Response and explanatory variables

StatPREP Class Lesson

# Orientation

A variable is a quantity or characteristic that varies from one person to another, or more generally, from one *unit of observation* to another. There are two distinct *types* of variables: quantitative (numeric) and categorical (labels/words/etc.).

There’s another distinction to be made about variables, which is not about the type of the variable itself but the *role* the variable will play in the statistical methods we use to describe a situation of interest and the relationships among the variables involved. This distinction is between the *response variable* and the *explanatory variables*.

Think about how we describe human relationships. For example, consider two women: Eliana and Rabia. One possible relationship: Eliana is Rabia’s aunt. Another possible relationship: Rabia is Eliana’s niece. As you know, these two relationships are exactly the same thing, just expressed differently. Each of the expressions involves a reference person, that is, a person with respect to whom the relationship word (“aunt”, “niece”) is used.

Similarly, when we describe a relationship between two variables, it’s helpful to consider one variable to be the reference. We describe the pattern of the other variable *with respect to* the first variable. Now phrases like “one variable,” “other variable,” and “first variable” are ambiguous. It’s hard to keep track of which variable is which. So, to simplify things, we identify one of the variables as the *response variable* and the other as the *explanatory variable*. Just as the two women can be related by either “aunt” or “niece”, the roles of the two variables can be described as “Response is a *function* of Explanatory” or “Explanatory *determines* Response.”

How to decide which variable should play the role of the response and which the explanatory? There’s no absolute rule; generally you can describe relationships either way, keeping in mind that the statement of the relationship (“aunt”, “niece”) will depend on which is which. Here are some rules of thumb.

* If you are trying to *predict* the value of one variable by another, call the variable to be predicted the *response* variable. Example: Price of a rental apartment as a function of the city in which it’s located.
* If you believe that one variable *causes* the other variable, call the variable being caused the *response* variable. Example: Risk of cervical cancer as explained by whether or not the person got a human papilloma virus (HPV) vaccine.
* If there’s an *outcome* you’re interested in, make that the *response* variable. Example: Test scores as explained by socio-economic status.

Most statistics books have traditionally been written without making much use of the concepts of response and explanatory variables. Mathematicians, in particular, seem to like to talk about “bivariate relationships” and to use statistics like the correlation coefficient that are the same whichever order you put the variables in. But most two-variable statistics – for instance the slope of a regression line, or the difference between two proportions – depend on the order of the variables. As such, it’s helpful to be clear indicating the particular variable you want to predict or that’s being caused, or that you have a particular interest in.

And, when there are more than two variables, it’s particularly helpful to distinguish between the single response variable and the multiple explanatory variables (which are sometimes called *covariates* or *confounders*.)

# Activity

There are two main types of variables: quantitative and categorical. Strickly as a matter of logic, there are four possible ways that these two types can be arranged as the response and explanatory variables. It’s important to know this, since the choice of an appropriate statistical technique should be shaped by the types of the response and explanatory variables.

|  |  |  |  |
| --- | --- | --- | --- |
| Response | Explanatory | Statistical technique | Little App |
| Quantitative | Quantitative | linear regression | [linear regression](https://dtkaplan.shinyapps.io/LA_linear_regression/) |
| Quantitative | Categorical | group-wise means | [t-test](https://dtkaplan.shinyapps.io/LA_t_test/) |
| Categorical | Quantitative | proportion regression | [proportions](https://dtkaplan.shinyapps.io/LA_proportions/) |
| Categorical | Categorical | group-wise proportions | [proportions](https://dtkaplan.shinyapps.io/LA_proportions/) |

Depending on where you are in you statistics course, you might have not yet encountered one of these techniques or the other.

Here are a few pairs of variables:

## Some pairs of variables

These variables are found in the data available through the Little App. For simplicity, we’re using only those categorical variables that have just two levels.

* Births\_2014
  1. Whether the mother is covered by the WIC program and the age of mother.
  2. The age of the mother and of the father.
  3. The length of gestation and the baby’s weight at birth.
* NHANES
  1. Systolic blood pressure and sex.
  2. Diabetes and age
  3. Weight and body mass index (BMI)
  4. Income and home\_type
* diamonds
  1. The price of a diamond and its weight in carats.
  2. The weight of a diamond and its clarity.

Answer the following questions for each of the 9 pairs of variables listed.

1. View the data with a point plot, using the [point-plot Little App](https://dtkaplan.shinyapps.io/LA_point_plot/). This will let you readily see which type each variable is. (The choice of one variable as response and one as explanatory isn’t critical here, since the reverse choice would generate the same plot but turned around the diagonal to reverse the axes.)

* Use the point-plot Little App to visualize the relationship. Characterize it as “strong”, “moderate”, “weak” or “none”. You don’t need to give a more detailed description. We just want you to decide whether you can see a clear relationship or not.
* *For each of the nine variable pairs, characterize the relationship as ‘strong’, ‘moderate’, and so on.*   .  .  .

1. In order to give a detailed description of the relationship, or sometimes in order to detect any relationship at all, you need to calculate appropriate statistics on the data. For this, you have to pick the appropriate Little App, which often means that you need to designate one variable as the response and the other one as the explanatory variable. (You can refer to the table above.)

* For each of the nine pairs of variables, make an appropriate choice of response and explanatory variables using the rules of thumb above. Since some of these depend on the shape of your particular interest in the variables, your choice might be different from a classmate’s.
* *For each of the nine pairs of variables, write down your choices for response and explanatory variables here, along with the appropriate LittleApp from the table above.*   .  .  .

1. Now you are going to try to describe the relationships using statistics. Just like a forest can be obscured by the trees, so a relationship can be obscured by the scatter of data. Often, the appropriate statistic makes clearer the relationship.

* Open up the appropriate Little App from step (3) and turn on whatever statistics you can use to describe the relationship. (You can choose the sample size to be whatever you want, but generally a larger sample size makes it easier to see a relationship. You may also choose to *stratify* the sample.)
* Give an appropriate English-language description of the relationship. Some of the terms you might use are “upward sloping function,” “downward sloping function”, “difference of means,” “difference of proportions,” “no relationship.”
* *For each of the nine pairs of variables, write down your English-language description of the relationship.*   .  .  .

Version 0.3, 2019-05-23, Danny Kaplan, [Word version](/word-versions/foo.docx)