2typesofvariables.pdf	Michael Hallstone, Ph.D. hallston@hawaii.edu

Lecture 2: Types of Variables

Recap what we talked about last time

Recall how we study social world using populations and samples.

Recall that when we study the social world using statistics we need to define characteristics that we wish to study so that they can be expressed using numbers. And when we do this we "code" variables.

Types of variables

There are two ways to classify variables that will be important to us in this course. One is to decide whether a variable is *continuous* or *discrete* and the other is to decide whether a variable is *nominal, ordinal, interval, or ratio.*

continuous vs. discrete

Some of the statistical tests assume the variable is continuous or discrete. For example, the theory behind a test called Chi-Square requires discrete variables. Therefore if the test you use assumes a discrete variable you must use one, or you are violating a key assumption of the test. This is why we need to learn to distinguish between continuous or discrete.

discrete variables

Consider the definition of a discrete variable first. A variable that is discrete is coded with "either/or" categories where there is NOTHING "in between" the categories. <u>There is nothing in-between these categories. You are in one or the other but not both.</u>

For example, imagine the variable *T-shirt size: 1= small 2= medium 3=large*.

There is no way to be in between categories in this variable, so it is discrete. You either wear a small or a medium or a large and there is nothing in between.

Any variable with a yes or no answer is discrete. Have you ever been convicted of a crime? 1= yes 0=no. The answer is either yes or no and there is nothing in between. Did you go to the doctor in the past 30 days? 1= yes 0=no. The answer is either yes or no and there is nothing in between.

Any variable where the categories are "either/or" is discrete. Who is your health insurance provider? 1=HMSA 2= Kaiser 3=other 4= uninsured. You choose one of the four and there is nothing in between.

To use a public administration example, pretend you had a variable the measured under which branch of government a person's job in the state government was organized. (Generally speaking, there are three main branches of state or federal government under which all departments are organized: executive, judiciary, and legislature). So pretend there was a variable *branch: 1= executive 2= judiciary 3=legislative.* Now pretend you had a respondent who worked in the prison system for the State of Hawaii (in Hawaii the prisons are run by the Department of Public Safety that falls under the executive branch of government). So that person would be coded *1= executive.* <u>The important thing is these categories are discrete. There is nothing in-between these categories. You are in one or the other but not both.</u>

continuous variables

Continuous variables exist on a continuum, thus the term "continuous." There are NOT discrete categories in a continuous variable, but are an infinite number of responses along a number line.

So pretend you had a variable that measured *length of longest fish ever caught _____(fill in the blank).* If you think about it, if you had a precise enough measuring instrument there are an infinite number of possible responses between 68 and 69 inches. It is possible to have caught a fish that was 68.25 inches or 68.251 inches or 68.2512 inches, etc. etc. on and on and on.

Many variables that "code naturally" can be considered continuous variables: *age _____, height____, weight_____, years worked in your current department _____, etc.*

Some times we "pretend" it is continuous

Now to complicate matters a bit. Take the variable age for example. Even if you measured a person's age to the second, the distance between seconds is, theoretically discrete. Or if you measured a person's age in whole years the distance between years is discrete. However, in practice it okay to

use a *variable age in whole years* when using some tests that are designed for continuous variables. (This is because *in theory* the variable is continuous, even though in this case the data collected was collected in a discrete manner).

On a test I will make it very clear – with this one "trick" question

Bubble test questions will mostly look like this

> Kelly measures the exact length of a person's surfboard. He uses a really sophisticated measuring device that allows very precise measurement. He will not use either/or categories, but will measure them exactly. Classify the variable. ***Here I am giving you the hint that the correct answer is continuous because he does NOT use either/or categories.***
>
> Kelly measures many years a person has been a surfer. He collects the data in whole years. Classify the variable. ***Here I am giving you the hint that the correct answer is discrete because he collects in whole years.***
>
> Blu will measure how long prisoners have been incarcerated. Choose the best statement below regarding this variable. There will be answers that say "this variable must be continuous" and this variable must be "discrete" but **the correct answer is Depending upon how he collects the information, it could be either continuous or discrete.**

Test yourself to see if you REALLY understand the concept continuous vs. discrete

This is a bit theoretical, but it will create a self-test and increase your critical thinking skills. Classify these variables as continuous or discrete:

number of deaths in US prisons last year _____ (insert #)

number of people treated in the Emergency Room in Queen's Hospital last year _____ (insert #)

number of felony convictions in your life _____ (insert #)

At first glance all of these might appear to be continuous as they are number variables that "naturally code." And many times natural coding number-type variables are continuous in theory, but THESE ARE NOT. Each of these are discrete as there is no such thing as "half a death" or "half a person treated" or "half a felony conviction." You should know this concept for the test, because I want to see if you "really" understand the concept.

That being said, each of these variables could be used on most tests in this course that are designed for continuous variables.

The *really important* way to classify variables for this course is next.

Is the variable nominal, ordinal, interval, or ratio?

All of the statistical tests you will learn in this course will require you to classify variable into one of 4 categories *nominal, ordinal, interval,* or *ratio.* So for example, some statistical tests you learn are designed for nominal level variables, for example something called *chi-square*. So when you do a chi-square test you MUST use a nominal variable or you are violating a key assumption of this test and your results will be, essentially, meaningless.

Thus, it is very important that you be able to decide whether a variable is either *nominal, ordinal, interval,* or *ratio.*

Most important of all is whether or not the variable is fit into one of 3 categories: *nominal, ordinal, interval/ratio.* (As you will see later in the course many of the statistical tests are designed for interval or ratio level variables, for example, a *t-test or a hypothesis test of means* requires a variable that is at least interval/ratio.) By the way, the statistical software program SPSS uses only three categories: nominal, ordinal, and "scale" where scale includes interval and ratio level variables.

By the way, some books will simplify these four categories into something like "categorical" or "attribute" and "numerical," but let's learn the "real way" instead. If you go on to use statistics later, learning this way will help. (Essentially, it appears that nominal and ordinal are attribute and interval and ratio are numerical.)

NOIR spells "black" in French!

The word NOIR is black in French and that allows you to keep these in their proper order. As we go from Nominal to Ratio, the variables are getting more complex. You must keep the categories *nominal, ordinal, interval,* or *ratio* in that exact order to understand how to classify these variables.

N - nominal
O - ordinal
I - interval
R -ratio

N - nominal

A nominal level variable has categories only. There is no dimension or numbering to the categories. The numbers assigned to each category are MEANINGLESS and are just a grouping procedure.

Consider the following variables, they are all nominal.

Gender: 1=male 2=female
Ethnicity: 1=Haole 2=Polynesian 3= Asian 4= other
Favorite plate lunch: 1=chicken katsu 2=mixed plate 3=kalbi ribs 4=other
Have you ever surfed: 1=yes 0=no

We assign numbers to each category, but the numbers do not mean anything. The numbers are meaningless. You could swap the numbers any way you want and it would not matter. 1 is not "less" ethnicity than 2 and vise versa. 1 is not "less gender" than 2. Any variable with a "yes/no" answer is nominal.

O - ordinal

The numbers have an order! The order is meaningful. The numbers mean something. We have the beginnings of a dimension but no common units of measurement -- no way exactly measure distance between categories. We have an order but they are not real numbers (because there are not common units to the numbers you cannot "do math" on the numbers").

Consider the variable: *t-shirt size: 1=small 2=medium 3=large 4=extra large.*

A large t-shirt (3) is NOT ACTUALLY three times larger than a small (1) t-shirt is it? But the number 3 is actually "3 times the size of the number 1." An extra large (4) is NOT ACTUALLY twice as large as a medium (2), but the number 4 is actually twice a big as the number 2.

The professor explains things in an easy-to-understand manner.
1=strongly disagree
2=disagree
3=neutral
4=agree
5=strongly agree

As you "go up" in numbers you do have "more agreement" although you cannot say that strongly *agree (5)*is five times as much agreement as *strongly disagree (1).*

I - interval

An interval variable has categories that are “real numbers” with common units of measurement, but there is NO TRUE ZERO. The zero does not indicate “a complete absence” of something. There are not lot of truly interval variables. Two I know of are *temperature in degrees* (it does not matter whether or not it is measured in Fahrenheit or Celsius scale) and *calendar year (in relation to the birth of Christ).*

Temperature: _____ (insert # of degrees).

Year you were born:_____

For both variables, you have numbers and they are real numbers. For example 40 degrees is half the temperature of 80 degrees. 81 degrees is one more degree than 80 degrees. The year 2000 is twice as many years as 1000. You could do math on these numbers. You have real numbers however there is no “true zero.” Zero degrees (F or C) does NOT indicate a “lack” or an “absence” of temperature. Temperature will be on the test! As I say in class, “When you see temperature you say interval.”

Years are also measured in exact intervals, meaning each year is exactly the same size. But the year zero is not an absence of years. It is merely marking the year of Christ’s birth.

You could also take a ratio variable and recode it into categories that were exactly the same size and you could turn a ratio variable into an interval variable. So you could recode age into five year intervals and you’d have an interval variable. But age is not an inherently interval variable.

The fact that there are very few actual interval level variables is why, in practice, for many statistical tests we simply need an interval or ratio level variable.

R -ratio

A ratio level variable is the same as an interval variable, but they have a “true zero.” You have real numbers with real intervals and a zero means “a lack of” or an “absence of” what ever you are measuring.

# of cars you own _____
Amount of money in your wallet right now_____

# of felony convictions in your life
# of children you have ____
# of times you have been surfing ____
# of times you have jumped out of an airplane ____

In each of these cases if you answer 0, that means a complete absence of that thing. If you own zero cars, that's an absence of cars, if you have no money, that is an absence of money, if you have no children that is an absence of children, and so on.

Both interval and ratio are "real numbers"

So you can see that both interval and ratio level variables are composed of "real numbers." You can "do math" on real numbers. This is why earlier I wrote, "it is most important to know whether or not the variable is fit into one of 3 categories: *nominal, ordinal, interval/ratio.* (As you will see later in the course many of the statistical tests are designed for interval or ratio level variables, for example, a *t-test or a hypothesis test of means* requires a variable that is at least interval/ratio.) By the way, the statistical software program SPSS uses only three categories: nominal, ordinal, and "scale" where scale includes interval and ratio level variables."

"Dummy" variables

"Dummy" nominal variables to pretend they are "ratio"

Sometimes we can code a nominal (or ordinal) variable into 1's and 0's for statistical techniques like regression and "pretend" it is a sort of ratio variable.

When we cover multiple regression we recode gender into a dummy variable. And as I explain elsewhere, I'm hip to the idea that gender can exist on a continuum, but to keep this simple we shall use an example that can be easily understood. I'll use a cis-gender manifestation of a socially constructed concept as institutional sexism exists in many, but of course not all, parts of society. (Facts exist whether we like it or not.) I'm not trying to offend sensibilities here, just teach statistics in an easy-to-understand manner, so please let's not politicize this. That being said, instead of male and female as described below we could pretend we used "conforming and non-conforming" (and then you'd have to assume that construct affected income).

Most students code gender into 1 and 2 in some fashion and for regression we need it coded into 1 and 0 to dummy it. This is true for any nominal variable by the way: we can recode it to 1 and 0, where 1 is considered 100% of the characteristic and 0 as 0% of the characteristic. So if you have

been paying attention, in American society women are often paid less than men ***even when controlling for education and experience etc***. (When you control like this it is evidence that creates evidence that contributes to creating a "fact.") There is no simplistic sound bite explanation I can do here but the reasons are complex. But the fact remains, that as a group this is basically true for women in many fields of employment.

So for a test problem for lecture 22 we will examine income and one of the variables we will use will be gender. We use women as the "characteristic of interest" as they are the group who often sees their salary affected, so we code female=1 and men=0. This way we can analyze what is called a coefficient and say something like "When all other independent variables are held constant, when the person is female (as compared to male), monthly income decreases by $1266.40."

By the way there are examples when men are the characteristic of interest. Men are more likely to have more criminal convictions than women and if we wanted to examine this variable in the context of criminal convictions, we would code male=1 and female=0. Then we could say something like say something like "When all other independent variables are held constant, when the person is male (as compared to female), the number or criminal convictions increases by 1.2."

This is true for any nominal variable by the way: we can recode it to 1 and 0 , where 1 is considered 100% of the characteristic and 0 as 0% of the characteristic. The nominal variable can have more than 2 categories.

So pretend we had a simplistic coding of race in the US prison system. (I am also aware that race is a social construction and not biological when you examine the DNA of the human genome. I won't bore you with the details but am happy to explain what this means verbally.) But systematic racism exists in the US prison system in the sense that Black and Hispanics are over-represented in our prisons compared to Whites. (These are the terms used by the Department of Justice for which I can show data to prove this.). Again, there is no simplistic sound bite explanation I can do here but the reasons are complex.

So pretend the Dept. of Justice coded race like this and we wanted to look at Blacks compared to Whites and Hispanics: 1=White 2=Hispanic 3=Black. We would code Black =1 and Hispanics and Whites would be coded 0. Then we could say something like "When all other independent variables are held constant, when the person is Black (as compared to White or Hispanic), the incarceration rate increases by..."

You can dummy an ordinal variable in this same manner as noted below.

Dummy" ordinal variables to pretend they are "ratio"

Pretend you had an agreement variable like this and you wanted to compare people who agree to some statement as those who are neutral or disagree.

1=disagree
2=neutral
3=agree

You recode so agree =1 and neutral and disagree =0. Then you can make a statement like "When all other independent variables are held constant, when the person is agrees (as compared to those who are neutral or who disagree) to the statement that 'Professor Hallstone is really confusing me right now'…"☺