Slide 1 – Title Slide

Hello and welcome to Week 6, Part 1 of EGM101: Variables. In this lesson, we'll learn a bit more about what variables are and what we can learn from them.

Slide 2 – Week 6 Outline

Over the remaining lessons this week, we'll cover correlation and how it absolutely does not mean causation; we'll also learn about topics like: linear regression; the coefficient of determination; outliers and how to deal with them; before wrapping up by discussing interpolation and extrapolation.

Slide 3 – Variables

Early on last week, we introduced the term "variable" with the following definition: a variable is some characteristic that will have different values for each member of the population – that is, it's a characteristic that *varies*.

Some of the examples that we discussed were things like length, weight, age, sex, color, or other attributes that we might be interested in studying in the context of studying fish.

For humans, we might be interested in variables such as annual income or opinion of government or voting intention.

Slide 4 – Relationships and Association

So, variables are characteristics that we study in a population. Sometimes, we might study them on their own, but more often we're going to be studying how different variables *relate* to each other. For example, in this graph I've plotted fish lengths, measured in centimeters on the horizontal or x-axis, and I've plotted fish weights in kilograms on the vertical or y-axis. From this, we can see that the points appear to make some kind of a shape – that is, there's some kind of relationship between the length and the weight measured.

Before we get too much deeper into relationships, we need to provide two more definitions. The first is the independent variable – this is the variable in the relationship whose value does not depend on another variable.

You will often see an independent variable referred to as the "explanatory" variable, but keep in mind that while these two terms are often used interchangeably, that does not mean that they are always interchangeable. As we will cover more next week when we talk about probability, if a variable is truly "independent" it means that the variable has no dependence on *any* other variable; in practice, we don't always know if this is the case, and "explanatory" is often a better term to use.

The other term to introduce is "dependent variable", also called a "response" variable. This is the variable in the relationship that does depend on the value of the explanatory variable – it "responds" to changes in the explanatory variable.

On a graph, we typically plot the value of the explanatory variable on the horizontal or axis, while the response variable is plotted on the vertical axis.

When we are looking at the relationships between two (or more) variables, we are looking at the variation of the response variable based on the value of one or more explanatory variables.

In the example here, we would say that the length variable "explains" the weight variable – in other words, as the fish's length increases, the fish's weight increases according to some as-yet undefined relationship. That is, there is some kind of association or relationship between these two variables.

It's very important to stress here that the existence of an association or relationship between two variables does not imply that the relationship is causal – that is, we cannot say that an increase in length *causes* an increase in weight; all we can say from looking at this graph is that there is a relationship.

Slide 5 – Causality

We introduced this word on the previous slide, so it's good to put a definition down. As the name implies, causality means that a change in one variable (the explanatory variable) *causes* a change in the other variable (the response variable).

I've said this already, but it's an incredibly important point that needs to be stressed: when we find a relationship or association between two variables, it does not mean that there is causality!

For example, there might actually be no relationship at all – the association may only be due to random chance over a narrow range of time or values. Or, there might be something called a confounding variable – a variable that has an effect on both of the variables that we are examining that causes the appearance of some kind of association or relationship.

One of the classic examples here is an example of shark attacks and ice cream sales, which I'm pretty sure I am legally obligated to use as an example in an introductory statistics lecture.

So, let's pretend that we're researchers looking at incidences of shark attacks on a particular beach, and we notice a strange association. Every year, we notice that as sales of ice cream at a particular ice cream shop increase, we see a corresponding increase in shark attacks on the beach. Over the winter, ice cream sales are very low, and so are shark attacks. That is, there is a clear association between these two variables.

Now, we could draw a diagram of this as follows. We have ice cream sales here, and shark attacks here. Now, we might be very tempted to say that the increase in ice cream sales is what causes the increase in shark attacks, indicated by an arrow in this diagram. But, as you are hopefully thinking, this is absolutely bonkers. For one thing, we could just as easily (maybe more likely?) say that people are buying more ice cream to help them cope with anxiety about shark attacks. This, however, would be equally bonkers, because these variables have no causal relationship.

Instead, there is a far better explanation: temperature. As temperature increases in the summer months, more people buy ice cream because it's hot. Also because it is hot, more people go to the beach, which means there are more people in proximity to sharks, and a higher incidence of shark attacks.

When you read news articles or even research articles that talk about a “link” between two variables, it’s important to think about what the possible cause of that link or relationship might be – without one, it’s very difficult to conclude that there is a causal relationship.

Slide 6 – What’s in a Relationship?

Now that we’ve discussed relationships and causality, let’s look at some of the ways that we can describe the relationship between two variables, starting with something called the direction.

The direction tells us the “sign” of the relationship – that is, if the direction is positive, it means that as the or explanatory variable increases, the response variable also increases. Or, a positive change in one variable is also a positive change in the other. This also means that a negative change in one is a negative change in the other – the change is the same for both variables.

A negative sign means that we have opposing changes: if we increase the value of the explanatory variable, the value of the response variable decreases; along the same line, if we decrease the value of the explanatory variable, the value of the response variable goes up.

And, as you might guess, a direction of zero means that we don’t see any change in the response variable as we change the value of the explanatory variable – there doesn’t appear to be any kind of relationship here.

The other property we’ll look at is the strength, or consistency, of the relationship. Looking at a fixed value of the explanatory variable, if we have a large spread of response variable values, we say this is a “weak” relationship. By the same token, then, if we have a small spread of response variable values for a fixed value of the explanatory variable, we say this is a “strong” relationship. We will discuss this more in the next lesson when we talk about correlation.

Slide 7 – Types of Relationships

We might also see different forms of relationships. The one that we will be working with most common in this module are linear relationships, where the form of the relationship makes a straight line on a graph.

But, this is far from the only possibility – the relationship could be nonlinear, which is sometimes called “curvilinear” because it has a curved shape. Some different examples of this are shown here, such as a polynomial, where the shape depends on the explanatory variable raised to some power;

other common forms are exponential relationships, where the response variable is equal to some constant value raised to the value of the explanatory variable; or logarithmic relationships, where the response variable is the logarithm of the explanatory variable.

Slide 8 – Summary

In this lesson, we've discussed how a big part of statistics, and science more generally, is understanding the relationships between variables. This includes response variables, which are variables that we're hoping to learn more about by using explanatory variables, which we hope provide some "explanation" for the changes that we see.

As part of investigating the relationship between variables, we can also think about causality – whether the changes in an explanatory variable actually cause changes in the response variable or not.

In addition to causality, we also consider the direction, strength, and form or shape of the relationship – whether changes in the variables are the same or opposing, what the spread of values in the response variable is for a given value of the explanatory variable, and whether a plot of the two variables looks like a straight line or some other shape.

Finally, you will eventually get sick of me saying this, but it's very important to remember that not all relationships are causal – this is a theme that we're going to keep coming back to this week (and beyond).

Slide 9 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapter 12.1; Caswell, Chapters 9.1 and 9.2; and Weiss, Chapter 4.2.

I've also provided links to a YouTube video that provides more information about explanatory variables, and a page on a great web resource for statistics that also provides additional information and explanation about variables.

That's all for this lesson – I hope you found it interesting, and you have any questions, please don't hesitate to e-mail me or post in the discussion forum on blackboard. Bye!