The Themes of Statistics

In a statistics course, it can feel like you're covering many diverse topics that don't have anything to do with each other; that you're starting over again with each new topic.

There are a number of recurring themes in your statistics course, and if you look out for them as you study the material, they'll help you draw parallels between different topics so you can build connections between them.

WE DO STATISTICS TO DETERMINE NUMBERS THAT WE CANNOT MEASURE DIRECTLY

If you wanted to know the average number of children in a Canadian household, how would you determine this answer? You could do what the government of Canada does and hold a **census**: a survey that asks literally everyone in the population the question. Even the governmental census happens only every five years, and it's terribly expensive. Without those resources and the law forcing everyone to answer the question, how could you determine such a number?

Without the census, you could never measure the number, but with the tools of statistics you can estimate such a number with a high degree of accuracy and confidence in your answer.

THERE ARE SOME NUMBERS THAT WE NEVER GET TO KNOW

All that being said, the answer you get from statistics is still just an estimate. How can we know how good the estimate is?

Even with the power of a national government on its side, the census still doesn't quite answer the question. Every day, babies are born, young adults move out of the house, and people make mistakes. There's no single point in time when we know the answer to that simple question, let alone a question that people might lie about. (Could we get a true national average income through a government census when welfare cheques and taxes are going to be affected by people's responses?)

The harsh reality is that in nearly all cases, the statistical estimates are the best answers we're ever going to have.

STUDIES AND EXPERIMENTS ARE PRONE TO "NOISE"

You may be familiar with the idea of a "signal-to-noise ratio", the idea being that we're trying to tune into a broadcast signal, but there may be static or other noise making the signal harder to hear.

Inaccuracies can creep into data, even without the human element. Changes in temperature, atmospheric pressure, and humidity can affect physics experiments. Biological systems are notorious for being affected by all sorts of external forces. When trying to act in their own best interest consumers and investors make decisions emotionally rather than rationally. The data simply will not sit still while we try to examine it, so even data that is free of **bias** (a systemic difference between the data we're

collecting and the data we're trying to collect) will still have experimental variation adding noise to the signal, obscuring the answer we're trying to see.

WE NEED TO BE ABLE TO DISTINGUISH BETWEEN THE NOISE AND THE SIGNAL

If all data has this noise issue, then how can we tell the difference between experimental variation and a trend or change in the data that we weren't expecting? There are a few ways we can do this. Being very careful with the way we design experiments and collect data can limit mistakes. Using multiple observations and combining the results can reduce the chances that variation in one direction will skew our results in that direction—abnormally high values should be balanced by abnormally low values in a large enough sample. Finally, we can test our results to see whether they stand up to scrutiny.

WE ESTABLISH A COMMON-SENSE "LINE IN THE SAND"

We scrutinize numbers all the time. When we're buying something, we have an idea of what we might expect to pay for it. We know that $300 is too much for a week's groceries for one person, and too little for any car worth buying, but reasonable for a laptop computer. For the groceries and the car, the price tag of $300 triggers the sense that the number is too far away from the norm to be believable without a good reason. Even if we're not aware of it, there's a "line in the sand" for numbers we're familiar with, and if a number crosses that line, we know something is strange.

Statistical experiments are still experiments, so with any results we publish, we'll want to include information about how our numbers were arrived at, so they can be replicated—and scrutinized—by others. Since we're always uncertain about any statistical conclusion we may draw, we can decide how uncertain we wish to be. We may say that a result which is consistent with our expectations will have a greater than 5% chance of occurring if we assume that those expectations are correct. This gives us an unambiguous way of deciding whether our results bolster our hypothesis or undermine it.

ESTIMATES THAT ARE OVER THE LINE ARE "SIGNIFICANT"

The idea of the line in the sand also gives us a way to challenge our expectations. If we discover that we really are paying $300 for groceries a week, then that behaviour becomes something we'll want to examine more closely to find out why, and whether something needs to be done about it. In the language of statistics, we've found a **significant** result. If we start by assuming our expectations are right, and we get reliable but anomalous results, inconsistent with those expectations, then our assumption might be wrong.

DATA TELLS STORIES; DATA CANNOT LIE, BUT IT CAN BE MANIPULATED OR MISINTERPRETED

Every once in a while we'll see a company advertise that they're the fastest rising company in their field. If we assume they're not lying, should we be impressed? If they had 2 sales last month and 22 sales this month, they can truthfully say that they've

experienced 1000% growth, and this might be more than their competitors (who already make hundreds of sales a week).

Legitimate statistics will always include context so the reader can judge whether their conclusions are valid. They may also include raw data so that there's transparency surrounding any calculations made. It's just as important to learn how statistics can be done badly in order to recognize it when we see it. Skepticism is healthy.

If you're coming to your stats course from a nursing or health science background, then the idea of applying critical thinking to what a patient tells you is a vital skill. Statistics is the science of applying that same critical thinking to numbers so that we're not fooled by them.

TOOLS ONLY WORK IN SITUATIONS WHERE THEY APPLY

That same fastest-growing company might project that they'll be making 200,000,000,000 sales a month by the end of the year, if their 1000% growth stays steady. The mistake here is applying a linear model to their growth rate when there's no good reason to believe that's the right model. This is a glaring example of misuse of statistical tools, but more subtle ones abound. It's no use carefully collecting data if you analyse it with the wrong technique.

Learning when various processes and tools are appropriate to use and when they're not is the bread and butter of this course. The calculations you'll be performing will not be hard. It'll all be arithmetic, and a slight amount of algebra. A lot of the time, you'll just look up the answer in a table, rather than calculating something. This is the first math course that you'll encounter that people actually do for a job. The synthesis of ideas and the critical thinking about what circumstances generated a data set and the decision of what should be done with that data is more important than crunching the numbers. A computer can—and should!—crunch those numbers for you. You'll need to apply your common sense to the word problems in the course if you're going to succeed.

If you watch the new material you learn in your stats course to see these eight themes, you'll start to see that you're not really doing a variety of calculations and analyses, but rather you're doing the same ones over again, but with modifications for different circumstances, and you may come to understand why the modifications were necessary.

It may help you, in your review for the final, to look at these themes again and try to see how they apply to the overall subject of statistics, and the specific topics you learned. As an exercise, could you outline an essay taking specific examples from your notes on how these themes arise in the course?