Introduction to Statistical Thinking

Felipe José Bravo Márquez

March 22, 2021

Introduction to Statistical Thinking

- Statistical thinking is a systematic way of thinking about how we describe the **world** and use **data** make decisions and predictions.
- Taking into account the inherent **uncertainty** that exists in the real world. [Poldrack, 2019]
- The foundations of statistical thinking come primarily from mathematics and statistics, but also from computer science, psychology, and other fields of study. [Poldrack, 2019]
- **Statistics**, in particular, is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data [Wikipedia, 2021].

Statistical Thinking and Intuition

- Human **intuition** often tries to answer the same questions that we can answer using statistical thinking, but often gets the answer **wrong**.
- For example, in recent years most Americans have reported that they think that violent crime was worse compared to the previous year (Pew Research Center).
- However, a statistical analysis of the actual crime data shows that in fact violent crime has steadily decreased since the 1990's.
- Intuition fails us because we rely upon **best guesses** (which psychologists refer to as heuristics) that can often get it wrong. [Poldrack, 2019]

What can statistics do for us?

There are three major things that we can do with statistics:

- **Describe**: The world is complex and we often need to describe it in a simplified way that we can understand.
- **Decide**: We often need to make decisions based on data, usually in the face of uncertainty.
- **Predict**: We often wish to make predictions about new situations based on our knowledge of previous situations.

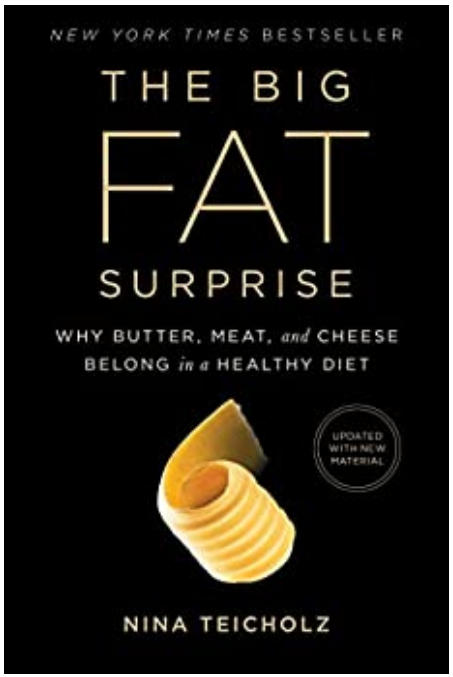
Example: Saturated Fat

- Suppose we want to answer the following question:
  Is saturated fat in our diet a bad thing?

- Common sense approach: If we eat fat, then it's going to turn straight into fat in our bodies, right?
- And we have all seen photos of arteries clogged with fat, so eating fat is going to clog our arteries, right?

Example: Saturated Fat

- Experts approach: The Dietary Guidelines from the US Food and Drug Administration have as one of their Key Recommendations that “A healthy eating pattern limits saturated fats”.
- You might hope that these guidelines would be based on good science, and in some cases they are.
- However, as Nina Teicholz outlined in her book “Big Fat Surprise” [Teicholz, 2014], this particular recommendation seems to be based more on the dogma of nutrition researchers than on actual evidence.

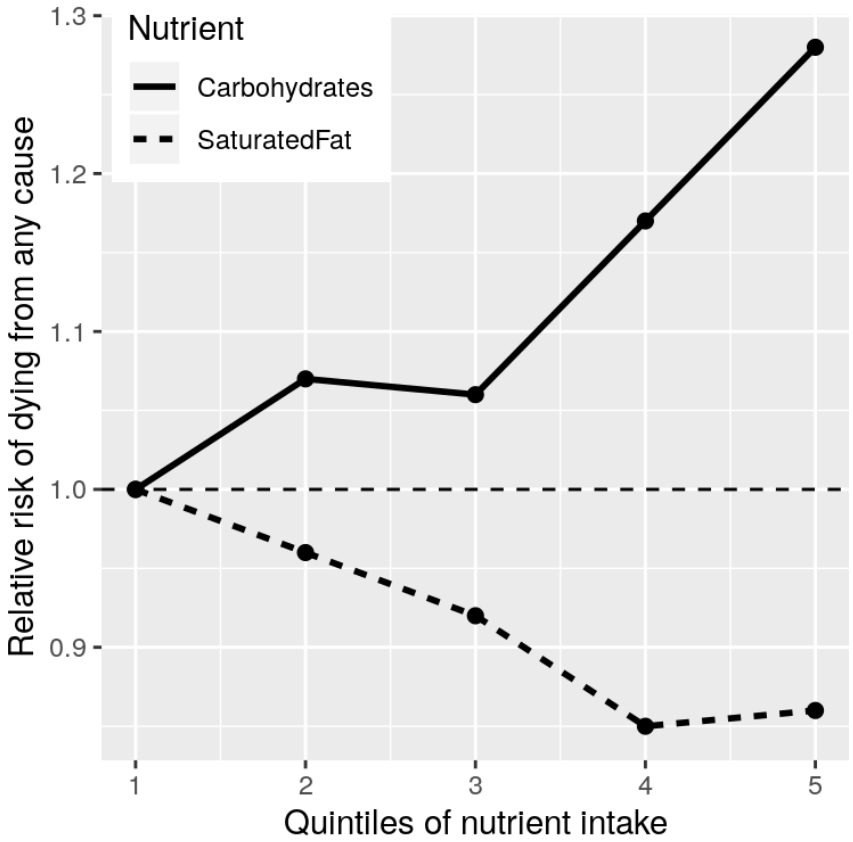


Example: Saturated Fat

- Statistical approach: we might look at actual scientific research.
- PURE: a large study that has examined diets and health outcomes (including death) in more than 135,000 people from 18 different countries.
- One of the analyses of this dataset reported how intake of various classes of macronutrients (including saturated fats and carbohydrates) was related to the likelihood of **dying** during the time that people were followed [Dehghan et al., 2017].
- People were followed for a median of 7.4 years, meaning that half of the people in the study were followed for less and half were followed for more than 7.4 years.

Example: Saturated Fat

- The following figure shows the relationship between the intake of both saturated fats and carbohydrates and the risk of dying from any cause.



- This plot is based on ten numbers.

Example: Saturated Fat

- To obtain these numbers, the researchers split the group of 135,335 study participants (which we call the “sample”) into 5 groups (“quintiles”).
- The first quintile contains the 20% of people with the lowest intake, and the 5th quintile contains the 20% with the highest intake.
- The researchers then computed how often people in each of those groups died during the time they were being followed.
- The figure expresses this in terms of the relative risk of dying in comparison to the lowest quintile.
- If this number is greater than one, it means that people in the group are more likely to die than are people in the lowest quintile.
- Whereas if it’s less than one, it means that people in the group are less likely to die.

Example: Saturated Fat

- The figure is pretty clear: People who ate more saturated fat were less likely to die during the study.
- The opposite is seen for carbohydrates: the more carbs a person ate, the more likely they were to die during the study.
- This example shows how we can use statistics to describe a complex dataset in terms of a much simpler set of numbers.
- If we had to look at the data from each of the study participants at the same time, we would be overloaded with data.
- It would also be hard to see the pattern that emerges when they are described more simply.

Example: Saturated Fat

- Despite the figures in the figure, we also know that there is a lot of **uncertainty** in the data.
- There are some people who died early even though they ate a low-carb diet, and, other people who ate a ton of carbs but lived to a ripe old age.
- Given this variability, we want to decide whether the relationships that we see in the data are large enough that we wouldn't expect them to occur **randomly**.
- Statistics provide us with the tools to make these kinds of decisions.
- Often people from the outside view this as the **main purpose** of statistics.
- But as we will see throughout the course, this need for **black-and-white** decisions based on fuzzy evidence has often led researchers down the wrong path.

Example: Saturated Fat

- If our conclusions were limited to the specific people in the study at a particular time, then the study would not be very useful.
- Based on the data we would also like to make **predictions** about future outcomes.
- For example, a life insurance company might want to use data about a particular person's intake of fat and carbohydrate to predict how long they are likely to live.
- An important aspect of prediction is that it requires us to **generalize** from the data we already have to some other situation, often in the future.
- In general, researchers must assume that their particular sample is **representative** of a larger **population**.
- This requires that they obtain the sample in a way that provides an **unbiased** picture of the population.
- For example, if the PURE study had only recruited vegetarian participants, we would probably not want to generalize the results to non-vegetarians.

The big ideas of statistics

There are a number of very basic ideas that cut through nearly all aspects of statistical thinking.

1. Learning from data
2. Aggregation
3. Uncertainty
4. Sampling from a population
5. Causality and statistics

Next, we examine each of them in more detail.

Learning from data

- One way to think of statistics is as a set of tools that enable us to learn from data.[1]
- In any situation, we start with a set of ideas or hypotheses about what might be the case.
- In the PURE study, the researchers may have started out with the expectation that eating more fat would lead to higher death rates, given the prevailing negative dogma about saturated fats.
- Later in the course we will introduce the idea of prior knowledge, which is meant to reflect the knowledge that we bring to a situation.
- Statistics provides us with a way to describe how new data can be best used to update our beliefs.

[1] As you may have realized statistics is closely related to machine learning.

Aggregation

- Another way to think of statistics is as “the science of throwing away data”.
- In the example of the PURE study above, we took more than 100,000 numbers and condensed them into ten.
- It is this kind of aggregation that is one of the most important concepts in statistics.
- Statistics provides us ways to characterize the structure of aggregates of data, and with theoretical foundations that explain why this usually works well.
- However, it’s also important to keep in mind that aggregation can go too far.
- Sometimes, a summary can provide a misleading picture of the data.

Uncertainty

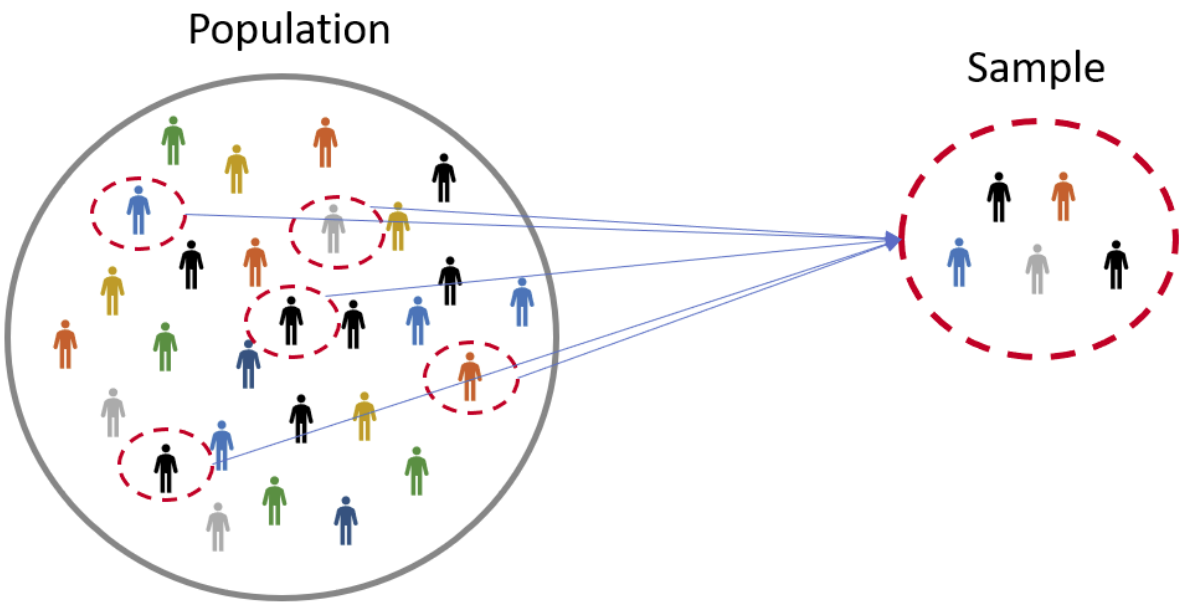
- The world is an uncertain place.
- We now know that cigarette smoking causes lung cancer, but this causation is probabilistic.
- A 68-year-old man who smoked two packs a day for the past 50 years and continues to smoke has a 15% risk of getting lung cancer.
- This is much higher than the chance of 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.

Uncertainty

- Statistics provides us with the tools to characterize 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.
- Statistics can provide us with evidence, but it’s always tentative and subject to the uncertainty that is always present in the real world.

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 summarize 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.

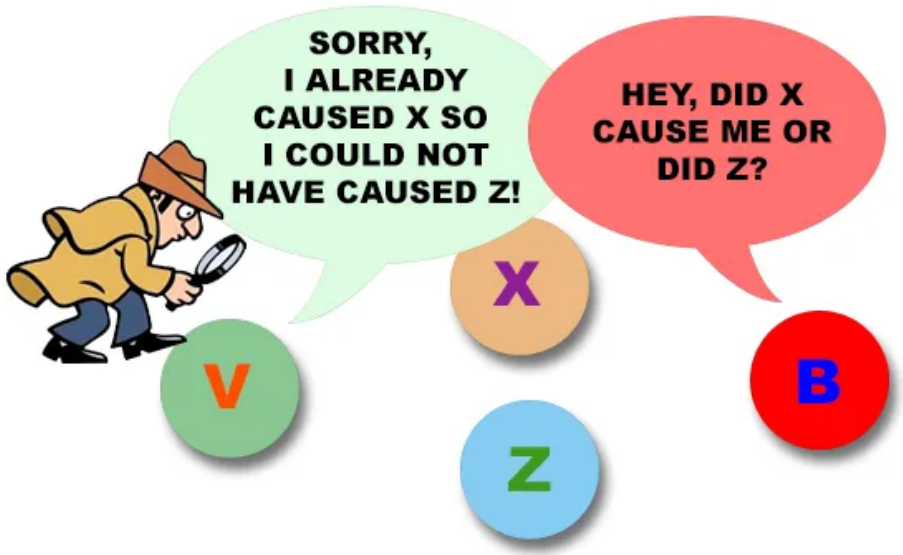


Sampling from a population

- For example, the PURE study enrolled a sample of about 135,000 people.
- But its goal was to provide insights about the billions of humans who make up the population from which those people were sampled.
- The way that the study sample is obtained is critical, as it determines how broadly we can generalize 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 in order to double the quality of our data we need to quadruple the size of our sample.

Causality and statistics

- The PURE study seemed to provide pretty strong evidence for a positive relationship between eating saturated fat and living longer.
- However, 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 are consistent with such a relationship, but they are equally consistent with some other factor causing both higher saturated fat and longer life.



Causality and statistics

- For example, it is likely that people who are richer eat more saturated fat and richer people tend to live longer.
- Their longer life is not necessarily due to fat intake though.
- 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 the claim that "correlation does not imply causation".

Causality and statistics

- 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 randomized 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 or a control group.
- Treatment group: people told to increase their saturated fat intake.
- Control group: people told to keep eating the same as before.

Causality and statistics

- 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 behaviors as well.
- We would then follow the participants over time and see how many people in each group died.
- Because we randomized 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 randomization yields treatment versus control groups that do vary in some important way.

Causality and statistics

- 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 amongst 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.

Course Philosophy

There two radical approaches to teach an introductory course on statistics.

Mathematical Statistics

- This is usually taught in mathematics departments.
- The focus is on the mathematical aspects of statistics (e.g., asymptotic properties).

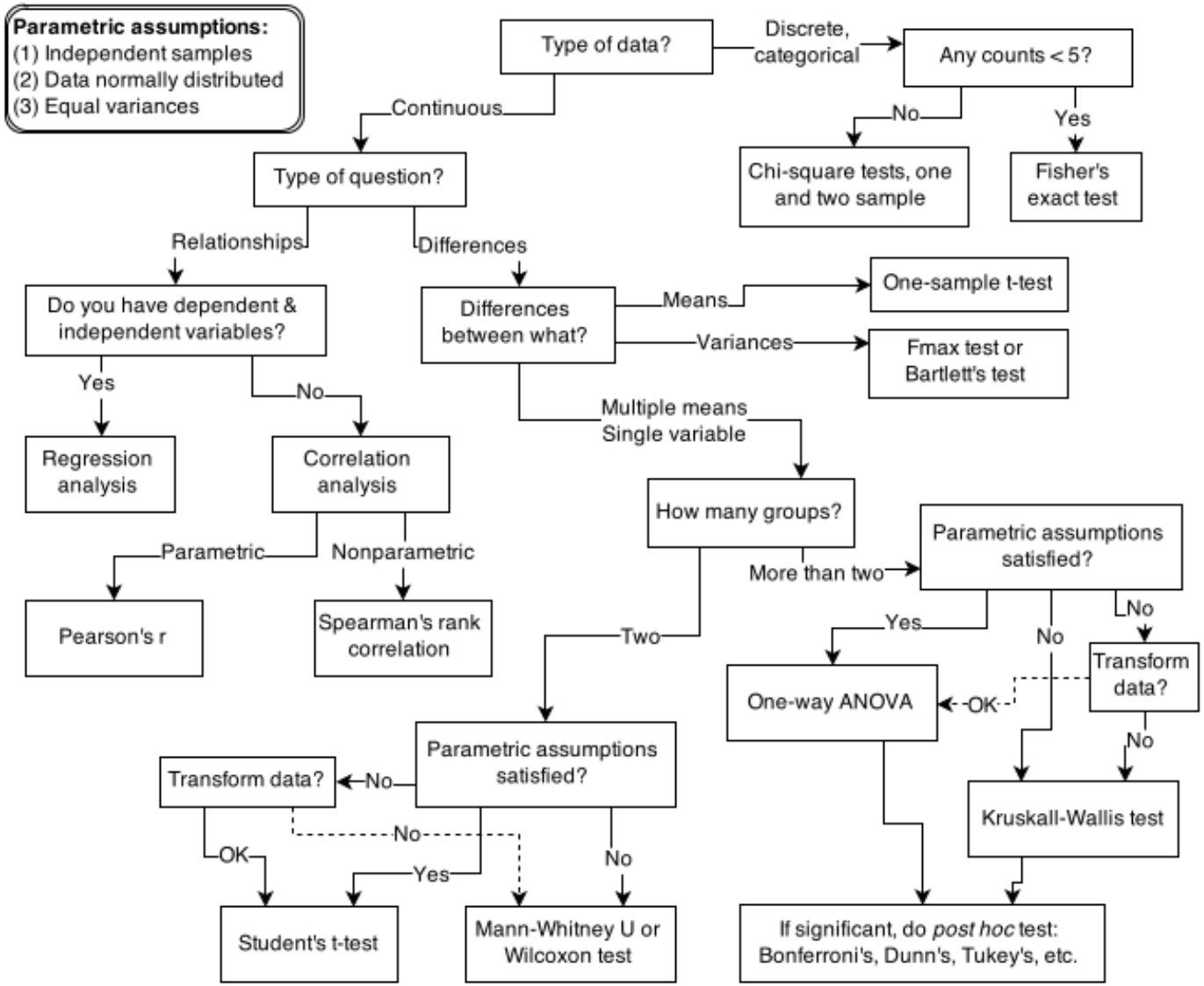
Applied Statistics

- This is usually taught in health and social science departments (e.g., psychology).
- The focus in on the methods (e.g., t-test, ANOVA, linear regression) and how to apply them to the specific field.

Course Philosophy

- In my personal opinion mathematical courses are in many cases hard to grasp for non-mathematical students and don't pay enough attention to the use of statistics in real world problems.
- On the other hand, applied courses sometimes omit the fundamental assumptions on which the methods are based and end up being a kind of recipe book of methods for ad hoc problems.
- This course attempts to bring a balance between both paradigms.
- We will reflect and discuss the fundamental assumptions and ideas on which statistical models are based (without going too deeply into the mathematics).
- We will be aware at all times that no method or tool fits all problems perfectly, or as it is commonly said: "all models are wrong but some are useful".
- Finally, we will discuss thoroughly the two major schools of statistical inference: **frequentist** and **Bayesian**.

Example of flowchart (recipe book) taken from [McElreath, 2020]



Roadmap

The main topics that will be covered in this course are:

- Statistical Programing in R
- Descriptive Statistics
- Probability
- Frequentist Inference
- Bayesian Inferece

References I


Dehghan, M., Mente, A., Zhang, X., Swaminathan, S., Li, W., Mohan, V., Iqbal, R., Kumar, R., Wentzel-Viljoen, E., Rosengren, A., et al. (2017).
Associations of fats and carbohydrate intake with cardiovascular disease and mortality in 18 countries from five continents (pure): a prospective cohort study.
The Lancet, 390(10107):2050–2062.


McElreath, R. (2020).
Statistical rethinking: A Bayesian course with examples in R and Stan.
CRC press.


Poldrack, R. A. (2019).
Statistical Thinking for the 21st Century.


Teicholz, N. (2014).
The big fat surprise: why butter, meat and cheese belong in a healthy diet.
Simon and Schuster.


Wikipedia (2021).
Statistics — Wikipedia, the free encyclopedia.
http://en.wikipedia.org/w/index.php?title=Statistics&oldid=1011947657.
[Online; accessed 19-March-2021].