Chapter 1: The World of Data Around Us

Steph de Silva

06/11/2017

Introduction

Data is all around us. In a world where data is being generated on an unprecedented scale, the ability to find insights that data is a significant competitive advantage in business.

This unit will teach you how to explore, discover and search for those insights.

Statistics

One of the first things you will need to understand is the concept of *statistics*. What exactly do we mean by statistics?

Statistics is the scientific study of how data behaves. It provides a framework for decision making and for discovery. There are two main kinds:

- Descriptive statistics: using numbers to describe the way things are. That's about it, really!
- Inferential statistics: using numbers to extrapolate about the way things *may be*. We don't have all the information, but we have some and we can use that to make *inferences* about the way the world works.

Populations vs samples

One of the fundamental concepts of statistics is that of a *population* and of a *sample*.

A population contains all the items of interest. That's a pretty vague term, let's try some examples:

- If we're interested in Koala health, it may be all the koalas in a certain region.
- If we care about diabetes treatments, it may be all the people with diabetes.
- If we care about phytopthera disease in avocados, it may be all the avocado trees in a certain place.
- If we care about a company's return on equity, it may be each return on equity figure published on the balance sheet of the annual report for that company every year since IPO!

When I worked for the World Bank, we studied the early grade reading of school children in many different countries. In our case, the population in our studies was all the school children in the country we were working in. You can imagine, that's alot of kids in countries like Papua New Guinea or Laos!

Can you imagine testing the reading of every school child in Papua New Guinea, Tonga or Laos? We couldn't do it. There were lots of reasons for that:

- Many of the places where these kids lived were inaccessible, or would become inaccessible at some times of the year. One of our studies was nearly ruined because it was simply too dangerous to send people to islands during summer storms! It just wasn't always possible to get there.

- It was simply too expensive to send people out to every single school to test every single child. We were there to improve reading, so saving most of the cash for that was the right thing to do, obviously.
- It would take too long to test every child. By the time we tested every kid, they all would have been at different times of the year. It wouldn't have made for comparable data comparing a child in Grade 1 in January with another in Grade 1 in December. Kids learn alot and fast!

While these were the specific problems we faced at the World Bank, they are in fact very commmon ones. Cost, accessibility and time make population studies very difficult. But we have a solution for that! It's called *sampling*.

All about sampling

Sampling means taking a small subset of a population and collecting data on that, instead of the whole population. It's cheaper, faster and can be extremely accurate when done correctly.

There are many kinds of sampling. Some of the simplest are:

- Simple random sampling. Every member of the population is equally likely to be chosen. Easy as! Unless some of the population is hard to get to, may choose not to participate or there may not be enough of some parts of a population to capture enough information about them… then we need a new plan.
- Stratified random sampling. We used this one alot at the World Bank. For instance, we knew that small schools tended to have less money, more diverse language use and students from more remote backgrounds compared to large schools. But we didn't have the time to visit as many small schools, so we stratified the sample between small and large. Later on, we could use a statistical technique called *weighting* which allowed us to return the sample to a representative one, even though we had fewer small schools.
- Cluster sampling. This is what we use when the individuals of the population form clusters in some way. For instance, for school children, they are clustered in schools. So we randomly selected schools and then randomly selected students within those schools.
- Convenience sampling. This is where you grab anyone that walks past, when you make your online survey open to anyone or when you ask six of your closest relatives and all your Facebook friends at 9:15pm on a Friday night. It's not a statistical form of sampling, but it happens alot anyway. It's much harder to make sensible inferences about this kind of data and what this kind of data describes is also difficult to define.
- Another non statistical technique for inference or description is judgement. It's not a sampling technique. But often people make inferences based on their judgement or experience. That's not a bad thing necessarily, but it's not a statistical method.

Samples allow us to make inferences

The key point of sampling is that it allows us to make inferences about a larger population. When we took samples of school children, we used these to understand what was happening in a country or a region more generally. The schools and kids were important: but what they could tell us about all the other schools and kids was **really** important!

There are a few different categories of inferential statistics, but it comes down to two basics for our purposes:

- Estimation. We use sample statistics to estimate characteristics about our population. You may have heard of a mean or an average: we use the sample mean to make inferences about the population mean. More on that in coming weeks!
- Hypothesis testing. We can use the sample statistics we create to test ideas and theories about our population. One thing we did often at the bank was test hypotheses like:
    - “Are boys reading better than girls in this country?” (The answer was always no in my experience!)
    - “Are there differences between a child’s ability to read in school, depending on the language they are speaking at home?” (The answer was usually yes.)

As you can see, we can answer some quite important questions about a population with a good sample.

What about in a business context, what might you want to know? Well, everything really!

- Did the change to website generate more revenue than the old version of the website? (This is called A/B testing in web development- it’s still hypothesis testing!)
- What’s the relationship between advertising spend and revenue? Does it matter? Does it matter differently for different customers?
- What’s the forecasted demand for a company’s product next year? How to plan for the future?

All of these things can be explored with inferential or descriptive statistics.

Two important concepts: variables and parameters

Two concepts will come up often over the course of this book. They are the concepts of a variable and a parameter.

- A variable is simply a collection of data. In Excel or another spreadsheet, it’s usually a column. A variable may be revenue of a firm over time or it may be something like the number of clicks a post gets on Instagram.
- A parameter is a way we can describe a data set. There are lots of parameters you can use to do this and not all of them are relevant at any time. This can include things like the mean which you may have heard of, but it can also include quite esoteric and weird things we’ll leave to a more advanced course. All you need to realise at this stage is that it’s a way to describe data.

A note on data types

It’s often useful to divide data up into types, depending on the way it’s ‘packaged’. Bear in mind that data can be altered so its type changes, but it can be a good way to start categorising how we can treat it! Some common data types include:

- Integer or discrete data. This just means the data comes as whole numbers. For example, 3 is an integer and so is 5. But 4.1 is not an integer. The concept of *discrete* just means the data comes in neat packages with space between it- those are the integers!
- The simplest form of discrete data is *binary*. It just means there are only two options: yes/no,

true/false and so on.

- Discrete data can be used to represent categories, for example colours. We may say red = 1, blue = 2 and green = 0. It doesn't make alot of sense to decide that purple = 4.78 though! That's why we use discrete data for categories. Other common examples of categorical data is gender (don't get me started on this one, it shouldn't be a binary category!), ethnicity, customer segment and so on. We also call categorical data *nominal* because it has names.
- Discrete data can also be *ordinal*, that is ordered. This is for things like first, second, third; or high, middle, low; or hot, medium, cold. Any discrete set of categories that have some kind of intrinsic ordering is ordinal.
- Continuous data. This is just data that can have any number of values. For example height is continuous, a person can be 176.1 cm tall or 176.2 or 184.2521, depending on how finely you want to measure it. Length, time and often money are all expressed as continuous data.
- There are two other kinds of continuous data, but they are rarely discussed:
    - Interval data. Data where we express the distance between two points. For example, a temperature of 30C is further away from zero than one of 20C. The distance between the temperature and zero is the critical part.
    - Ratio data. This is where we express outcomes as the ratio of two other variables.

Review Questions

1. What's the difference between descriptive and inferential statistics?
2. If we are interested in understanding fuel consumption in disel cars in Australia, what is the population?
3. If we are interested in understanding about whether a new tax change will impact women in the workplace, what is the population?
4. If we wanted to know about disadvantage in remote communities, what kinds of sampling could we use?
5. If I make a twitter poll asking people what they think of statistics in business, what kind of sampling am I using?
6. What is the point of samples? What kinds of things do they allow us to do?
7. I have a variable describing the proportion of women on boards in Australia. What variable type do I have?
8. I have a variable describing the bonus pool of a team over time. What kind of variable do I have?
9. I have a variable describing whether or not an employee has taken sick leave this year, what kind of variable do I have?

Review Answers

1. Descriptive statistics describes the data itself, inferential statistics uses the data to describe what may be happening- this usually happens on a broader scale, going from sample to population.
2. The population is diesel cars in Australia. In this case, not "cars" or "global cars" - there were two specifics, diesel and downunder!
3. Women in the workplace. But be wary with things like this: does the study actually care if women will choose to enter the workplace because of the changes? If so, the population is women of working age.
4. We could use stratified sampling - capturing communities small and large or by other key variables.

Or we could use clustered sampling as we expect communities to be clustered together. Choosing a sampling method is a fine art and often there is no clear cut answer. Also, we can combine different methods together!

5. Convenience sampling and possibly self-congratulatory, ego-boosting social media use.
6. Samples allow us to get data cheaper, faster and sometimes better than the population data. They allow us to make inferences about the population.
7. You have a ratio variable here! Women : total board members
8. Continuous data most likely, bonuses are measured in $ and theoretically these could be fractional!
9. We have a binary variable: either they have or they haven’t taken sick leave.