# 1.3: Big Ideas in Statistics

Now that you've understood some of the problems we usually come across when conducting statistics and interpreting the results, let's look at the main ideas that cut through nearly all aspects of statistical thinking. These ideas will be threaded throughout the whole book.

Before going into the "meat" of the book, introducing these main ideas can help you on your path to understanding statistics. My goal for you is to not just *remember* concepts but to *understand* them. By giving you the fundamental ideas first, hopefully, it will help you organise the knowledge you will learn in this book and make them more flexible and powerful.

Several of these ideas came from Ji Y. Son, a learning scientist who co-wrote a fantastic interactive textbook, *Introduction to Statistics: A Modelling Approach* and James Stigler's (2016) outstanding book *The Seven Pillars of Statistical Wisdom*, which are augmented here.

## Big Idea 1: Aggregation

One way to think of statistics is "the science of throwing away data". In the example of the PURE study previously, we took more than 100,000 numbers and condensed them into 10. This kind of *aggregation* is one of the most important concepts in statistics. However, you maybe asking yourself: If we throw out all of the details about every one of the participants, then how can we be sure that we aren't missing something important?

Statistics offers us ways to describe the structure of aggregated data, backed by theoretical principles that explain its effectiveness. However, it's crucial to remember that aggregation can go too far and later we will encounter cases where a summary can provide a misleading picture of the data being summarised.

## Big Idea 2: Variation

Statistics can also be thought of as "the study of variation". If variation didn't exist, then we wouldn't need statistics.

But variation is everywhere. If everyone experiencing chronic pain took a particular drug and their symptoms improve and those who didn't take the drug got worse, then we wouldn't need statistics. However, this scenario rarely occurs. Typically, some individuals who take the drug improves, while others do not. Conversely, some individuals who don't take the drug can recover. It can be challenging to determine if the drug truly cures the ailment or if the recovery is simply a coincidence. Statistics helps us understand such scenarios by providing a set of concepts and tools that have developed over centuries to detect patterns and make sense of variation.

## Big Idea 3: Modelling

When my niece, Bella, turned 1 year old, my brother and I bought her a 6V battery-powered Frozen toy car that was advertised for 18 months and up. She had to wait another 8 months to actually ride her toy car. This toy car is a model of a real car. It has nearly all the components to make it a car – it has four wheels, a steering wheel, a pedal, side mirror. You can basically drive it like you would a real car (but you need to be 5 or under, otherwise you would not fit. Trust me, I tried). One could argue that my niece's toy car is a good 'model' of a real car because this toy car was constructed using existing information about a real car.

As quantitative social scientists, we can also build (statistical) models of *real-world processes in an attempt to predict these processes in certain conditions.*[1] Unlike ride-on toy makers, however, social scientists are mostly working with *non-physical constructs* and therefore, we can only make *inferences* about the psychological process that our models are based upon. Like ride-on toy makers, we do still want our models to be representative of reality (or at least close to it). The closer the model is to reality, the better the *fit* of the model. We want to build a model that has *a better fit* (i.e., closer to reality), because if we use this model to make predictions about the real world then, we can be more confident that these predictions will be accurate.

Modelling is an important concept that we will touch on over and over again in this book.

## Big Idea 4: Uncertainty

The world is an uncertain place. We now know that cigarette smoking causes lung cancer, but this causation is probabilistic: A 68-year-old man who smoked two packs a day for the past 50 years and continues to smoke has a 15 per cent (1 out of 7) risk of

getting lung cancer, which is much higher than the chance of lung cancer in a nonsmoker. However, it also means that there will be many people who smoke their entire lives and never get lung cancer. Statistics provides us with the tools to characterise uncertainty, to make decisions under uncertainty, and to make predictions whose uncertainty we can quantify.

One often sees journalists write that scientific researchers have "proven" some hypothesis. But statistical analysis can never "prove" a hypothesis, in the sense of demonstrating that it must be true (as one would in a logical or mathematical proof). Statistics can provide us with evidence, but it's always tentative and subject to the uncertainty that is always present in the real world.

## Big Idea 5: Sampling from a Population

The concept of aggregation implies that we can make useful insights by collapsing across data – but how much data do we need? The idea of sampling says that we can summarise an entire population based on just a small number of samples from the population, as long as those samples are obtained in the right way. As we already discussed above, the way that the study sample is obtained is critical, as it determines how broadly we can generalise the results. Another fundamental insight about sampling is that while larger samples are always better (in terms of their ability to accurately represent the entire population), there are diminishing returns as the sample gets larger. In fact, the rate at which the benefit of larger samples decreases follows a simple mathematical rule, growing as the square root of the sample size, such that to double the quality of our data we need to quadruple the size of our sample.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

---

1. Field, A. (2017). *Discovering statistics using IBM SPSS statistics*. Sage. ↵

---