An Introduction to Stata – Section 1.2
Najib Mozahem

I am sure that many of you have heard of the term data analytics. Today, this topic is trending. However, contrary to what you hear in the media, the topic is not new. Some of the concepts in it are new, but data analytics has been around for a very long time but under a different name, which is statistics.

Thanks to the spread of computer networks, companies can now collect a vast amount of data, whether it is about their customers, employees, or even competitors. The amount of data collected is really huge, which is why some of you have probably heard of the term Big Data. The problem with having too much information is that it becomes difficult to make meaning of the information. It's just like when five people are speaking at the same time. There is a lot of information, but you need to concentrate really hard to understand what is being said. In statistics, we use equations in order to find meaning in the data. There are a lot of tools in statistics, and knowing which tool is to use is extremely important. Think of statistics as a toolbox in your house. If all you have is a hammer in that toolbox, then when something needs to be fixed you are going to end up hitting it on the head whether or not this is the right thing to do. If however you have another tool, then you will be able to solve the problem with the appropriate tool. The more tools you have, the larger the number of problems that you can solve. The more tools you have, the larger the number of perspectives that you can have. You can look at the data from many different perspectives, depending on what statistical tools you know.

Collecting data helps us understand what is going on. There is a large difference between someone who believes that something is right because he "feels" that it is right, or because in his or her opinion it is common sense, and between someone who collected the data, searched for the story that is being told by the data, and then argued based on the evidence. I will give you an example about this. In one of my classes, we talk about the concept of failure. We always hear that failure is part of the growth process. It is how we learn from our mistakes. This sounds nice, but is it true? Just because something sounds nice doesn't make it true. Maybe we like to hear that failure is necessary because many of us are struggling in certain aspects of their life and want to reassure themselves that everything will work out fine eventually? In order to see whether the data supports this I collected information about football players (or soccer player as it is called in the U.S.). At the end of each season, we are told who were the top goal scorers. I was interested in the opposite question. Who were the biggest failures? To find the answer, I got the data, loaded in into Stata and performed some analysis. From This analysis, I was able to produce the below two figures. I showed the figures to my students and asked them what they represent. Since the figures don't state what is the variable that is being measured, the students started to guess. As you can see from the figures, the players and clubs mentioned are among the top and most popular in the world. This is why students would usually guess that the figures measured salary, revenue, and number of followers on social media networks. No one would guess that the figures measure the failed shots on goals. The figures clearly show that the top goal scorers in the world are also the ones who miss the most chances. The same goes for the clubs.

Students are usually shocked by this. They had always heard that failure is necessary, but it is one thing to hear something and a completely different thing to actually see it. I usually do not stop there. We can use statistics to see whether there is a relationship between two variables. I do

that and show the students that there is a strong and significant relationship between the number of shots missed and the number of goals scored. Some students would argue that there would also be a relationship between the number of goals scored and the number of minutes a player spends on the field. In order to see if this were true we looked at the data and it turned out to be true. However, statistics allows us to see which two variables have a stronger relationship, and when we did this analysis we found that when we include both the number of failed attempts and the number of minutes played, the stronger variable turned out to be the number of failed attempts.

By looking at the data, we were able to tell a story, and that story was the story of players who fail and who succeed. The media concentrates on the success part, because it sounds more interesting and would generate more sales. But this is not the whole story. Using statistics, and a powerful software like Stata, we can extract the whole story, or at least do a better job than other people. It might turn out that some datasets don't have an interesting story, but the process to finding whether there is a story or not is a very interesting one and is, in my experience, very enjoyable. To be able to do that, we have to know how to use a statistical software such as Stata, and this is the aim of this course.