Week 7: Data and Decision-making

Transcripts

Week 7: overview

Hi there and welcome back. This week we will discuss statistics and exploratory data analysis. Our motivation here is to enable you to move from data to decision-making, which is the core of this program.

We start by introducing some simple examples in the form of case studies that will help us motivate the study of statistics and its applications. The first example we will consider consists of a data set of house prices and their total square meterage. The second example will consist of an admissions data set from UC Berkeley in the US. We'll see that depending on how you look and interpret the data, you could conclude that there was gender discrimination or not.

Finally, we will consider a case study consisting of mortality rates for house patients versus hospital patients. The question is, if you're sick, you're more likely to survive at home or in the hospital.

Then I would like you to collectively create your own data set by submitting guesses for how much you think an elephant weighs. And then we're going to apply some of the tools we've learnt to make an intelligent guess about the weight of an elephant.

All those case studies are meant to help us understand how the interpretation of a data set and the models we use can affect our decisions. Let's get started.

Moving from data to decisions

Our goal now is to learn some statistics or perhaps for some of you to review some statistics. We're going to take advantage of the fact that we know some programming and we know some probability theory by now. And we're going to use all those tools we've learnt recently to try to understand how to interpret data, and how to extract important information or how to extract interesting and relevant information from a data set.

So, our goal here is then we're going to start talking about statistics and our goal is to move from data to decision. And to do so, we're going to use computer programs, we're going to use probability, we're going to use some statistical models, etc.

So, to motivate our study of statistics, let's start with an example. Let's say you are given some data. Let's say you have here some data relating to the size in square feet of a house and how it

—the costs vary with the size. So, for example, I have 1,400 that house will cost 112,000. Okay? So, those are sort of like more or less random numbers here. So I have 240 (the professor incorrectly reads '2400' as '240'), 192,000. So, this is in thousands of dollars here, and so on so forth, 1,800, 144,000, 1,900, this is a 152,000, 1,300, yeah? This is a 104,000, 1,100 here. Okay. Now, let's say you're trying to derive some consequences from this observed data. What can you say and how can you get to those conclusions?

So, to motivate it, I'm going to have to suggest one of the most sort of like elementary ways of looking at a pair of data columns, which is for example, since I have these two columns of data, how about we look at how cost varies as a function of size? Based on that data set above, how do you think that variation is going to look like? Would it look like this, an increasing function? Will it stay more or less the same? Will it increase, or will it decrease? I would like you to make a guess.

And if you thought about it, you can check yourself that, yes, in fact, price is approximately increasing as a function of size. For this concocted data set, that relation is identical linear, but in general, cost will be a function of size. And one of our goals in this module is trying to figure out how to determine or how to build models for how certain variables or how some features, data features are related to each other, and how to use that information for doing decision-making.

So for example, let's say that I had a house. You've observed that cost should be more or less a linear function of size and that — and for some reason, for a house of a fairly decent size, imagine that that house is below what you would expect would be sort of like the regular market size for that house, then obviously you can see here that that is a purchase of opportunity for a house, assuming that the house is in a similar neighbourhood to the ones you've used for plotting the relationship in the data. This house is cheaper than usual. And if you're a home seeker, that's probably a good opportunity for you.

So, yes, our goal then in this module is to look for, look at tools and ways of thinking that will enable us to extract information from a data set and how to use that data set for making decisions.

Statistics theory and practice: the stories data tell you

So, another big drive for this module is how do we interpret the data that we are given access to? How do we interpret the data that we obtain in our jobs or in our day-to-day lives? So, let me share with you a case study that is based on a real data set and a real situation that happened at UC Berkeley a few years ago. So, this is the so-called admissions case study.

So, here's some data.

I have male and female students. They've applied for two majors, major A and major B. And within each major, some students applied, some students were admitted. So, let's take a look at some numbers here.

For major A, in the male population, 900 male students applied to major A, 450 of them were admitted. And for major B, 100 students applied and only 10 of them were admitted. Whereas if I look at the female student population, I had that 100 female students applied for major A, 80 of them were admitted. And, for major B, 900 of them applied and 180 of them were admitted. So, this is, again applied and admitted here. So, the question that the admissions committee were asking at the time was, was there gender bias in this admissions process? Let's think about it for a second.

So, if I look at the male population for, say major A, the admission rate for men was 50 per cent, and the admission rates for women were 10 per cent. So, it looks like men were considerably admitted more frequently. Same if I look at, for example, and on the other hand, it looks like women here, and I am going to use a different color, women were admitted, 80 per cent of the women that applied were admitted. And, on this other case here, 20 per cent of the women were admitted.

So it seems like, from looking at the data in a longitudinal way, so to speak, or if we look at the data at the gender level, major by major, it does seem that there was a bias in terms of like more women being admitted than men. But was that so? Was that really the case? So now, let's look at this data from a different angle.

Now let's look at the aggregates of majors A and B. So, for example, in absolute terms, now, so this is men, and this is here women. Now, let's say, we look at the percentage of men that applied and the percentage of men that got admitted. So, in that case, I would have here something like 460 divided by 1,000. So that's 46 per cent of all men that applied got admitted. But the women, on the other hand, I've only had 260; 26 per cent of them were admitted, when I look at aggregates of majors A and B.

So, does that mean that the admissions process was biased towards men? Well, on the other case, when we looked at the majors' level, it seems like it was biased towards women. So, the big lesson here is that, so the way we look at data and the way we interpret the results, the numbers that the data is telling us, shapes our conclusions.

So, it's important to be aware of how we're drawing conclusions from data. And the Berkeley case study is a good example of that. Depending on how we're looking at the data, we arrive at two completely different conclusions related to gender discrimination.

Correlation is not causation

Correlation and causation. One of the biggest sins in statistics or in drawing conclusions from data is the confusion between correlation and causation. There is often a confusion that

correlation and causation are interchangeable words, but in fact they're not. One of the points I would like to make in this presentation is that the relation does not imply causation. So, let's see an example to highlight the importance of understanding that correlation is not equal to causation.

So, let's say, I give you some data related to mortality rates in a hospital or at home. So, let's say that in a hospital, I have that out of four (the professor incorrectly reads 40 as four) patients, four people died. Okay? And let's say that I have similar data that, at home, out of 80,000 (the professor incorrectly reads 8,000 as 80,000) people, only 20 people died. We don't have to think of this in absolute numbers; we could also think in terms of percentages, for example. This is 10 per cent and here, this is one in 400. So, this would be a lot less, right? So, for example, this would be here 20 divided by 8,000. So, this is 0.0025, or 0.25%. So, imagine that you're sick. It looks like your chances of dying in a hospital are a lot bigger than dying at home.

But is that the real story? Think about it for a second. Well, now, let's say that if I am not fully convinced about these statistics and that I will, in fact, look for other factors and look, for example, whether those people were sick or healthy. And when I do that, when I breakdown my data that way, what I end up with is the following. I have 40 people who were sick and four of them died. They are all in a hospital. And out of four people were healthy in the hospital, none of them died.

On the other hand, if I do a comparison, a similar comparison, for people that decided to stay home when they were sick, I get a data breakdown that looks like this. So, this is home. And of 8,000 people that were sick at home, 20 of them died. But out of 7,960 people who were healthy at home, 20 of them also died. And it turns out that, for example, if I take a sample of people that were in the hospital and that were sick and healthy, then, yes, out of those people, for the sick people, four in 40 died in the hospital. But, for example, which is about 10per cent, but for a sick person at home, it looks like that their chance of dying if they don't go to the hospital, now it's about 25 per cent, as opposed to a small fraction. Also, a healthy person from this data is unlikely to die in the hospital.

So, the moral of the story here is the following. When we got that first version of our data set, we were tempted to make a jumpy conclusion that people are more likely to die in a hospital than they are likely to die at home. In fact, because we were missing information about the data set, it was once we looked at actually the health status of those patients that we figured out that, in fact, the data set implied a different story. So, here is 36 out of 40. And here of 8,000. So, 7960. And then, here, it's 40. So, from our first look at the data, it looked like it was more likely for you to die at home than it was for you to die in a hospital. But once we include an extra feature, which is whether those patients or people were sick or healthy, you see a completely different story here.

So, for example, if I breakdown my population now into sick and healthy patients, you see here that you're more likely to, if you're sick, to die at home than to die at a hospital. And you can understand the reason why. A hospital is much more likely to provide you with care and assist you in those situations. So, the moral of the story here is whether — we have to be very careful about how we interpret our data.

Firefighters' example

Do not confuse causation with correlation.

So, let's consider a fictitious example here. Let's say you have some data about firefighters and fire sizes. So, you know, you have some building, different sizes, and this building could be on fire. And the data that the fire department compiled over the years is the following, which is firefighting crew size and fire size. So, imagine here that this is in square metres and this is in numbers of people. And the data looks something like this. So, for a fire of, let's say, 100 square metres, they had a crew of size five for a fire of size 500. You could even think of multiple floors. They had a crew of size, say 30, and here let's say here a fire of size 1,000. So, let's say, you had a crew here of size 75 so on so forth.

So, the question is, do you think that a larger firefighting brigade is implying a larger fire size? And I know it does sound like a word problem, a wording puzzle, but yeah. The idea here is to highlight the difference between causation and correlation.

So, yes, you would expect that as the fire size increases so does your need for a fire crew, for the size of a fire crew. But that does not mean that a larger fire crew causes the size of the fire to increase. This is a simple and somewhat contrived example, but it does help you to sort of highlight and really sort of punctuate the fact that correlation is different from causation.

And even though nowadays there are mathematical tools for measuring causations between variables, most of the time what people talk about are correlations, and most of the time there's a confusion between them two.

So, as an exercise, I would like you to post, in our hub, examples of news articles or blog snippets, whatever you could find in the wide web, of examples of people explicitly confounding correlation with causation. And unfortunately, for everyone's loss, that causes a lot of confusion and a lot of problems, for us, seeps in sometimes.

All right. So, go out, do your research, and try to figure out if you can find some examples of correlation being confused with causation.

Week 7: summary

Hi again! I hope you had fun this week when we began learning about statistics and decision making. We have learnt this week that it's not only important to summarise, for example, by using the means and to visualise data, for example, using scatter plots, but also how our own interpretation of the data or the story it tells can completely skew the results.

We have also learned how in the media, politics and in management there's a constant confusion between correlation and causation. My hope is that after completing this module, you're more aware of this confusion and you will be able to spot this cognitive bias when confronted with it. Let's move on.