Science and Inference

1 WHY ARE QUANTITATIVE METHODS IMPORTANT?

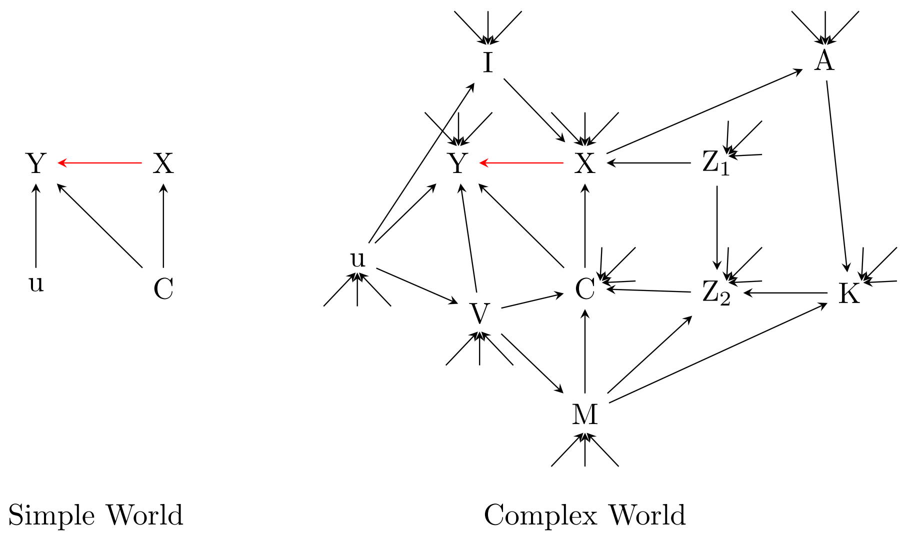
Before we start doing any statistics, it is important to understand a bit about the philosphy of quantitative thinking and how it links to scientific thinking. For many of us statistical reasoning is equivalent to critical thinking. Of course, not everyone agrees, and that is fine. But here is the argument for this claim laid out. In essence, this is all about decision making under uncertainty and depends on your view of the complexity of the world.

2 HOW COMPLEX IS OUR WORLD?

You often here Statisticians scoff and roll their eyes when the hear phrases like "with 95% probability the true effect of x on y is in this interval". This is wrong on many levels; exactly on how many levels depends on your view of science.

To understand this, step back for a second and think about how complex you **think** the world is. For example, how complex do you **think** are human interactions? Simplifying a bit, you can get two extremes: relatively clear delineated causes and effects, or a jumbled mess where everything causes everything. See for example Figure 1:

Figure 1: How complex is reality?



The point, is we do not know how complex the world really is! We have some guesses. Many philosophically minded Statisticians tend to think that the complex version probably describes reality better, but we don't know.

If we take science seriously, we should err on the cautious side and imagine the world being very complex. However, in the complex world, it is hard to learn about a relation, say between x and y, because there might be many possible forms of that relation. It might not even be completely sensible to talk about **the** relation between x and y, if the link between x and y is so highly contextual as in the complex world model in Figure 1. Summarizing: the complex world makes it very hard to learn how it works, because:

1. Well, by the nature of being complex, there are very few simple relations.
2. Very few big, first-order effects. Many many small, interacting effects.
3. Learning from data is considerably more difficult if the data generating process has so many moving parts (We will talk a lot about that in the course)

3 It is models all the way down

Learning how the world works is hard. Statistical inference is one way of trying to do so. Statistical inference basically means **drawing conlcusions from data**. Figure 2 describes the process:

Figure 2: Drawing conclusions from data

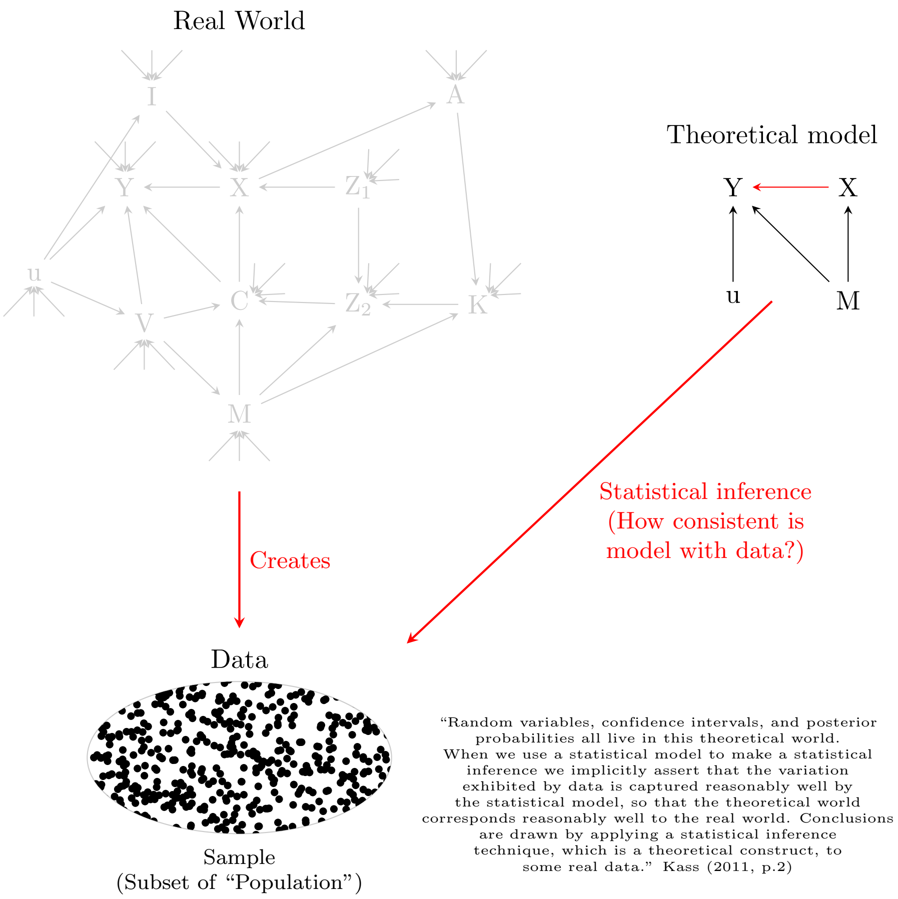


Figure 2 says that learning about the world via statistical inference is indirect. The steps are:

1. Formulate a theory of an aspect of the real world.
2. Turn it into a statistical model (Introduce random variables, etc.)
3. Check how well the model fits the data (Those pesky hypothesis tests).

Think about this for a while. There are a couple of key insights in here.

3.1 All models are wrong

George Box said: "All models are wrong but some are useful". I hope the meaning becomes clear from Figure 2. Depending on how complex the world is, a model that fully describes it would be, by definition, reality again and thus not useful. A model is a simplification – an idealization – of reality. It grabs a certain mechanism and tries to isolate and learn about that particular mechanism. Depending on how isolatable that mechanism is, this is more or less sensible. But we only know this after we have tried lots of hypotheses and model fit tests and found that our proposed models do not really fit well to the data. This is how quantitative science creates knowledge.

3.2 Talk about "true effects" is a mine field

Even if it feels disappointing, "learning the truth about how the world works" is not possible. But we can be unsure to different degrees. After we have fit lots of models to good data, we can better judge how the world probably doesn't work. A common mis-step is then to turn around and act as if we then know "true value regions". That is wrong, we only know which models don't work well. But there are always many other models. Maybe some of them predict a completely different effect and might better fit the data.

3.3 Talk about causal effects is tricky

The universe is deterministic as far as we can tell, so things must have a cause. But they may have lots of causes that interact with each other. Thus, it is important what causal effect one is interested in. And, depending on how complex you think the world is, your only chance of studying causal effects might be an experiment. Because experiments allow you to isolate certain arrows in the complex world in Figure 1 by randomization. That is the greatest strenght and the greatest weakness of experiments. For example, Nobel Prize winner Angus Deaton cautioned that it might be dangerous to draw policy recommendations from field experiments alone in development economics. Think of questions, like early child support, micro finance, etc. His point is basically that by isolating the red arrow in Figure 1, you cut off all the arrows that are there and modify the causal effect of x on y in reality. These additional relations represent important contextual effects that could significantly affect how benefecial or detrimental treatments would be (You can think of these as "context"; what some people mean when they argue that a treatment effect might be highly contextual) For example, there is a debate about what policy implications can be drawn from an incredibly involved field experiment: “Labor Market Returns to Early Childhood Stimulation: a 20-year Follow-up to an Experimental Intervention in Jamaica” by Paul Gertler, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeerch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor. The take-away of the study is: early childhood stimulation raised earnings by 25%. If the world is complex, the effect of early childhood stimulation should be highly contextual, so what is this 25%? If the experiment is sufficiently randomized, it would be an average effect over all contexts that were in effect during the experiment. Which where those? How would it extrapolate?

3.4 In order to do good science you need a good model and good data.

This is absolutely crucial, because it used to be taught differently in the past! You always need both: A good model and good data. On the one hand, interesting data alone does not guarantee a good scientific study. If you only have a vague theory (aka, x could have a positive or a negative effect, or maybe none?), then you are essentially just data mining. Data mining can be a very useful form of exploratory research, but it mustn't be labeleld theory testing. On the other hand, even if you have a good theory, bad data quickly destroys your chance of learning something about how well your model describes reality. First, if you cannot tell whether your model does not fit because your data is bad or your theory is wrong, you cannot learn much. Second, noisy data (tons of other effects going on) create so many "random" patterns that even if your model is completley misleading chances are high you find enough (spurious) patterns that concur with it. How noisy is noisy? That depends on the strength of your theory and how many researcher degrees of freedom it offers you.