1.2: If I Apply Statistical Thinking all the Time then why do I find it Difficult?

> *If you tell students that math is logical and students don't understand something that's supposed to be logical, then students will start thinking that there must be something wrong with them – Bruce Hoskins*[1]

I hear you ask: "If I apply statistical thinking all the time – then why do I find it difficult?" Bruce Hoskins, a great educator who I look up to, argued that high school mathematics education is to blame for anxiety about statistics and numbers.[2] We have been told as young children that math is supposed to be logical. So, if students do not understand something that's supposed to be logical, then there must be something wrong with them.

Therefore, in this book, we will treat statistics as learning a new language. Think of this learning process as developing a new way of speaking and a new way of thinking. Even if you have never learned another language before, it is like visiting a new country, with different cultures and different norms. Similar to learning a new language, there are many ways to say the same things (e.g., an independent variable is also called a predictor variable).

More importantly, to be able to understand complex ideas, you need to have a solid grounding in the fundamentals. Therefore, statistics is not something you can cram at the last minute.

> *Statistics is not intuitive – Harvey Motulsky*

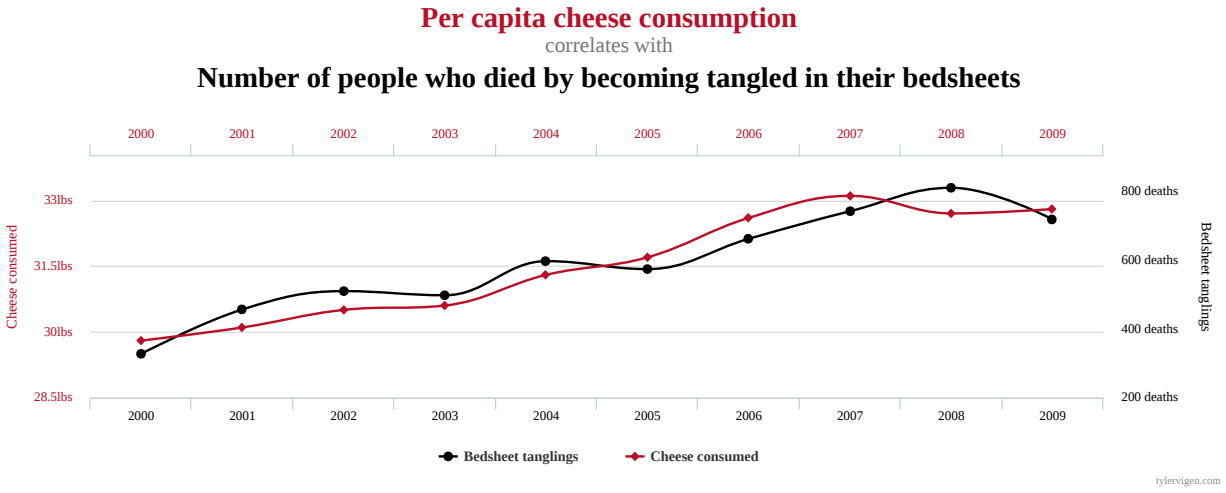
Harvey Motulsky, another great statistics educator, tells us how statistics is not intuitive. There are many biases within the human mind that often lead us astray when dealing with statistical concepts like probabilities. For example, human beings tend to jump to conclusions. My niece (who was five at that time) learned that I do boxing as a hobby and told me that, "boxing is for boys only"! She must have only seen boys to be doing hard-contact sports (like boxing) and made conclusions based on her small sample.

There is evidence to suggest that we may be hardwired to generalise from a sample to a population – even 8-month-old babies do it! (Xu & Garcia, 2008).[3] Hopefully, by the end of this book, you will understand that there are many problems with these types of generalisations. As Motulsky stated, *to avoid our natural inclination to make overly strong conclusions from limited data, scientists need to use statistics*.

> *Correlation does not imply causation – age-old statistical wisdom*

If I ever do a class in person, I would get everyone in the class to shout the above quote "Correlation does not imply causation!"

Just because two variables are correlated and the difference is statistically significant (we will learn more about this later), it does not mean that changes in the X variable caused the changes in the Y variable. Let's look at the following chart:



The above chart shows that cheese consumption has a positive relationship with the number of people who died from being tangled by their bedsheets. In other words, as cheese consumption increases, more people die from bedsheet entanglement.

Ummm… Is cheese to blame for people dying on their beds?

The figure above is an example of spurious correlation. A spurious correlation, also known as a false correlation or a coincidental correlation, is a relationship between two variables that appear to be related, but in reality, the relationship is not causal. In other words, the correlation between the two variables is a random occurrence that is not caused by any underlying mechanism or factor.

While the above example is a bit of a ridiculous one, we sometimes make the mistake that, just because two variables are highly correlated, we assume that one of the variables can cause changes to another. Let's go back to the PURE study.

Causality and Statistics

The PURE study seemed to provide pretty strong evidence for a positive relationship between eating saturated fat and living longer, but this doesn't tell us what we really want to know: If we eat more saturated fat, will that cause us to live longer? This is because we don't know whether there is a direct causal relationship between eating saturated fat and living longer. The data is consistent with such a relationship, but it is equally consistent with some other factor causing both higher saturated fat and longer life. For example, it is likely that people who are richer eat more saturated fat and richer people tend to live longer, but their longer life is not necessarily due to fat intake — it could instead be due to better health care, reduced psychological stress, better food quality, or many other factors. The PURE study investigators tried to account for these factors, but we can't be certain that their efforts completely removed the effects of other variables. The fact that other factors may explain the relationship between saturated fat intake and death is an example of why introductory statistics classes often teach that "correlation does not imply causation", though the renowned data visualisation expert Edward Tufte has added, "but it sure is a hint."

Although observational research (like the PURE study) cannot conclusively demonstrate causal relations, we generally think that causation can be demonstrated using studies that experimentally control and manipulate a specific factor. In medicine, such a study is referred to as a *randomised controlled trial* (RCT). Let's say that we wanted to do an RCT to examine whether increasing saturated fat intake increases life span. To do this, we would sample a group of people, and then assign them to either a treatment group (which would be told to increase their saturated fat intake) or a control group (who would be told to keep eating the same as before). It is essential that we assign the individuals to these groups randomly. Otherwise, people who choose the treatment might be different in some way than people who choose the control group – for example, they might be more likely to engage in other healthy behaviours as well. We would then follow the participants over time and see how many people in each group died. Because we randomised the participants to treatment or control groups, we can be reasonably confident that there are no other differences between the groups that would *confound* the treatment effect; however, we still can't be certain because sometimes randomisation yields treatment versus control groups that *do* vary in some important way. Researchers often try to address these confounds using statistical analyses, but removing the influence of a confound from the data can be very difficult.

A number of RCTs have examined the question of whether changing saturated fat intake results in better health and longer life. These trials have focused on *reducing* saturated fat because of the strong dogma among nutrition researchers that saturated fat is deadly; most of these researchers would have probably argued that it was not ethical to cause people to eat *more* saturated fat! However, the RCTs have shown a very consistent pattern: Overall there is no appreciable effect on death rates of reducing saturated fat intake.

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. Hoskins, B. (Host). (2020, October 29). *(Re)Teach: Teaching statistics pt 1* [Audio podcast]. https://reteach.buzzsprout.com/428977/6117973-teaching-statistics-pt1 ↵
2. I would like to clarify that this comment is more about the systemic issues we face rather than blaming individual math teachers. Teachers, I know you're already underfunded, stressed and constantly burned out, I don't want to place this burden on you! ↵
3. Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012-5015. ↵

This page titled 1.2: If I Apply Statistical Thinking all the Time then why do I find it Difficult? is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Klaire Somoray (Council of Australian University Librarians Initiative) .

- **1.5: Causality and Statistics** by Russell A. Poldrack is licensed CC BY-NC 4.0. Original source: https://statsthinking21.github.io/statsthinking21-core-site.