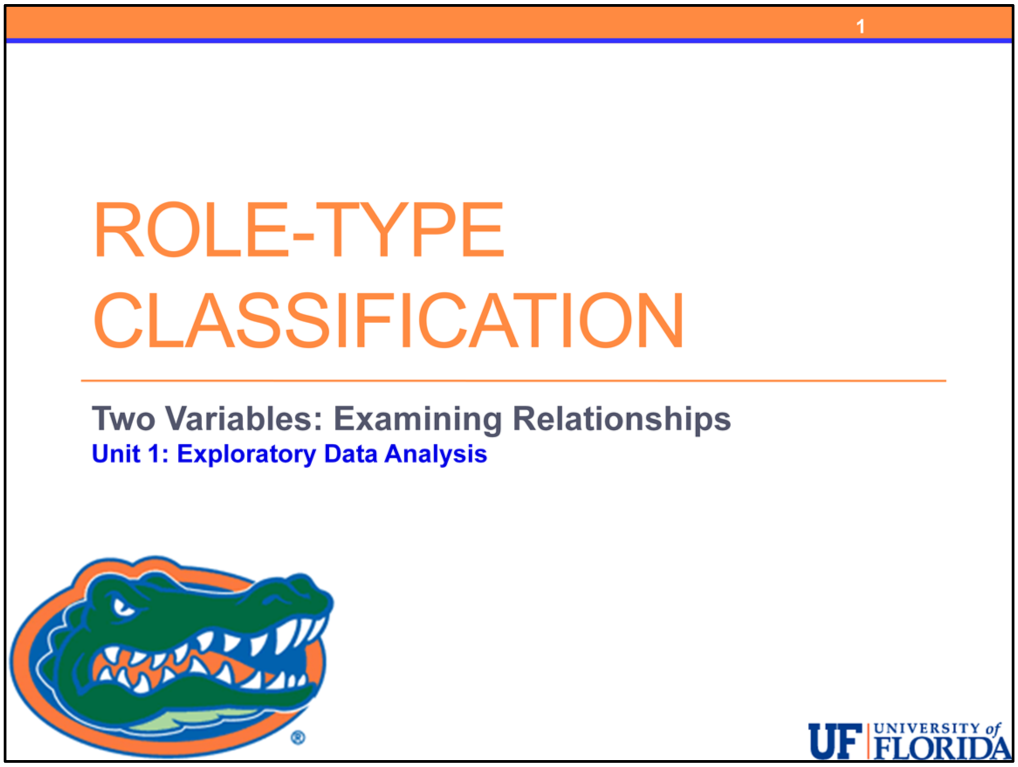


1

ROLE-TYPE CLASSIFICATION

Two Variables: Examining Relationships

Unit 1: Exploratory Data Analysis

UF UNIVERSITY of FLORIDA

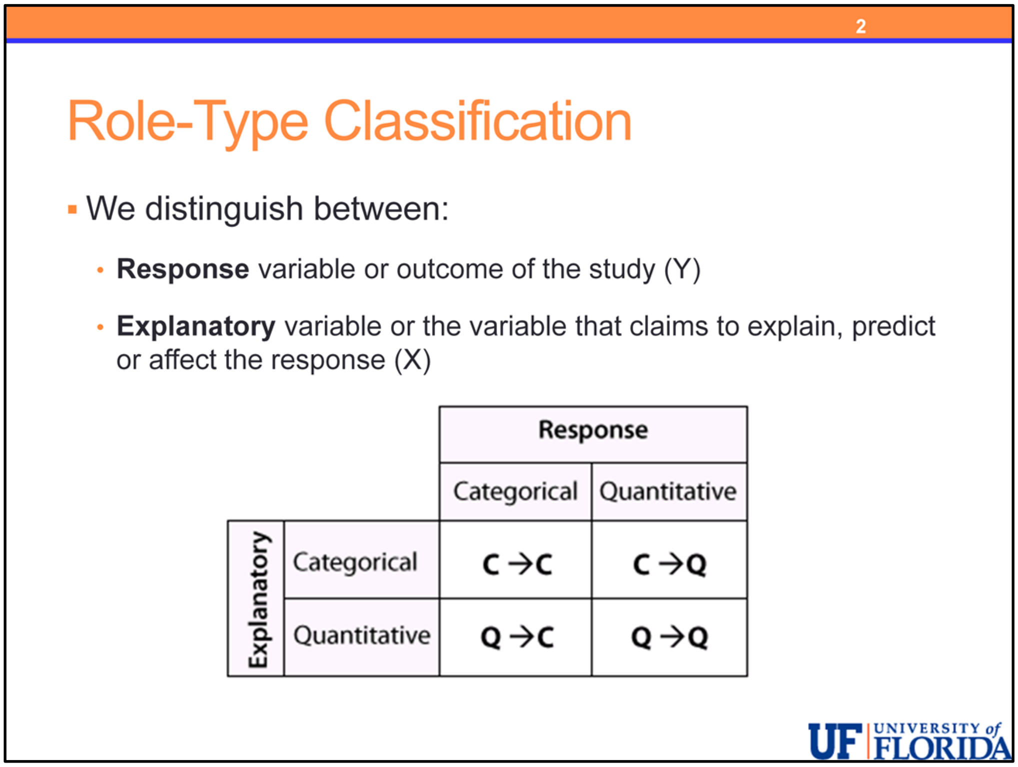
Now we begin our discussion of exploratory data analysis for examining relationships between two variables.

Our first task is to define the “Role-Type classification” which plays a crucial role in determining the correct methods - both for exploratory data analysis and later for inferential statistics.

There are two components to this classification – the ROLE and the TYPE

We already know that variables can be classified by their type (categorical or quantitative) and that the types of variables are important in choosing the appropriate methods.

In most of the studies that you do in real-life, each of the variables usually has a specific role to play in your analysis. We briefly mentioned this in the preliminaries – “What is data?” and will now elaborate further and discuss a few examples.



Role-Type Classification

- We distinguish between:
  - **Response** variable or outcome of the study (Y)
  - **Explanatory** variable or the variable that claims to explain, predict or affect the response (X)

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C →C	C →Q
	Quantitative	Q →C	Q →Q

We distinguish between the response variable, or outcome of the study and the explanatory variable. The response variable is what we are interested in learning more about or proving something about.

Then, usually, we are trying to determine if there is an effect on or difference in the response for different values of the explanatory variable.

The explanatory variable claims to explain, predict, or affect the response variable.

The explanatory variable we usually denote X and the response variable or outcome, we usually denote Y.

In order to determine how to proceed we need to combine the information about the Role of the variable with the information about the types of variables.

We can have two categorical variables. Case C-C.

We can have a categorical explanatory variable being used to predict or explain a quantitative response variable. Case C-Q – this case we have already mentioned in our discussion of boxplots.

In the reverse order of Case Q-C, the methods we present for Case C-Q can be used, however, we will never get to the point in this course where we can literally “predict” the value of the response variable in Case Q-C. Formally, that analysis would be logistic regression.

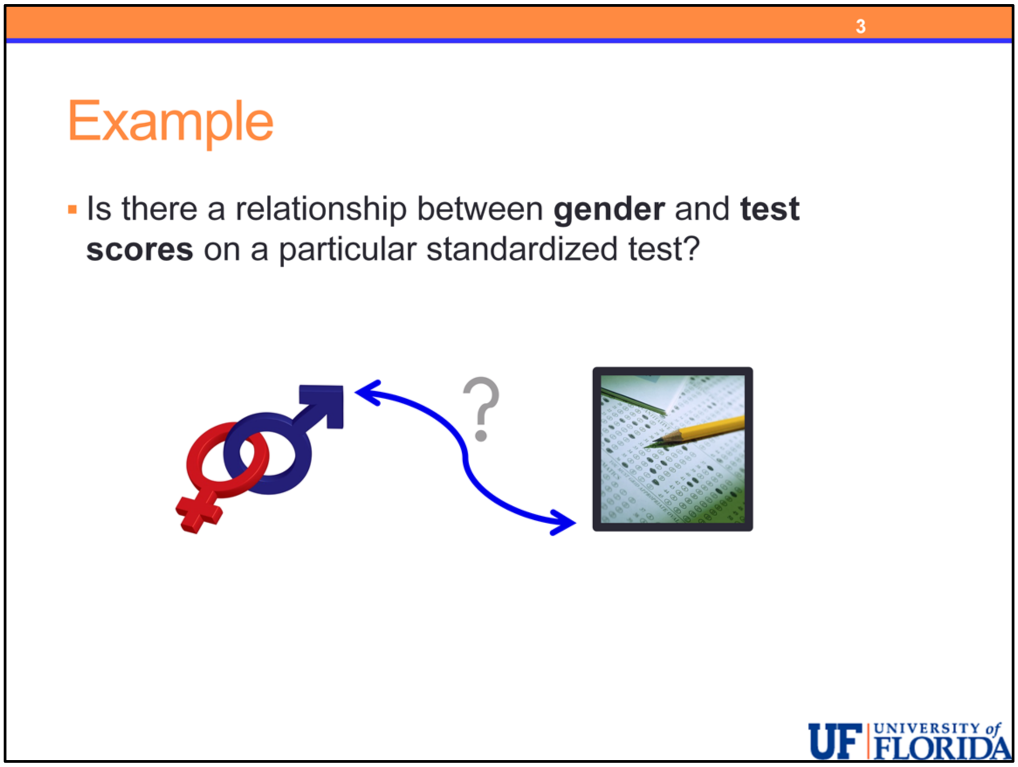
And finally, we can have two quantitative variables. Case QQ - and there is a large amount of new information to discuss in this case.

In any problem involving two variables you need to determine:

What TYPES of variables do you have? Categorical or Quantitative … AND

What is their ROLE in your analysis? Which one of the two variables is the response variable and which is the explanatory variable?

Now let’s look at a few examples.



3

Example

- Is there a relationship between **gender** and **test scores** on a particular standardized test?

Is there a relationship between **gender** and **test scores** on a particular standardized test?

We could also say:

- Is performance on the test related to gender?
- Is there a gender effect on test scores?
- Are there differences in test scores between males and females?

We want to explore whether the outcome of the study — the score on the test — is affected by the test-taker’s gender.

Therefore:
Gender is the **explanatory** variable
Test score is the **response** variable

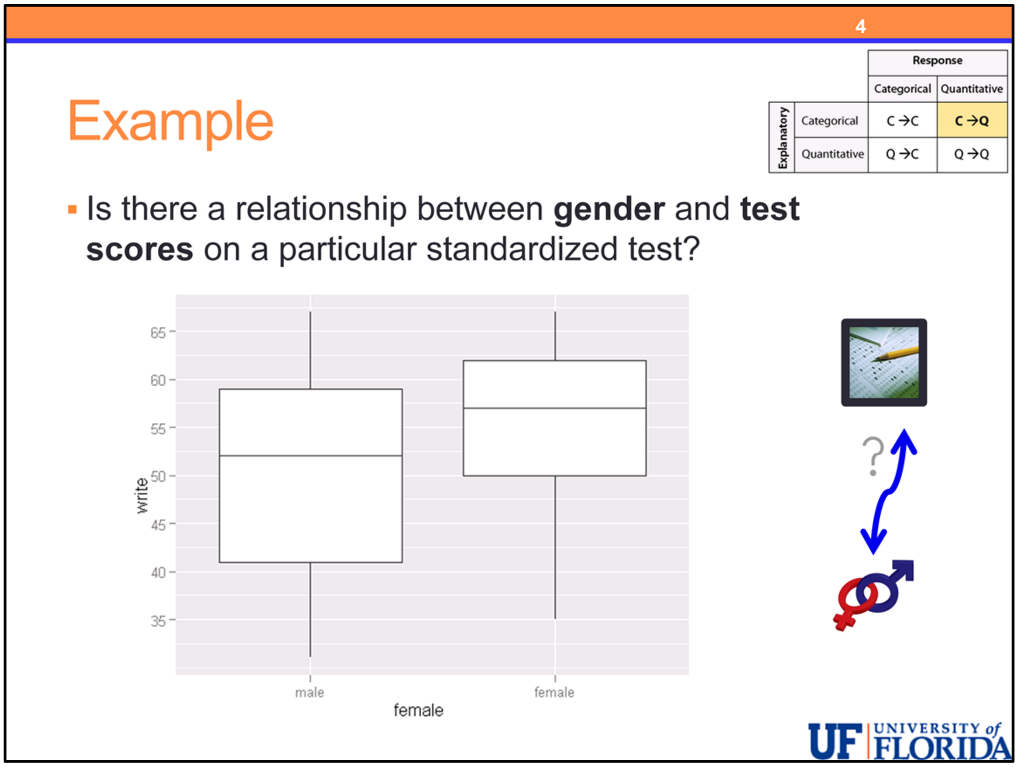
Whenever variables such as gender, age, race/ethnicity, etc. are one of the variables, it is almost always the case that these will be the explanatory variables in our analysis as we don’t expect to alter these values for individuals and we rarely have interest in predicting these variables in practice.

If the choice of explanatory and response variable for a particular question seems unclear, stop and consider if this kind of logic can help you decide.

We have determined the ROLE of the variables, now we need to determine the types of variables we have.

Gender is our **explanatory** variable and it is **Categorical**
Test score is the **response** variable and most likely we would agree that this will be **Quantitative**

Thus we are in **Case C-Q** where we have a categorical explanatory variable – Gender - used to explain or predict a quantitative response variable – Test Score.



4

Example

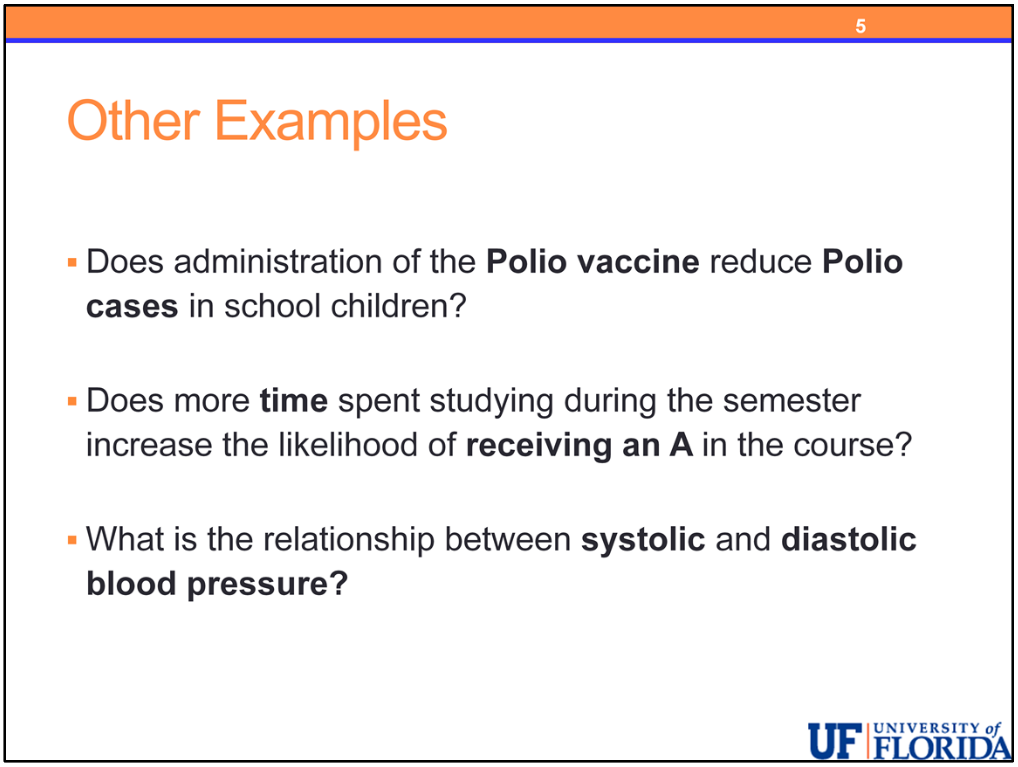
		Response	
		Categorical	Quantitative
Explanatory	Categorical	C →C	**C →Q**
	Quantitative	Q →C	Q →Q

- Is there a relationship between **gender** and **test scores** on a particular standardized test?

UF UNIVERSITY of FLORIDA

In our discussion on boxplots, we learned to compare a quantitative variable by groups using side-by-side boxplots. In Case C-Q this is how we can visually display the data.

How would you numerically summarize this data?



Other Examples

- Does administration of the **Polio vaccine** reduce **Polio cases** in school children?
- Does more **time** spent studying during the semester increase the likelihood of **receiving an A** in the course?
- What is the relationship between **systolic** and **diastolic blood pressure?**

UF UNIVERSITY of FLORIDA

Here we have a few more examples. One of each remaining Role-Type combination in fact.

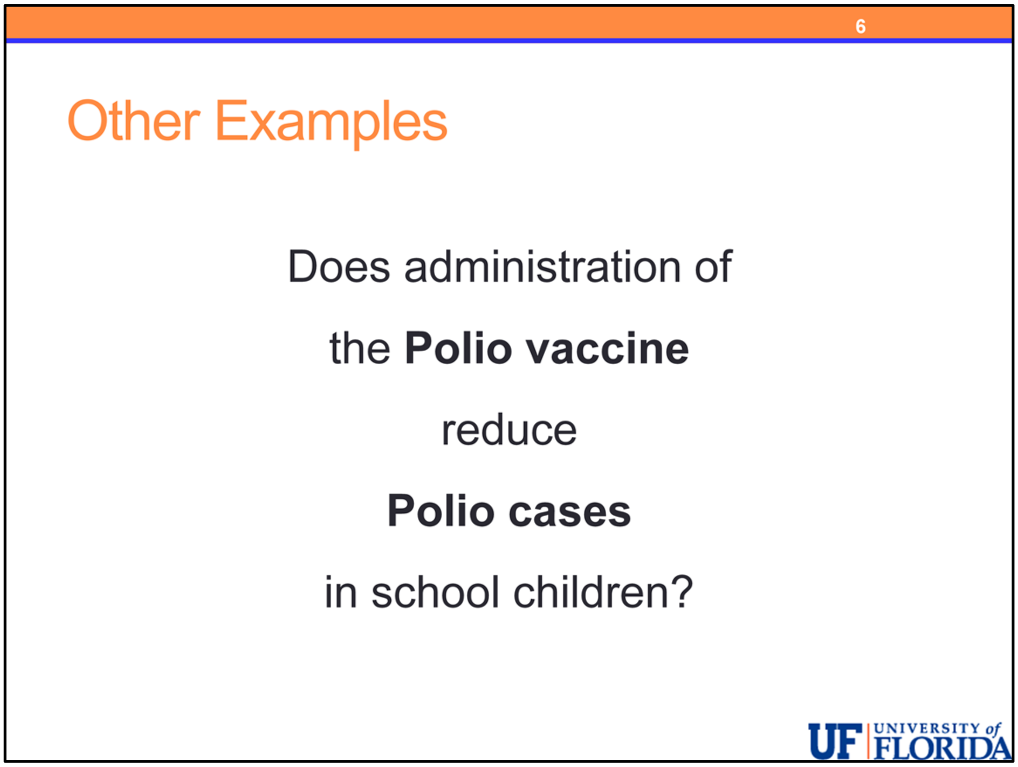
See if you can determine the answers as we read each scenario.

- Does administration of the Polio vaccine reduce Polio cases in school children?

- Does more **time** spent studying during the semester increase the likelihood of **receiving an A in the course**?

- What is the relationship between systolic and diastolic blood pressure?

Now let’s go through them one at a time.



6

Other Examples

Does administration of

the **Polio vaccine**

reduce

Polio cases

in school children?

In the first example, our two variables are:

- Was the student administered the Polio vaccine?

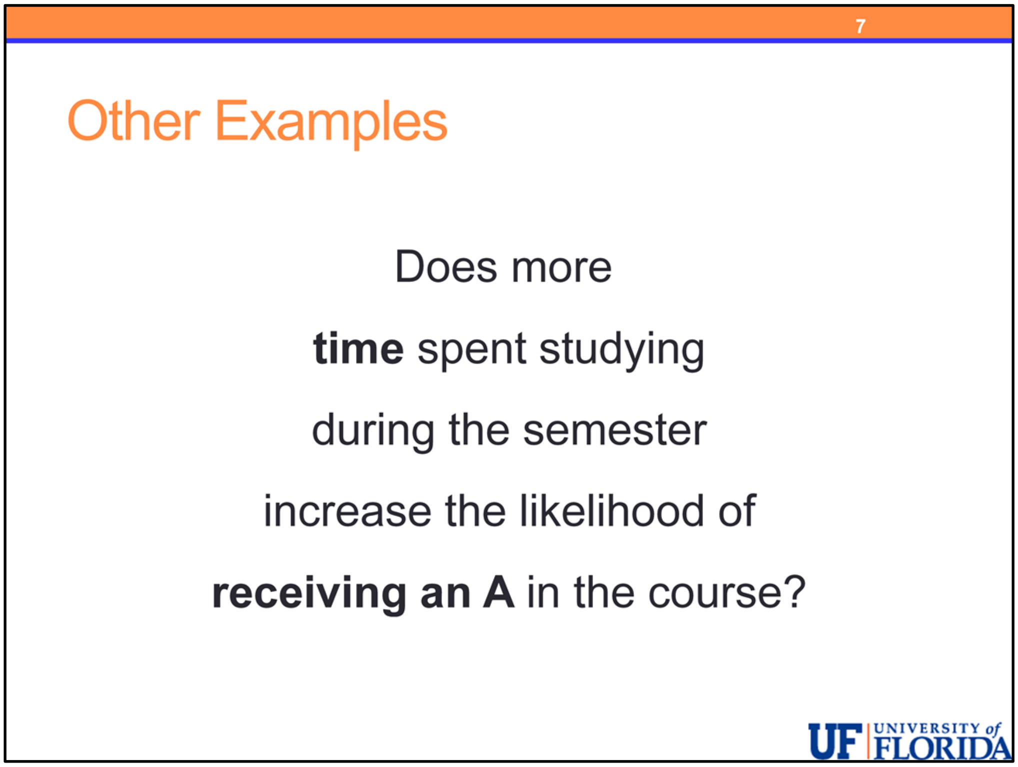
And

- Did the student develop polio (during the study period)?

Both of these variables are “Yes/No” variables and so are categorical. Therefore, regardless of the role each variable plays in the analysis, at this point, we know we are in Case C-C. However, it will be important to distinguish between the response and explanatory variables when we delve further into the methods involved in Case C-C.

In this case, clearly we intend to show that administration of the vaccine explains or predicts whether or not the student develops Polio. In particular, the study hopes to show that the vaccine reduces the chance of acquiring the disease. We will talk more about this example as well as the numerical measures and graphical displays which are appropriate in Case C-C.

What numerical measures do you think are appropriate? How would you graphically display the results of such a study?



Other Examples

Does more

time spent studying

during the semester

increase the likelihood of

receiving an A in the course?

In the second example, our two variables are:

- Time spent studying during the semester for a given student.

And

- Whether the student did or did not receive an A in the course.

We hope to address the question: Do students get A’s more often if they study more?

Let’s assume this is a perfect world and we can know the exact time (to the nearest hour) that each student studies during the whole semester. This “time” variable would be quantitative.

We wish to use this variable to predict the chance that a student will receive an A in the course, which is a categorical – Yes/No variable.

Thus the explanatory variable is the time spent studying and the response variable is whether or not the student receives and A in the course.

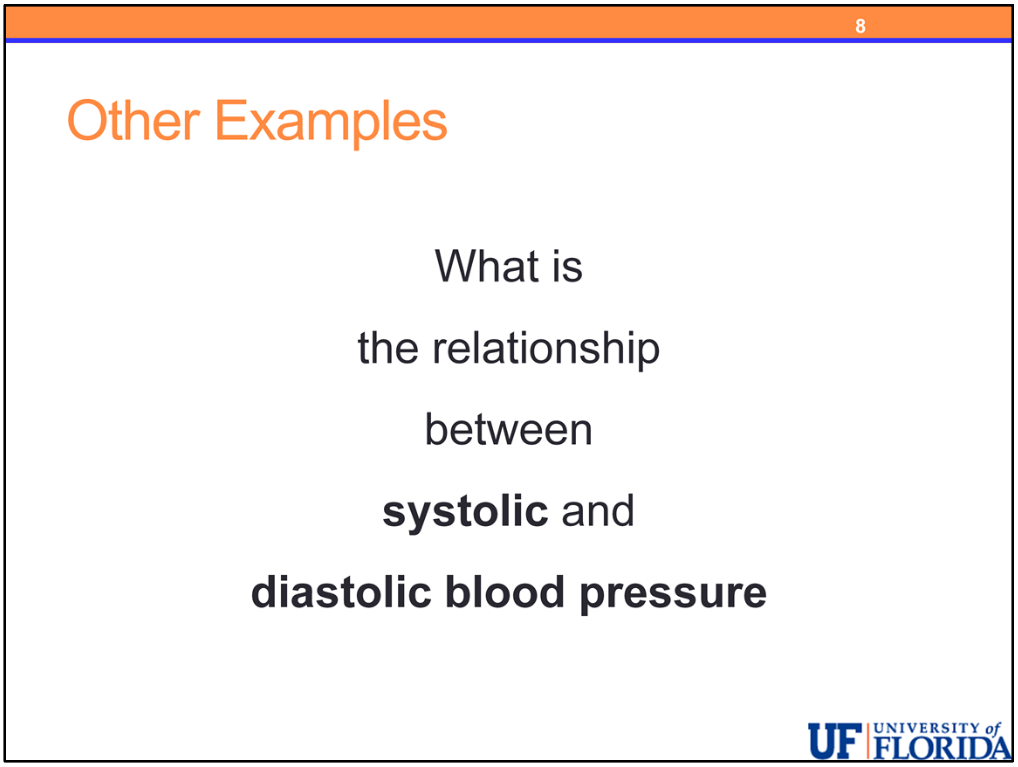
This is, therefore, Case Q-C. Although we will not specifically cover this case, we can determine whether the variables are related or associated using the methods discussed in Case C-Q, and this is often done in practice for some or all or the exploratory data analysis in this case.

One comment: Consider our question here which we have phrased two ways:

- Do students get A’s more often if they study more?
- Does more **time** spent studying during the semester increase the likelihood of **receiving an A** in the course?

Notice, although it may be easier for us to write questions in their logical order where the explanatory variable is mentioned first, it is entirely possible to rearrange any research question involving two variables and restate the questions in many different ways.

Keep this in mind and be certain to carefully consider both variables and their logical or stated role in the question or scenario.



Other Examples

What is

the relationship

between

systolic and

diastolic blood pressure

In the third example, our two variables are:

- Systolic blood pressure

And

- Diastolic blood pressure

Both of these variables are quantitative and thus we are in Case Q-Q. Again, however, we must consider the role the variables will play in our analysis when we make decisions about which analyses to use and how to carry them out.

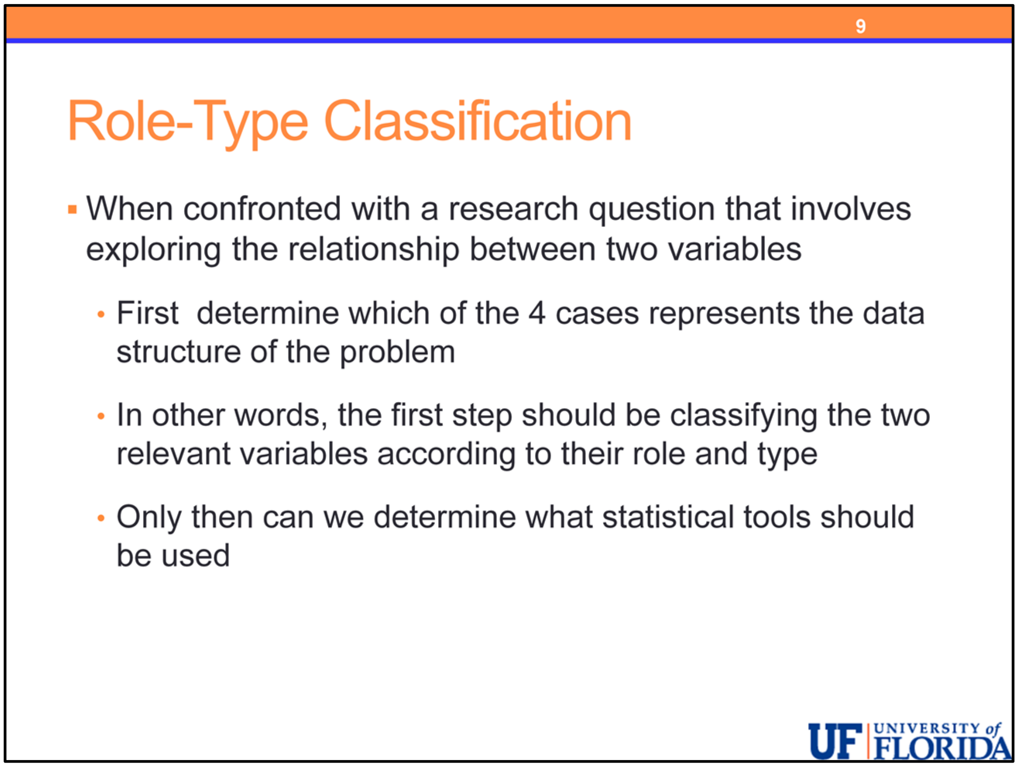
In this particular instance, nothing about the question or the variables tells us which will be the explanatory variable and which will be the predictor variable. In such a case, we really cannot give a “correct” answer to the ROLE portion of our classification. That is ok in all three main situations. Here, we know we are in Case Q-Q, which is all we need for the moment to continue our discussion!

Similarly, when we have two categorical variables, we will automatically be in Case C-C.

When we have one categorical and one quantitative variable, we could be in Case C-Q or Case Q-C, however, the type of exploratory data analysis we can apply in these two cases is the same in our course, and often in practice.

If you can distinguish which is the explanatory variable and which is the response variable then we may ask you to do so. If it is not possible for a given scenario, then we will not.

But we may still be able to ask questions about the types of analyses you might conduct and the types of questions you could answer from those analyses.



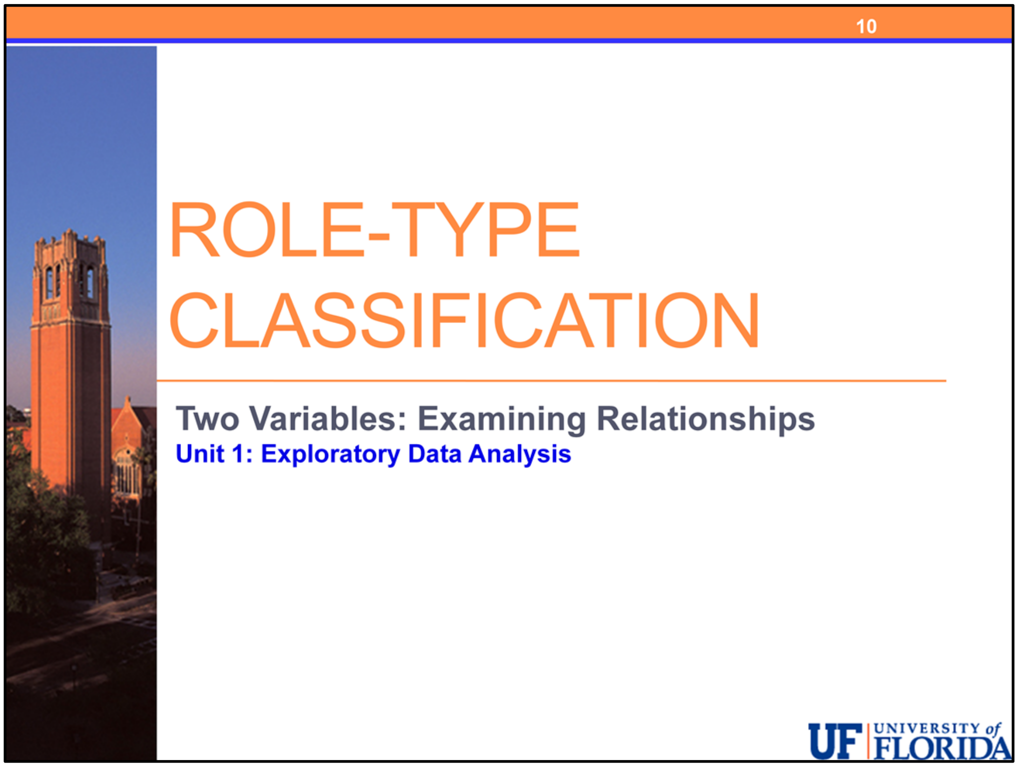
Role-Type Classification

- When confronted with a research question that involves exploring the relationship between two variables
  - First determine which of the 4 cases represents the data structure of the problem
  - In other words, the first step should be classifying the two relevant variables according to their role and type
  - Only then can we determine what statistical tools should be used

In summary, each time we are presented with a research question involving using one variable to explain or predict another variable, we need to determine which of the four cases represents our problem.

To do this we need to classify the two variables according to their ROLE (explanatory or response) and TYPE (categorical or quantitative).

Only after we know the Role-Type classification, can we make the right choice in terms of what statistical analysis we are going to conduct.



ROLE-TYPE CLASSIFICATION

Two Variables: Examining Relationships

Unit 1: Exploratory Data Analysis

You have seen the importance in knowing if a variable is quantitative or categorical before determining the appropriate exploratory data analysis for one variable scenarios. Now we embark on the specific visual displays and numerical measures for exploratory data analysis involving relationships between two variables.

And … being able to match the right kind of analysis to the data you have for a given scenario is - most definitely – a big part of what you are in this course to learn or improve upon!