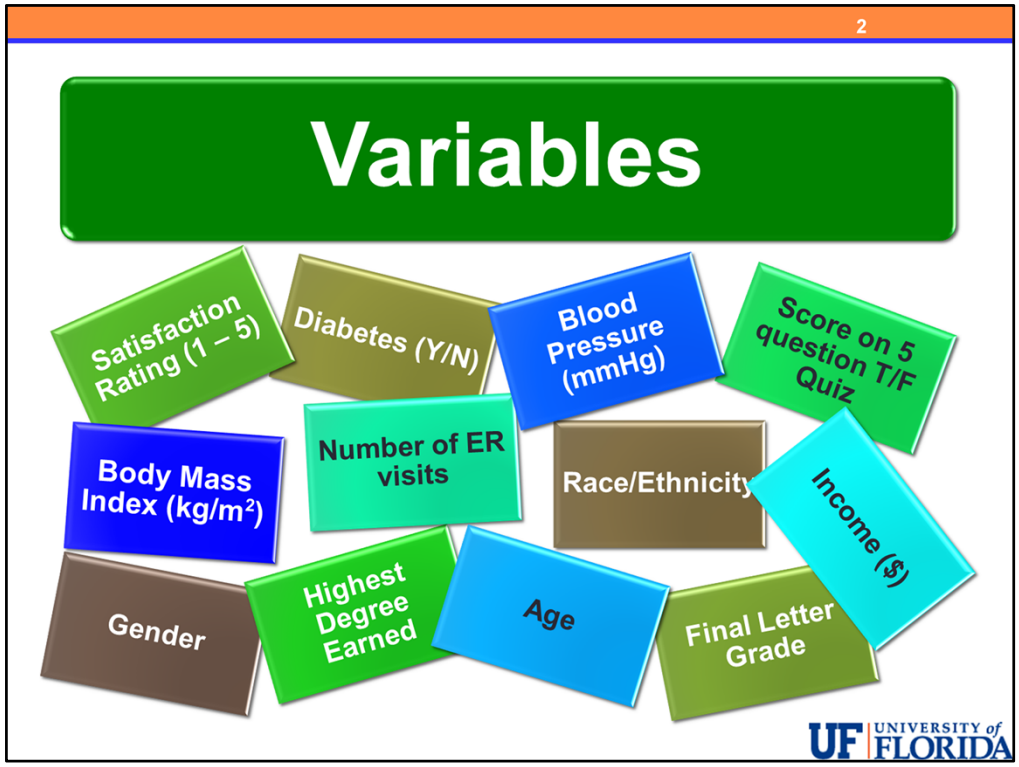
TYPES OF VARIABLES

UF UNIVERSITY of FLORIDA

Variables contain the information about a particular characteristic for all individuals in our dataset.

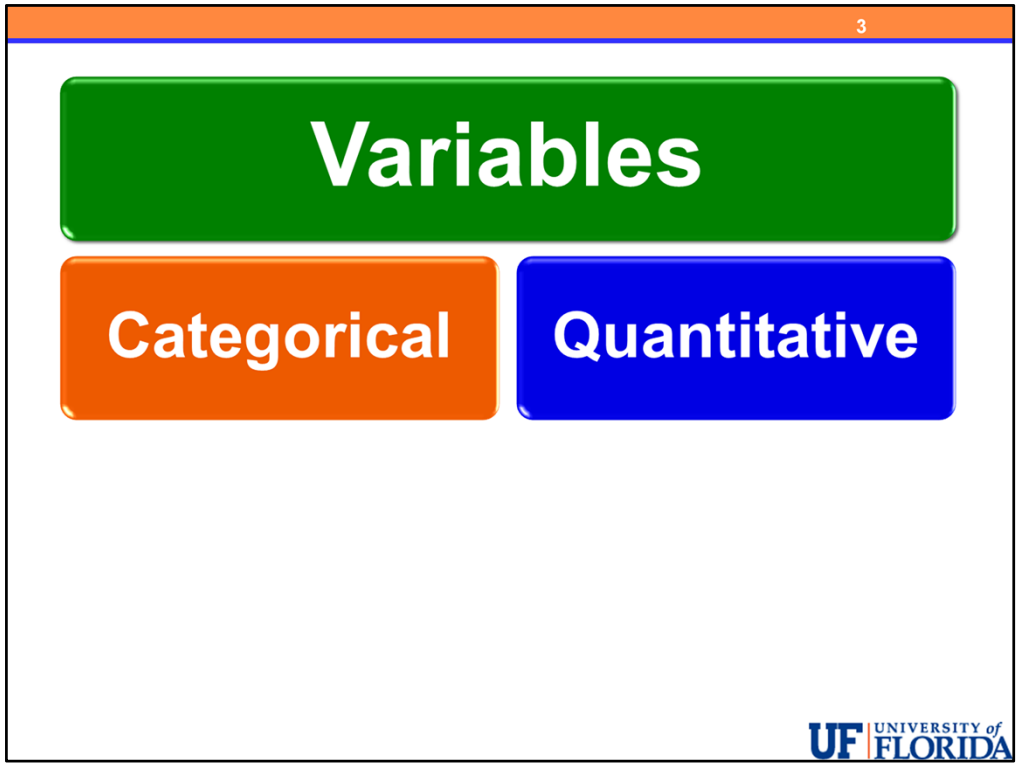
One of the most important skills in this course will be correctly classifying variables into one of two main types.



Here are some examples of variables.

Satisfaction ratings on a scale from 1 to 5.
Whether or not each individual has diabetes
Blood pressure measurements
Scores on a 5 question T/F quiz
Body Mass Index (or BMI)
Number of emergency room visits
Race/ethnicity for each individual
Income in dollars
Gender
Highest degree earned
Age
Final letter grade

The list of possible variables we can record on people is endless, add to that the possibility of studying animals, plants, objects, etc. and you should be able to see at least one reason that statistics and biostatistics are important!



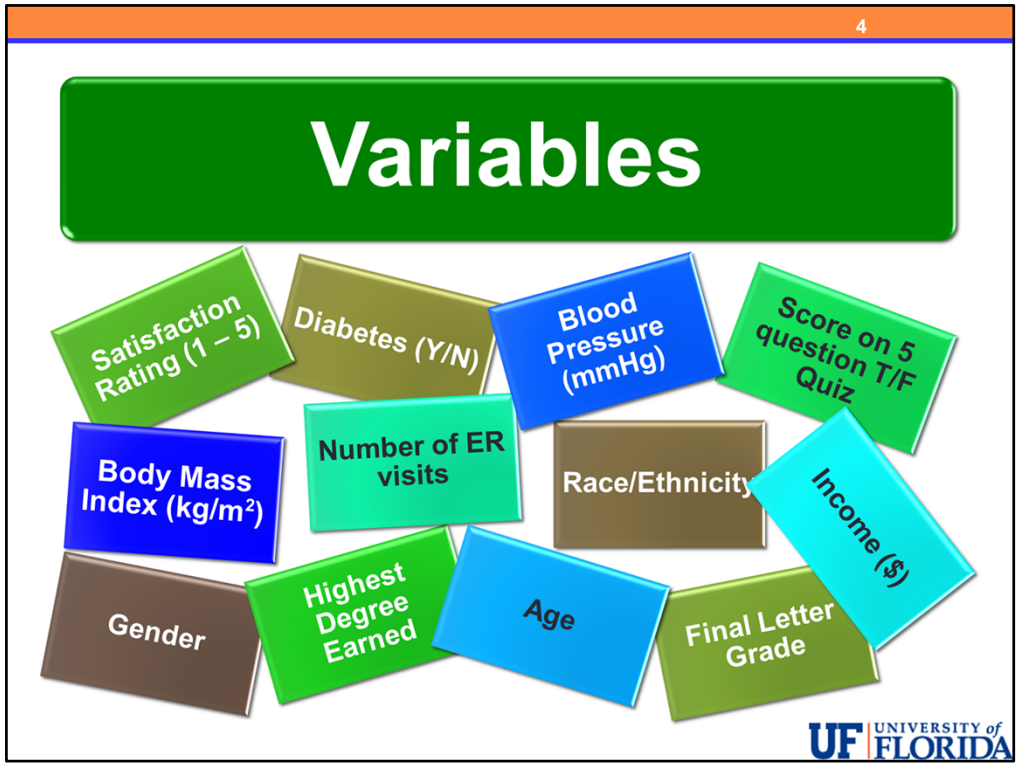
The two primary types of variables we will need to be able to distinguish between are those which are Categorical vs. those which are Quantitative.

Sometimes the terms used are Qualitative and Quantitative but I have found the similarity causes more problems than solutions for some students and in statistical circles, the term categorical is preferred.

Quantitative variables are true measurements or counts.

Categorical variables are NOT true measurements or counts but take on values which can only be categorized or grouped.

Let's return to our variables and determine whether each is categorical or quantitative.



Here are our variables.

Satisfaction ratings on a scale from 1 to 5.
Although these are represented by numbers, they do not represent a count or true measurement. Such ratings are categorical.

Whether or not each individual has diabetes
This is clearly categorical as there is no count or measurement and individuals are classified into Yes or No.

Blood pressure measurements
Here we have a true measurement. The units, millimeters of mercury, are clearly provided. This is a quantitative variable.

Scores on a 5 question T/F quiz
The scoring is not clearly defined but lets assume each question is worth a certain number of points (say 2 points each). Then, our scores do represent a sort of numeric measurement in that it gives the number of points the student scores out of 10 points and also, as long as the questions were equally weighted, we would have access to the number of questions the student answered correctly which most certainly would be a count. Under those assumptions (equally weighted points-based scoring), I would classify this as a quantitative variable.

Body Mass Index (or BMI)
BMI in this case is a measurement and thus would be quantitative. Sometimes, BMI is presented classified into categories such as underweight, normal, overweight, or obese, in which case it would be a categorical variable.

Number of emergency room visits
Here we have a count variable. The phrase "the number of" is usually a give away unless it is some other type of "number" such as a social security number, phone number, identification number, or similar.

Race/ethnicity for each individual
Would be categorical as we can only classify into groups. With genetics, there are some "measurements" which may be taken but here we simply mean the standard type of classification that we are often asked on surveys and forms.

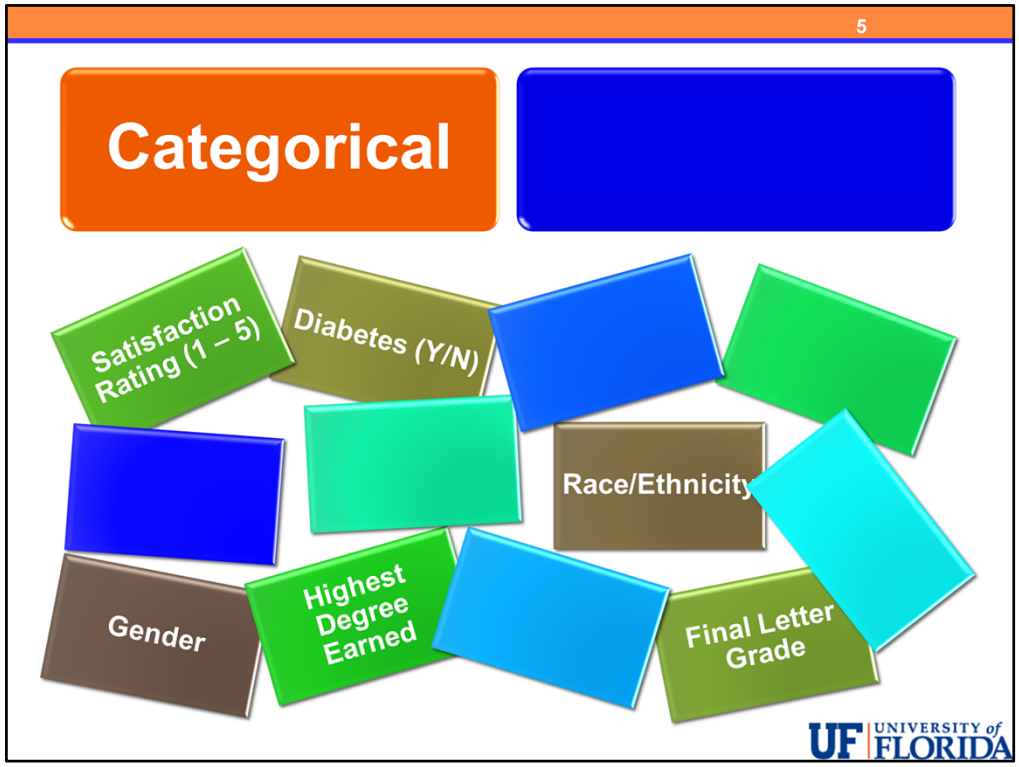
Income in dollars
Would be Quantitative

Gender
Categorical

Highest degree earned
Since it lists the highest degree (not the number of years) this would be categorical.

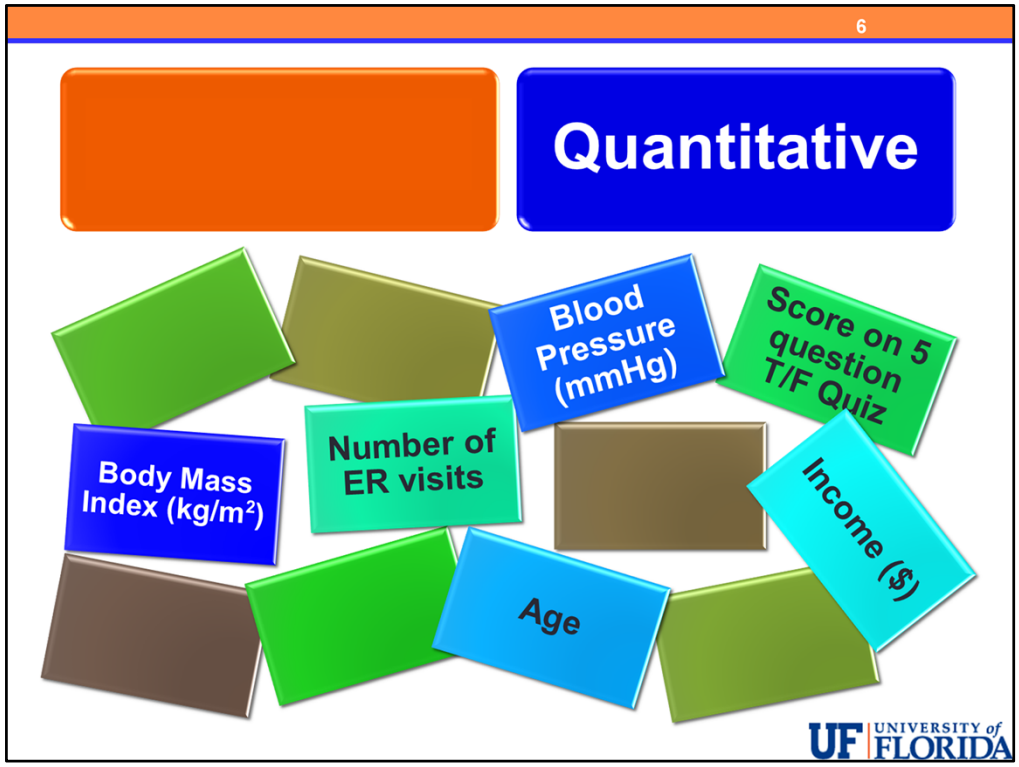
Age
Let's assume that we have the individuals age in years at a minimum (not a grouped classification of any kind). For age in years, the variables would be quantitative, however, in the description, the units should have been provided!!!

Final letter grade
Would be categorical



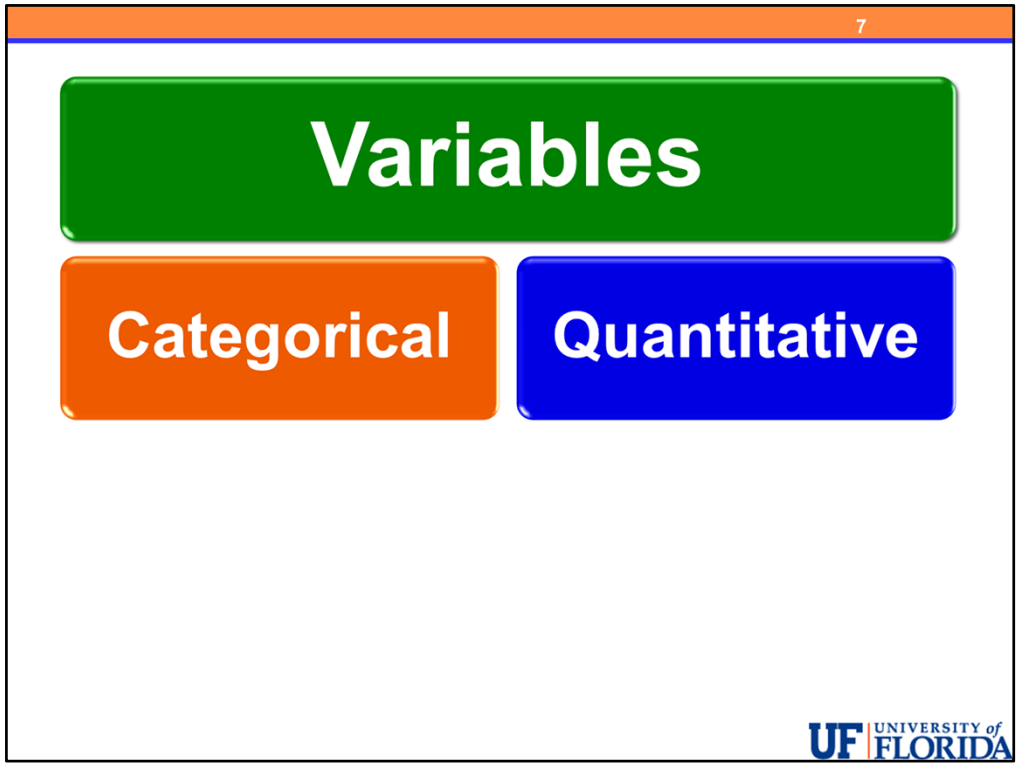
Here are our categorical variables.

Satisfaction ratings on a scale from 1 to 5.
Whether or not each individual has diabetes
Race/ethnicity for each individual
Gender
Highest degree earned
And
Final letter grade

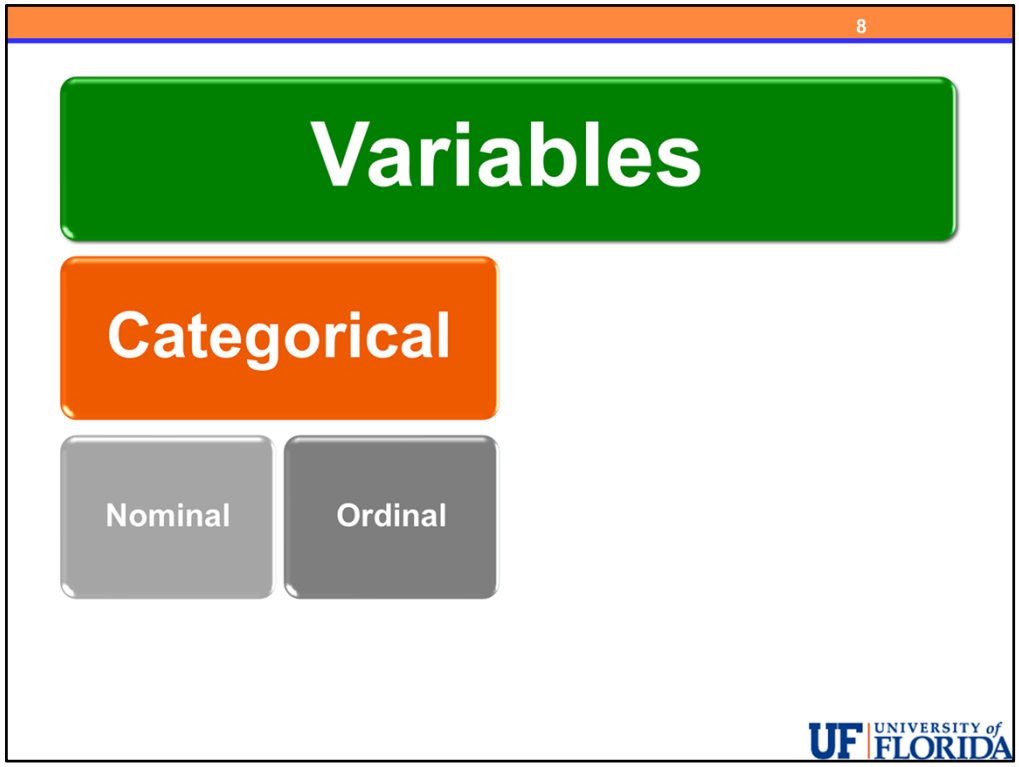


And here are our quantitative variables.

Blood pressure measurements
Scores on a 5 question T/F quiz
Body Mass Index (or BMI)
Number of emergency room visits
Income in dollars
And
Age



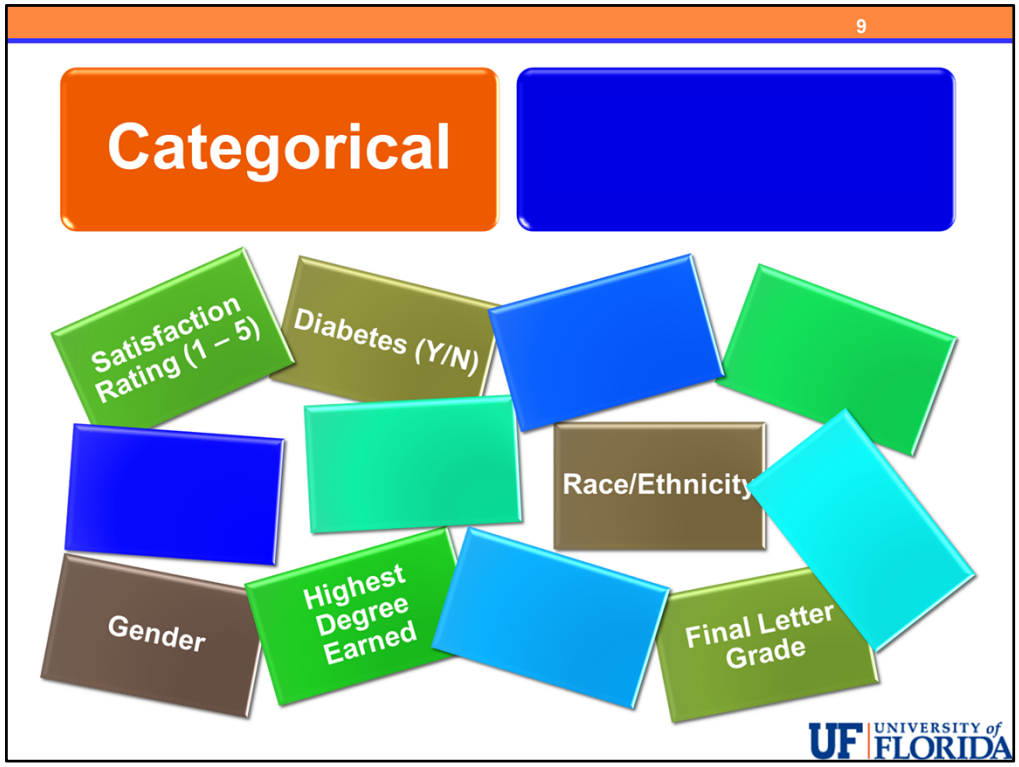
The primary sub-classifications we will use this semester result in two additional choices for each of the main types of variables.



For categorical variables, the two sub-classifications are nominal and ordinal.

Ordinal categorical variables have some natural ordering among the categories although differences are not precisely meaningful (due to the categorical nature of the data).

Nominal categorical variables do NOT have a natural ordering among the categories.



Here are our categorical variables again. Now lets determine which are ordinal (have a natural order) and which do not (nominal).

Satisfaction ratings on a scale from 1 to 5.
Would be ordinal as there is a natural ordering for each individual on the scale from 1 to 5.

Whether or not each individual has diabetes
Would be nominal as there is no natural order, only categorized into Yes or No

Race/ethnicity and gender are also nominal

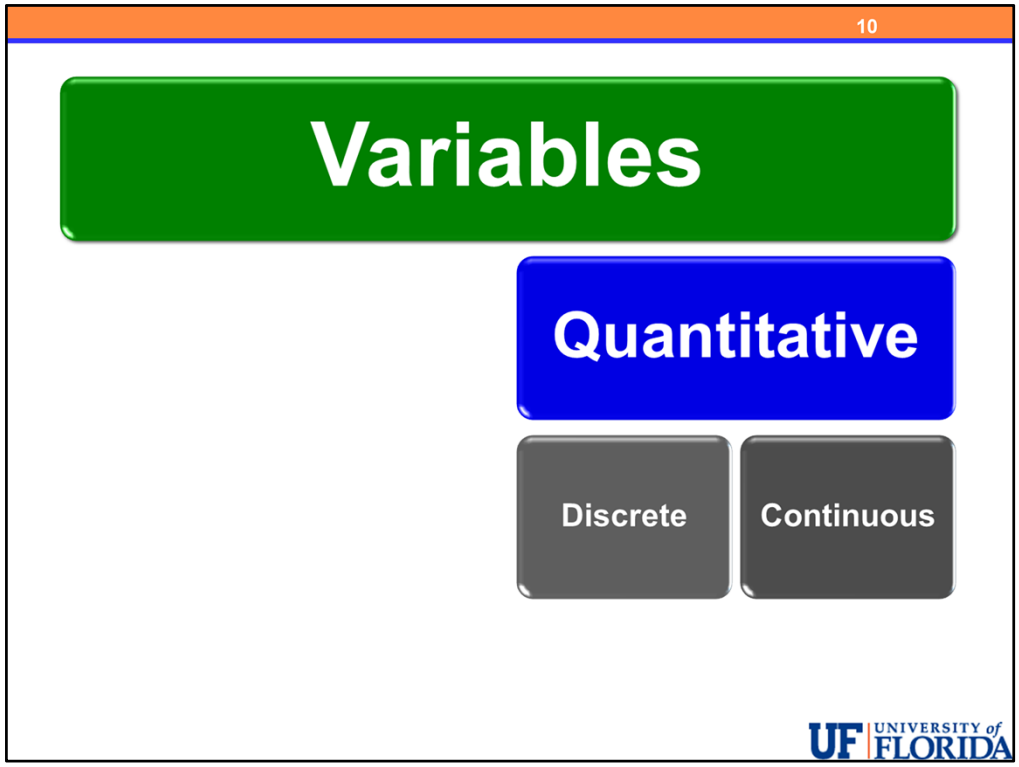
The highest degree earned could be ambiguous so let's assume it is generic in that it provides a list such as
Did not graduate high school
Graduated high school
Associates degree (or other 2 year degree)
Bachelors degree
Masters degree
Higher than a masters degree

In such case the highest degree earned and final letter grade would both have a natural order and would be classified as ordinal.

Remember that nominal and ordinal are only choices for categorical variables.

One special type of variable occurs when there are only two levels, although this could be a count (0 or 1), these variables are usually categorical in nature and are called binary or dichotomous variables.

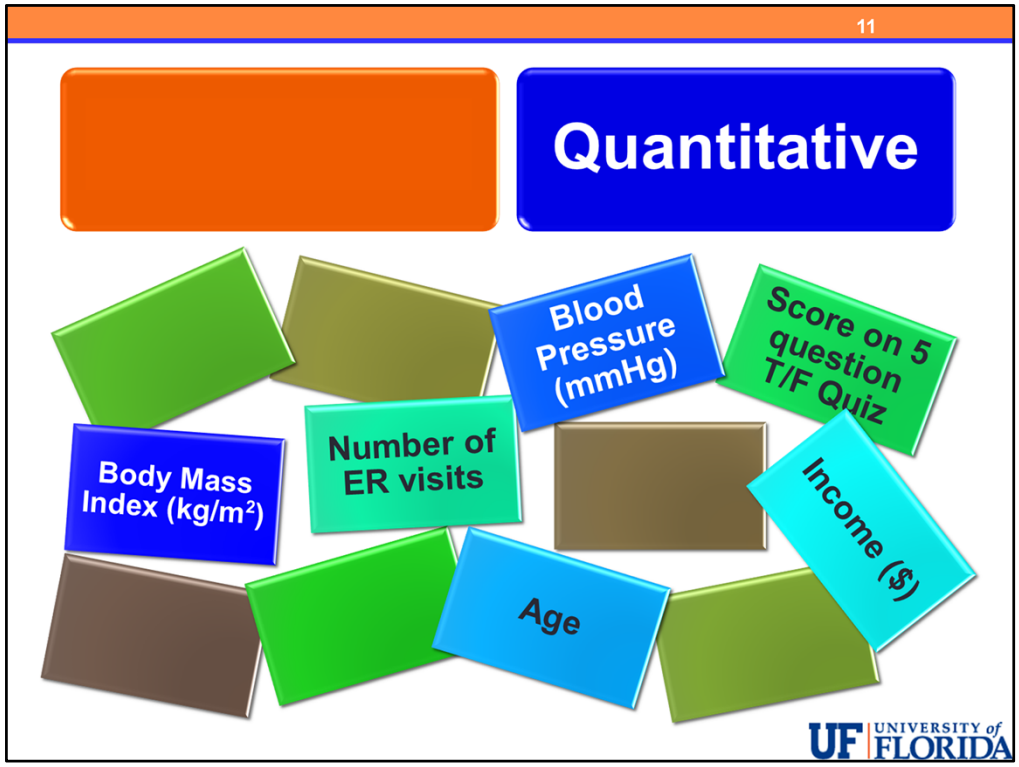
In our set of categorical variables above, Diabetes (Y/N) and Gender would be examples of binary variables, since there are only two possible levels.



For quantitative variables, the two sub-classifications are based upon the level of measurement.

Discrete quantitative variables can take on only a countable number of values, there are gaps between the possible choices. These are usually count variables of some kind.

Continuous quantitative variables can take on any value in an interval. Only our lack of need or lack of technology decides how precise we will or can measure these variables. Units should be provided with any analysis of continuous quantitative variables.



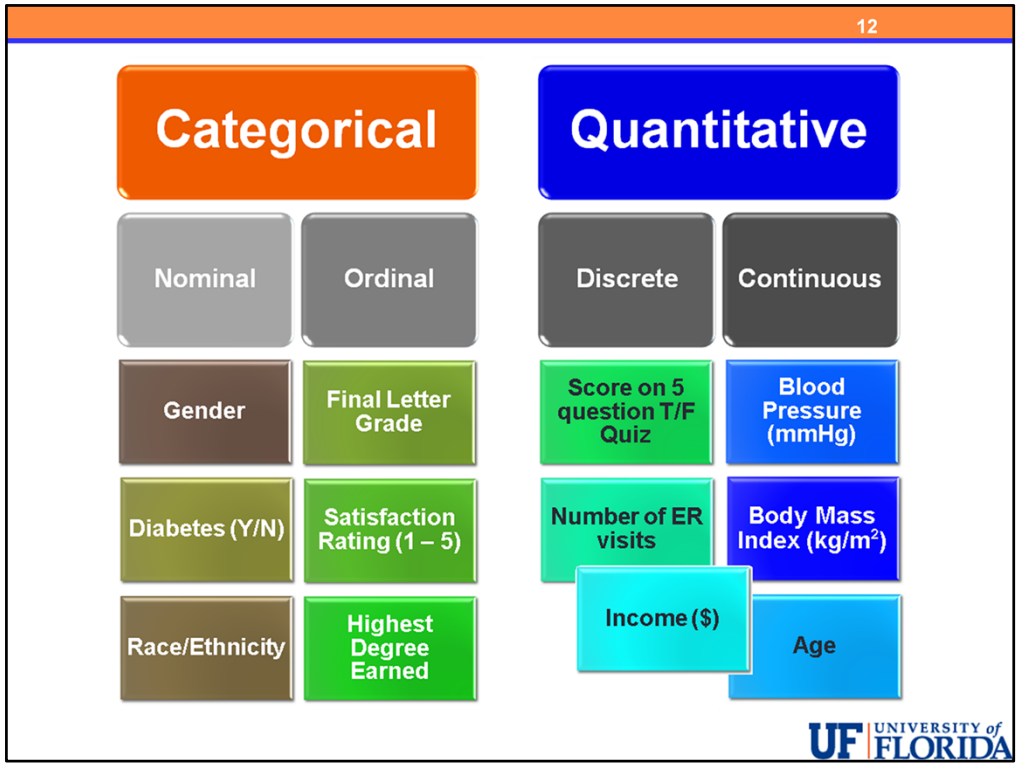
Here are our quantitative variables again. Let's determine which are discrete (think countable) or continuous (think true measurement)

In our set above, blood pressure measurements and Body Mass Index (or BMI) would be clearly continuous. These are true measurements, limited only by our desire or ability to measure more precisely.

Age and Income in dollars would most likely be treated as continuous variables in most analyses if there were a large number of possible values. Age is technically a continuous measurement since it is lack of desire for precision that usually produces "age in years."

Income however, could be technically considered to be discrete as it is almost never divided beyond one tenth of one penny!

And finally, the Number of emergency room visits and Scores on a 5 question T/F quiz would be discrete as these have a countable number of possible values.



Here is a quick review.

There are two main types of variables

Categorical and Quantitative. The easiest way to distinguish these is by determining if the variable represents a measurement or count, in which case it is quantitative, or if it does not. If it does not represent a measurement, it is categorical, in which case, be sure to think carefully about how individuals are grouped into categories.

Categorical variables can be sub-classified as nominal or ordinal with ordinal variables have a natural ordering, whereas nominal variables do not.

In our examples

Gender, Diabetes and Race/Ethnicity were nominal categorical variables, they have no natural order and individuals can only be put into categories for these variables.

Final letter grade, satisfaction rating, and highest degree earned were ordinal categorical variables, in that there some natural ordering yet they still do not represent a true measurement or count. Highest degree earned could be nominal if there turned out not to be a natural ordering in the choices provided, for example different types of masters degrees or Ph.D., M.D., D.V.M., etc. as these may not actually have a natural ordering!.

For quantitative variables, we can sub-classify them as discrete (countable) or continuous (can take on any value in an interval).

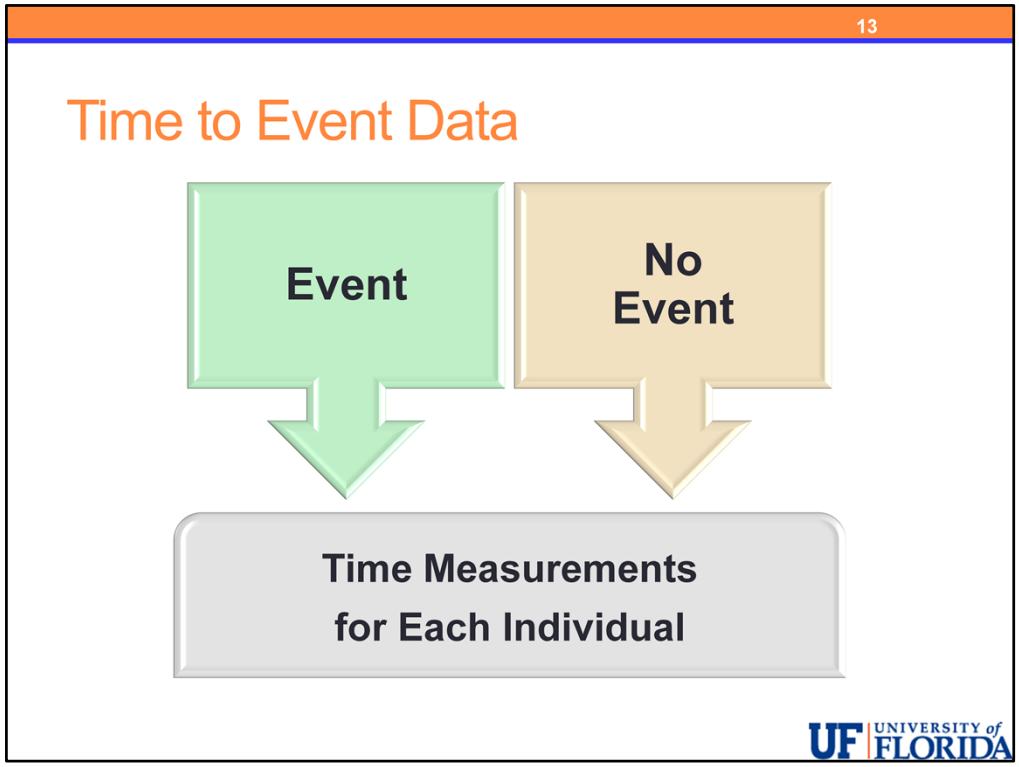
Our obvious continuous quantitative variables were blood pressure, body mass index, and likely age (but this was not clear without the units!).

Our obvious discrete variables are the score on a 5 question true-false quiz and the number of ER visits.

Income, although you may consider it to be technically discrete, would likely be treated as a continuous variable.

Other discrete variables (such as the number of ER visits per year for a sample of hospitals) may also be treated as continuous even though they are technically discrete.

The four sub-classifications will be more important to us later in the course but we may mention them as we go along when appropriate.

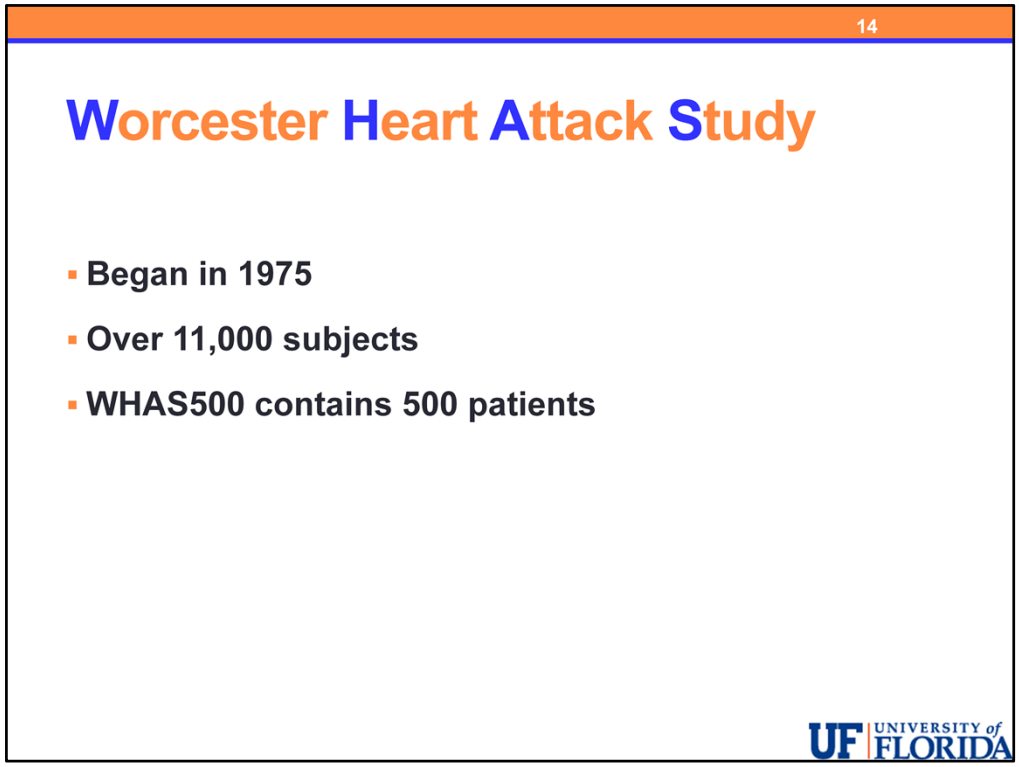


Although we will not cover this specifically in our course, it is worth noting another special type of data which, by design, actually records two variables simultaneously. When our response of interest is a time to event we record

Whether an event occurs
and
A Time, either to the occurrence or to the last follow-up without an occurrence

One must be careful in such a case to realize that the time to event measurement actually requires the status variable, whether or not the event occurs, in order to determine the meaning of the time variable for each individual. In this course, the best we could do is study each group of individuals separately with respect to any time-to-event measures.

We will look at the variables in such a dataset quickly for a little more practice.



The following list of variables is contained in a dataset from the WHAS study which was a study on the survival times of heart attack patients in an area of Massachusetts.

It began in 1975, and may still be continuing.

There are a large number of subjects, however, the dataset we have contains 500 patients and the following variables.

Variable Information

Variable	Description	Codes / Values
id	Identification Code	1 - 500
age	Age at Hospital Admission	Years
gender	Gender	0 = Male, 1 = Female
hr	Initial Heart Rate	Beats per minute
sysbp	Initial Systolic Blood Pressure	mmHg
diasbp	Initial Diastolic Blood Pressure	mmHg
bmi	Body Mass Index	kg/m^2
cvd	History of Cardiovascular Disease	0 = No, 1 = Yes
afb	Atrial Fibrillation	0 = No, 1 = Yes
sho	Cardiogenic Shock	0 = No, 1 = Yes
chf	Congestive Heart Complications	0 = No, 1 = Yes

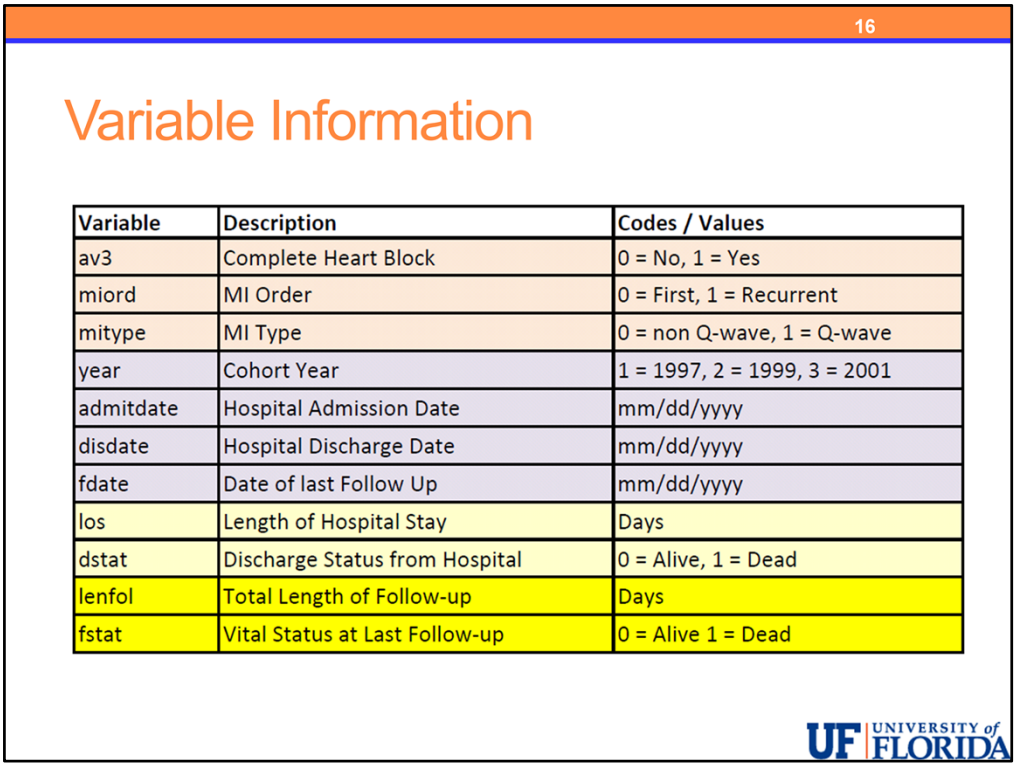
UF UNIVERSITY of FLORIDA

We will classify the variables as categorical or quantitative.

Age in years – quantitative
Gender – categorical
Heart rate in number of beats per minute – quantitative
Systolic and Diastolic blood pressure – quantitative
BMI – quantitative.

Then we have a few yes/no variables which are categorical
History of cardiovascular disease
Atrial fibrillation
Cardiogenic shock
Congestive heart complications

On the next slide we have a few more variables



Another few categorical variables in
Complete heart block (yes or no)
MI order (first heart attack or recurrent, meaning this was not your first!)
MI Type (q-wave or non q-wave)

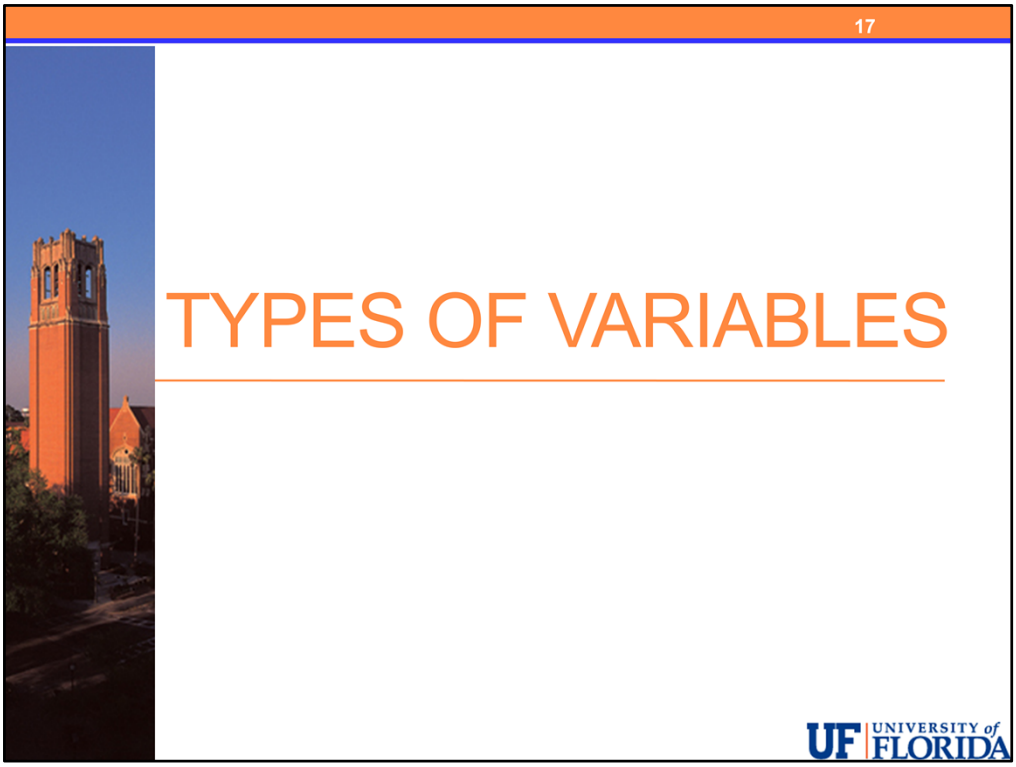
The cohort year (which is not an interesting variable to analyze) would also be categorical, however, it would be ordinal, it and the MI order variable are the only two ordinal categorical variables in the dataset.

We then have a few dates which are used to calculate two time-to-event measures. Which are color coded the same as the status variable for that time.

We have length of hospital stay in days as the first time-to-event – it is the time to discharge from hospital.
But in order to know what this means we must look at the DSTAT variable which tells us whether the patient was alive or dead at that time.

The second time-to-event is the length of follow-up which records the status variable FSTAT as the vital status at last follow-up, again, alive or dead.

In our course, we may return to this dataset to illustrate software skills and output, however, generally we will use only the variables listed through MI Type and avoid any concerns of dealing with the time-to-event outcomes.



Be sure return to this material and ask questions whenever you find yourself having a difficult time choosing between the two main types of variables.

This is an essential skill to be effective at statistical analysis. You will soon see that all of our decisions about how to summarize and analyze data begin with the answer to the questions “What types of variables do I have?” and “What role do they play in my research question?”