VARIABLES: Notes for M1100

I. What is a variable?

In basic algebra, a variable is a quantity that may be unknown or may be changing. It's the x that troubles middle-school students who wonder how a letter can be a number. In this course, we are concerned with data analysis. Here the word "variable" has a somewhat different meaning. It's just something that you go out and measure or observe in different individuals or in different objects or on various occasions. Here are some examples:

- Variables associated with people include: *date of birth, weight, height, years of education, annual income, hair color, sex, religion, marital status, whether or not a person ever smoked.*
- Variables associated with inanimate objects include: *weight, color, composition.*
- Variables associated with events include: *time, place, duration.*
- Variables associated with orders placed on `amazon.com` include: *time the order was placed, shipping address, items ordered, the total dollar amount.*

II. "Determining the value of a variable" means measuring it.

When we speak about variables, we use a specialized vocabulary and a language that is slightly artificial. The purpose is only to be clear and precise. When talking about variables we often refer to the *values* that a variable takes at different times or for different individuals. This is a convenient piece of vocabulary that I would like you to learn to use. To make the meaning clear, let me give some examples that show you how to use the words "variable" and "value" to say some things that you can also say without them.

Example. I weighed myself last Sunday. My weight was 162. I had gained two pounds since the previous Sunday. The variable here is my weight. This variable changes with time. *The value of the variable "my weight" was 160 on the previous Sunday and its value last Sunday was 162.*

Example. Imagine the people in a room. Each one has a date-of-birth, and this varies from person to person. This is why we call "date of birth" a variable. On the other hand, *your* date of birth is not a variable. Your date of birth is a specific date that is fixed and unchanging. If you were born, say, on May 15, 1980, then we would say that May 15, 1980 is the *value* that the variable *date-of-birth* that is associated with you.

Example. If the variable is hair color, then some of the possible values are black, brown, red, blonde, gray and white. Hair color tends to run in families. *Certain values of the variable "hair color" occur with greater frequency in some families than in others.*

Example. Mark is taller than Fred. The variable here is height. *The value of the height variable associated with Mark is greater than the value associated with Fred.*

Summary. When we make a measurement or observation, we determine a *value* of whatever variable it is that we are observing. On different occasions, or when dealing with different individuals, a given variable takes on different values.

Remark. You could probably get away without ever using this kind of terminology if you always restricted your attention to just one variable at a time. Why then do we need this vocabulary? In order to make general statements about variables, we need a uniform way to refer to the values that we record when we make observations or measurements. So the variable/value terminology is useful.

Problems

1. List a variable (other than any that I have already mentioned) that is associated with:

- **a.** people,
- **b.** objects,
- **c.** events,
- **d.** presidents,
- **e.** meals,
- **f.** shark attacks.

2. Make a statement about a value of each variable you mentioned in problem 1.

3. Is "clothing" a variable for people? If you think so, then what would you list as George Bush's clothing? If a person's clothing can change, then is clothing a variable of person, or is it a variable of person *and* and an occasion?

4. List some variables that you find interesting, and discuss your choices with others. (Students are likely to produce some examples that are particularly good illustrations of the idea and some other examples that are clumsy or that there is disagreement about. What is a clearer example of a variable: "date of birth" or "clothing"?)

III. Kinds of variables

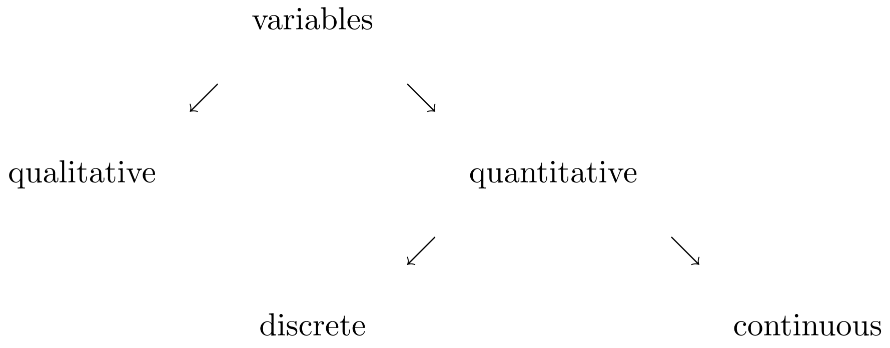
Variable may be either *qualitative* or *quantitative.*

Qualitative variables classify. *Examples. Hair color, blood type and race* are some qualitative variables that are applicable to people. *Make, model and color* are some qualitative variables that are applicable to cars.

A qualitative variable that is associated with a yes-no classification is called *boolean.* For example, married/not married is a boolean variable.

Quantitative variables either count or measure on a numerical scale. *Examples. Population* is a quantitative variable of cities. *Number of teeth* is a quantitative variable of people. *Age, height and weight* are quantitative variables of people. *Temperature* is a quantitative variable (of place and time).

A quantitative variable, such as population or number of teeth, that is always a whole number is called *discrete.* A variable such as age, weight or height that can take a value on a continuous scale is called *continuous.*



Summary. A discrete quantitative variable is a variable whose values are whole numbers. A continuous quantitative variable is a variable whose values may be anywhere on the number line. A qualitative variable is a variable whose possible values are categories or classes.

A note on qualitative variables.

The values of a qualitative variable must be non-overlapping classes. One way of determining whether a qualitative variable meets this requirement is to think of it as a question on a multiple choice test. There must always be one and only one correct response, no matter what we are examining. In contrast, a question that instructed respondents to "check all that apply" would not meet this requirement.

Example. Suppose you want to keep track of what people order at a restaurant. You should have a variable whose values tell you what a person has ordered during a visit. The values of this variable will *not* be the menu items, because a person might eat several items in one visit (or might order one thing two or more times). To actually gather data, a natural thing to do would be to give each diner a copy of the menu and ask them to mark how many orders of each item they had eaten. If we do this, we are not treating the menu as a variable; we are treating each separate menu item as a variable all by itself. Each menu item plays the role of a discrete quantitative variable. The reference class (see below) for each of these variables is the collection of all visits to the restaurant by individual diners. The value of the variable at any particular visit is the number of orders of that menu item that the diner ordered.

Example. In order to enforce civil rights laws, the federal government monitors access to housing, education and employment among different racial groups. For this reason, the government needs a clear and fair way of categorizing people according to race. The system that was used in the 1990 census was criticized for not giving adequate attention to mixed races. In response, the Office of Management and Budget (OMB) studied the problem and in 1997 created a new set of standards for federal data on race and ethnicity.

The new standards largely determined the way questions about race were constructed in Census 2000. Census questionnaires included the following six major racial categories: 1) *American Indian or Alaska Native*, 2) *Asian*, 3) *Black or African American*, 4) *Native Hawaiian or Other Pacific Islander*, 5) *White*, 6) *Some Other Race.* People filling out Census 2000 forms were instructed to select *one or more* of the six categories.

The census appears to have gone against the requirement of non-overlapping categories that I recommended above. The complex (but well-considered) reasons for doing so are discussed at the web site *Recommendations from the Interagency Committee for the Review of the Racial and Ethnic Standards*, which is at

`http : //www.census.gov/population/www/socdemo/race/Directive_15.html`

The standards set by the Office of Management and Budget satisfied the government's needs, but by allowing overlapping categories, the census created some significant bookkeeping problems. If we look at these, we can begin to understand why non-overlapping categories are desirable. Consider the following questions:

1. Suppose the census were to report the total number of boxes checked on the racial identity question. Could this number exceed the total population of the country?
2. From the data the census collected, how could you tell the percent of the population falling into each of the six major racial category? Would the percents add up to 100?
3. Some federal programs are such that they benefit only one race. (For example, some genetic diseases occur in only one race, and any money spent to find a cure would benefit only that race.) One might argue that the money put into such programs should be distributed according to the size of the racial groups. If the categories do not add up to 100%, then what's the right way to make the division?
4. Suppose we limit ourselves to the data from single town. Suppose in this town 1500 people indicate that they are at least part *Asian* while 1800 indicate that they are at least part *White.* Would a White candidate for mayor have an advantage over an Asian candidate?

You may have felt that the last question lacked some information. What prevents you from reaching a conclusion is the fact that the categories overlap, but you don't know how much they overlap. Some people may belong to several groups, or even all of them.

How should the census report the racial data collected? We get a much clearer picture of the population if we divide the population into mutually exclusive groups. One way to do this is to consider all the different ways of selecting one or more of the racial categories. There are actually 63 possibilities, giving us 63 different, non-overlapping categories. (There are six categories for those who report only one race and 57 additional categories for people who report two or more races.)

Some of the categories have very few members. Therefore in some census reports the 57 categories are combined as a single category called *Two or More Races.* When this is done, the entire population breaks up into seven mutually exclusive categories.

Projects.

Read the report at the web site mentioned above.

1. The OMB standards treat Hispanic ethnicity as a boolean variable that is independent of race. Explain what this means and why this standard for data on ethnicity was instituted.
2. Write a report summarizing the OMB standards for reporting racial and ethnic data and discuss the political considerations that may have helped to shape the standards.

IV. The Reference Class of a Variable.

Some variables only apply to certain limited classes of things. For example, *date inaugurated* makes sense for presidents, but not for other people. *Number of calories* makes sense for a piece of food, but you don't measure how many calories a shoe has. Similarly, events of a specific type may have certain special variables attached—magnitude on the Richter scale is a variable that makes sense for earthquakes, but not for sunsets.

We will call the class of all those things for which a variable makes sense the *reference class* of the variable. The members of the reference class are called *observational units.*

Sometimes the reference class is so clear that we don't even mention it. If the variable is grade earned on a certain quiz in this course, then the reference class is made up of the students who took the quiz.

Sometimes the reference class of a variable may be somewhat vague. For example, the variable *LSU grade point average* makes good sense for most LSU students. But do you need to be enrolled in a degree program to have a GPA? How about those students in their first semester? Do they have a grade point average before the end of the semester? Does a high school student who takes one LSU course after school have an LSU GPA? What we do to sharpen the boundaries depends on what we intend to do with the variable. For example, if we want to compare the average GPA of LSU students with the average GPA of Texas A& M students, we would want to make sure we were averaging comparable groups. We might, for example, restrict the reference class to students who had already earned grades in at least 12 hours and were presently enrolled full time.

There are times when the reference class of a variable is restricted to a group smaller than the full collection for which the variable makes sense. The meaning of any summary data about the variable—such as average value (for a quantitative variable) or the percents in various categories (for a qualitative variable)—depends on the reference class. If it is reported that a certain variable has such-and-such an average, or that such-and-such a category occurs in 90% of all cases, then it is important to know the reference class. If, for example, you are told that 9 out 10 patients benefit from a certain drug and your doctor is considering giving it to you, you may well question what the composition of the refernce class was in which the 9-out-of-10 benefit rate ocurred. How similar to the members of the reference class are you? If you are told that the rate of side effects is 5%, does that mean that 5% of the people who take the drug experience side effects, or does it mean that 5% of the times a dose is taken there is a side effect. Does the reference class consist of people taking the drug, or does it consist of the ocasions on which the drug is taken?

V. Measurement Procedure.

The meaning of the data we gather when we measure a variable depends on the procedure used. The name we give the variable may be misleading, because it may suggest a meaning that is only indirectly related to the procedure used to make the measurement.

Example. IQ stands for "Intelligence Quotient". The name suggests that IQ ought to be a good measure of intelligence. But how is IQ measured? Typically, it's calculated from a score on a written test of a very special kind. But there are many kinds of intelligence that might not be related to performance on such a test—for example, the ability to understand another person's feelings or the ability to improvise music. Intelligence, in fact, is a concept that has different meanings to different people. A cautious, critical attitude, therefore, takes IQ simply as reflection of a person's performance on a special test, and does not infer from the score that the person has (or fails to have) other attributes that thought of as intelligent.

Example. "Maternal mortality" refers to rate at which women die for reasons related to childbirth. It is a variable that can be measured in a given country and year. The observational units are countries in specified years. We can measure maternal mortality in Canada in 1930; we can measure it in Afghanistan in 2002, and so on. Maternal mortality is sometimes used as a indicator of the general health of a population.

How is maternal mortality measured? It's not appropriate simply to count the number of pregnancy-related deaths in a country, since a large country will have more deaths simply because the population is larger. We ought to take the population into account. But the age distributions and the proportion of women of child-bearing age may vary from country to country, as may the rate at which women become pregnant. So considering the population alone is not enough. To acoount for these factors, a common way to measure maternal mortality is to report the number of maternal deaths per 100,000 births. Even when this is accepted, there is still the question of how to decide if a death is pregnancy-related. Generally, whether or not to classify a death as pregnancy-related will be a doctor's decision, but doctors in different countries may use different criteria. Before comparing the maternal mopratlity figures from different countries, one should verify that the same criteria are being used.

Summary.

In describing a variable, one needs to consider the kind of variable, the reference class and the procedure used to make the measurement. If we are clear and explicit in describing these things, then we have a better chance of avoiding error or misinterpretation.

Homework.

Find a newpaper article in which a variable plays a prominent role, then provide answers to the follwoing questions:

1) What is the variable your article concerns?
2) What kind of variable is it? What are its possible values?
3) What is the reference class? If this is not clear from the article, say so. What *might* the reference class be?
4) What procedure is used to measure the variable? If this si not clear, say so. What are some procedures that *might* have been used?