can arbitrarily make up data to illustrate).
8. Give an example of some interesting data you might collect. Identify the (statistical) units, all variables, and, if applicable, the measurement units for the variables. Sketch out the beginnings of a data table.
9. A young girl has a sleep disorder. Her parents pay very close attention to her sleep patterns. Each evening, after the girl has fallen asleep, the parents fill out an index card with the following information: A rating on the child's behavior throughout the day ("5" = great, "4" = generally good with minor anger and frustration issues, "3" = multiple minor issues or one major tantrum, "2" = one major tantrum and multiple minor issues, "1" = two major tantrums, "0" = awful); The amount of sleep for the child since falling asleep on the previous day; Whether or not she attended day care that day; How long the girl took to go to sleep; How much television the child watched that day.

   (Note: Be careful identifying the units. One variable is "amount of TV watched." Now – from the parents' records, how would you fill in the following? "The amount of TV watched varies from _____ to _____." That should give the right units.)

   One theory about sleep patterns in children is that generally, the more sleep they get, the easier a time they have getting to sleep. Is this true? Here's the data from the cards on the amount of sleep the girl got since the previous bedtime, and the amount of time it took her to get to sleep that night. Both variables are measured in minutes.

Date	Sleep In Previous 24 Hours	Time To Get To Sleep	Date	Sleep In Previous 24 Hours	Time To Get To Sleep
1/1	695	20	**1/12**	575	20
1/2	675	20	**1/13**	680	55
1/3	625	20	**1/14**	665	30
1/4	655	40	**1/15**	585	30
1/5	715	40	**1/16**	705	55
1/6	710	25	**1/17**	690	55
1/7	555	45	**1/18**	700	70
1/8	750	90	**1/19**	600	35
1/9	*	*	**1/20**	665	40
1/10	725	40	**1/21**	690	55
1/11	600	10	**1/22**	530	25



a) Carefully plot these on the grid. (Obviously you cannot plot for 1/9.) The data from 1/1 has been plotted for you. Note that what comes first (Sleep in Previous 24 Hours) is treated as x, and what follows (Time To Get To Sleep) is treated as y. This is common practice in statistics when we want to see if we can predict what happens next (y) from what has already happened (x).

b) Now consider the line that has equation $y = 0.167x - 70.6$. This line was obtained, using statistical methods, from the data that are tabulated above.

   i) Carefully plot this line on the scatterplot. Use a straightedge.

   - First compute $y = 0.167x - 70.6$ for $x = 500$. This gives you one coordinate at the left of the grid. Plot this (x , y) point.
   - Now choose a convenient x value at the right of the grid. Find the corresponding y and plot this point.

   ii) Find the average of the x values. Find the average of the y values. (This choice of x and y might be the best for the parents to fill in for 1/9.) Plot the two averages on the scatterplot using an X. What do you notice?

   iii) Use the equation to predict the time it will take the girl to fall asleep on an evening where she has gotten 735 minutes of sleep in the previous day.

   iv) Determine the value of x that gives $y = 0$ in this equation. How does this apply to the girl's sleep situation?

c) Does the data support the theory that generally, the more sleep a child gets, the easier a time they have getting to sleep?

10. Read the following, as it appeared in the New York Times.

> Prevention: Evidence of Heart Benefits From Chocolate
>
> By NICHOLAS BAKALAR
>
> Published: August 29, 2011, New York Times
>
> An analysis of studies including more than 100,000 subjects has found that high levels of chocolate consumption are associated with a significant reduction in the risk of certain cardiovascular disorders.
>
> The seven studies looked at the consumption of a variety of chocolate — candies and candy bars, chocolate drinks, cookies, desserts and nutritional supplements. By many measures, consumption of chocolate was linked to lower rates of stroke,coronary heart disease, blood pressure and other cardiovascular conditions.
>
> But there was no beneficial effect on the risk for heart failure or diabetes.
>
> Over all, the report, published Monday in the British medical journal BMJ, showed that those in the group that consumed the most chocolate had decreases of 37 percent in the risk of any cardiovascular disorder and 29 percent in the risk for stroke.
>
> Still, the lead author, Dr. Oscar H. Franco, a lecturer in public health at the University of Cambridge, warned that this finding was not a license to indulge and noted that none of the studies reviewed involved randomized controlled trials.
>
> "Chocolate may be beneficial, but it should be eaten in a moderate way, not in large quantities and not in binges," he said. "If it is consumed in large quantities, any beneficial effect is going to disappear."

Conjecture reasonably on the variables that were assessed in the cited studies.

Solutions

1. Units = ball bearings. Variable = diameter. Diameter varies among ball bearings. Diameter is a quantitative variable. The measurement units are millimeters.
2. Units = tosses. It's not really stated what the variable is here. Most people would assume that what is recorded is whether the outcome of a toss is Heads or Tails. Variable = "Whether Heads or Tails" (or "the side of the coin that lands face up") "Whether Heads or Tails" varies from toss to toss (among tosses). "Whether Heads or Tails" is a categorical variable.
3. Units = tosses. It's not really stated what the variable is here. Most people would assume that what is recorded is the number of "pips" on the side of the die that lands face up. Variable = "Number of Pips." "Number of Pips" varies among tosses (from toss to toss). "Number of Pips" is a quantitative variable. The measurement units are "pips."
4. Units = the 500 people who recently bought this TV. There are 4 variables: Annual income (quantitative; measurement units probably $); household size (quantitative; measurement units of "people"); make of former TV (categorical); and "Whether or not they are satisfied with the TV" (categorical). For example: Annual income varies among people who recently bought this TV.

5. Units = calls. There are 3 variables: area code (categorical); length (or time) (quantitative; measurement units are seconds, or minutes…); resolution (or "How the issue is resolved") (categorical). For example, "How the issue is resolved varies from call to call (among calls)."

6.

Part	Units	Quantitative Variable(s)	Categorical Variable(s)
A	• batteries	• lifetime	• brand • environment
B	• on-campus student residents at SUNY Oswego	• number of times they visited home during the Spring of 2010	• class standing
C	• companies	• stock % gain in the last year • golf handicap of its chief executive officer.	• type of industry
D	• apes	• right ear temperature • left ear temperature • hand temperature	• species (bonobo or chimp)
E	• songs	• playing time • number of times played • star rating	• artist • genre
F	• 55 year old men	• heart rate	• whether or not he has a heart attack
G	• fills of the container or drives (of the car)	• distance driven • tire pressure	• [none]

7. Exercise 1:

	Diameter
Ball Bearing 1	10.24
Ball Bearing 2	9.98
:	:

Exercise 2:

	Side Up
Toss 1	HEAD
Toss 2	TAIL
:	:
Toss 100	TAIL

Exercise 3:

	Number of Pips
Toss 1	1
Toss 2	5
:	:
Toss 100	3

Exercise 4: (The names are made up.)

	Income	Household Size	TV Make	Satisfaction
Bill Evans	52,500	4	SAMSUNG	YES
Eric Dolphy	40,800	3	SONY	YES
John Gillespie	97,200	5	SONY	NO
:	:	:	:	:

Exercise 5:

	Area Code	Length	Resolution
Call 1	315	15.23	Satisfied
Call 2	315	4.25	Transferred
Call 3	404	18.25	Dissatisfied
:	:	:	:

8.

9. Units = days (or nights) – the variables are recorded, and vary from, day to day. The variables are pretty much listed in the description of the problem. It is not clear whether the "behavior rating" should be thought of as quantitative or categorical. The variable "Whether or not went to day care" is categorical, all the others are quantitative. For the time amounts given in the table the measurement units are minutes. b) ii) The averages are 656.7 and 39.05. When plotted, this falls exactly on the line. iii) $0.167 \times 735 - 70.6 = 52.15$ minutes. iv) 422.75 minutes which almost 7 hours and 3 minutes. If the child gets 7:03 (about 7 hours) of sleep the previous day, then she is predicted to fall asleep instantly tonight! This has to be somewhat nonsense. c) No. The plot shows that, at least for this child, generally speaking the

more sleep she gets, the longer it takes her to get to sleep.