6.1: Some preliminaries

Introduction

OK, you say, I get it: statistics is important, and if I am to go on as a biologist, I should learn some biostatistics. Let's get on with it, start with the equations and the problems already!

Before we review **probability theory** and introduce **risk analysis** I want to spend some time to emphasize that at issue is critical thinking, so please bear with me (Fig. 6.1.1).

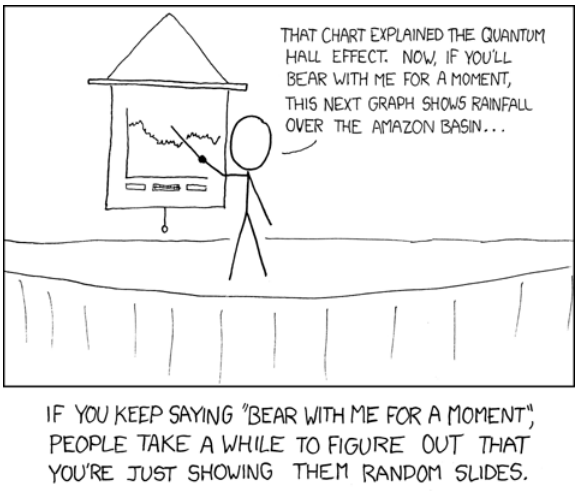

Figure 6.1.1: https://xkcd.com/365/

How likely?

As we start our journey in earnest, we start with the foundations of statistics, **probability**. Probability is something we measure, or estimate, about whether something, an **event**, will occur. We speak about *how likely* an event is to occur, and this is quantified by the probability. Probabilities are given a value between 0 and 1: at 0% chance, the event will not happen; at 100% chance, the event is certain to happen.

If probability is the chance that an event will occur, then **risk** is the probability of an event occurring over a specified period of time. I introduce our discussion of probability through a risk analysis. I like to start this discussion by relaying something I overheard, while we were all standing on a lava field on the slopes of Kilauea back in November of 1998 (Kilauea erupted with lava flow more or less continually between 1983 and 2018; on 5 Jan 2023, it started up again).

Figure 6.1.2: View of Kamokuna Lava Bench, eruption of Pu`u `O`o, Kilauea, November 1998. Photo by S. Dohm.

The Volcanoes National Park hadn't established barricades at the end of Chain of Craters Road, and people were walking to see new lava flows. The night we went, we met a park ranger who announced to us that the Park Service believed it was unsafe for us to walk out to see new lava flows because the area was unstable. Someone (not in my group), snapped back, "Oh, what are the chances that that will happen?" Of course, the ranger couldn't quote a chance between zero and 100% for that particular evening. The ranger was saying the risk had increased, based on their subjective, but experienced, opinion.

As you can see in Figure 6.1.2, we went anyway. I have been thinking about her question ever since. We were lucky — some of the same area collapsed two weeks later (USGS update 16 December 1998).

Cool picture though.

Multiple events

I have two coins, a dime and a quarter, in my pocket; when I place the coins on the table, what are the chances that both coins will show heads? A blood sample from a crime scene was typed for two **Combined DNA Index System** (CODIS) Short Tandem Repeat (STR) loci, THO1 (allele 9.3) and TPOX (allele 8), the same allele types for the defendant. What are the chances of a random match, that someone other than the defendant has the same genetic profile? For two or more **independent events** we can get the answers by using the **product rule**.

$$P\left(H_{\text{dime}} \text{ and } H_{\text{quarter}}\right) = P\left(H_{\text{dime}}\right) \times P\left(H_{\text{quarter}}\right) = 0.5 \times 0.5 = 0.25$$

The two coins are independent; therefore, the chance that both are placed heads up is 25% — we would expect to see this **combined event** one out of every four times. This is an illustration of the **counting rule**, aka **fundamental counting principle**: if there are n ways to do one thing (n elements in set A), and m ways to do another thing (m elements in set B), then there are $n \cdot m$ ways to do both things (combination of elements of A and B sets).

For the DNA profile CODIS problem (cf. Chapter 4, National Research Council 1996), the two alleles are both the most common observed in US Caucasian population at 30.45% and 54.7%, respectively (Moretti et al 2016). Assuming the individual was homozygous at both loci (i.e., $\text{THO1}_{9.3,9.3}$ and $\text{TPOX}_{8,8}$), then the genotype frequencies $\left(p^2\right)$ are:

$\text{THO1}_{9.3,9.3} = 0.3045^2 = 0.093$

$\text{TPOX}_{8,8} = 0.547^2 = 0.299$

Since THO1 is located on the p-arm of chromosome 11, and TPOX is on the p-arm of chromosome 2, the two loci are independent and therefore should be in linkage disequilibrium. We can use the product rule to get the probability of the DNA profile for the sample, 2.8%:

$$P\left(\text{THOI and TPOX}\right) = P(\text{THOI}) \times P(\text{TPOX}) = 0.093 \times 0.299 = 0.028$$

If two events are not independent, then the product rule cannot be used. For example, CODIS STR D5S818 and CSF1PO are both located on the q-arm of chromosome 5 and are therefore linked and not independent (the recombination frequency is about 0.25). The common allele for D5S818 is 11 at 40.84% and for CSF1PO the allele is 12 at 34.16%. Thus the chance of gettting the two most common alleles is not simply the product rule result of 14%; instead, we need to view this problem as one of dependent events.

Kinds of probability

So, how does one go about estimating the likelihood that a particular event will occur, whether it is the collapse of a lava delta, or that a person will have a heart attack? The probability of lava delta collapse or of heart attack are examples of **empirical probability**, as opposed to a **theoretical probability**. Despite many years of effort, we have no applicable theory that we can apply to say, if a person does this, and that, then a heart attack will happen. But we do have a body of work documenting how often heart attacks occur, and when they occur in association with certain risk factors. Similarly, progress is being made to determine markers of risk of lava field collapse (Di Traglia et al. 2018). Analogously, this is the essential goal of risk analysis in epidemiology. We know of associations between cholesterol and heart attack risk, for example, but we also know that high cholesterol does not raise the probability of the event (heart attack) to 100%. How is this **uncertainty** part of statistics? Or perhaps you are a molecular scientist in training and have learned about how to assess results of a Western blot where typically the results are scored as “yes” or “no.” How is this relevant to statistics and probability?

A misconception about statistics and statistical thinking

There’s a long history of skepticism of conclusions from health studies, in part because it seems the advice flips. For example,

- Coffee is bad for you (Medical News Today January 2008)
- Coffee is good for you (NBC News July 2018)
- Even light drinking can be harmful to health (Science Daily, January 2022)
- Seven science-backed reasons beer is good for you (NBC News August 2017)

- Meat and cheese may be as bad for you as smoking (Science Daily, March 2014)
- Cheese actually isn't bad for you (WIRED, February 2021)

The common thread is these studies are assessment of risk: about studies that seem to conclude only with statements of probability.

> **Note:**
>
> This "flipping" seems as much a function of reporting bias — the studies are not directly comparable — and may just be clickbait.

Perhaps you may have heard …? "*There are lies and then there is statistics*". The full quote reads as follows:

> Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: "There are three kinds of lies: lies, damned lies and statistics." – *Autobiography of Mark Twain* (www.twainquotes.com/Lies.html).

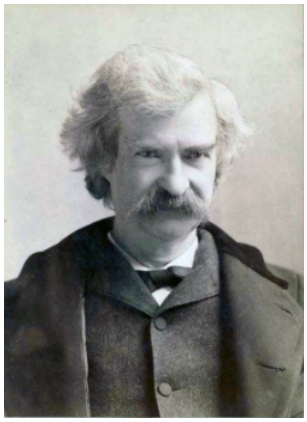

Figure 6.1.3: Mark Twain. Image from The Miriam and Ira D. Wallach Division of Art, Prints and Photographs: Photography Collection, The New York Public Library. "Mark Twain in Middle Life" The New York Public Library Digital Collections. 1860 – 1920. https://digitalcollections.nypl.org/items/510d47d9-baec-a3d9-e040-e00a18064a99

Twain attributed the remarks to Benjamin Disraeli, the prime minister of Britain during much of Queen Victoria's reign, but others have not been able to document this utterance to that effect. Still others believe that the "3-lies" quote belongs to Leonard Courtney, an English mathematician and statistician (1847-1929) (see University of York web site for more).

What is meant by this quote? Tossing aside cynicism — or healthy skepticism of authority that one should have, whether that authority is a scientist or a politician (within reason, please!) — what this quote means is that it seems that results of very similar studies are in conflict. To some, this is part of the **replication crisis** in science (Baker 2016). There's a perception that one can say just about anything with a number. Partly this is a matter of semantics, but also there is legitimacy to this concern. However, it is not necessarily the case that statistics have been intentionally done to mislead; rather, there is evidence that researchers are not always using proper statistical procedures.

One word, several meanings

We use the term statistics in multiple ways, all correct, but not all equal. For example, a statistic may refer to a number used to describe a population characteristic. From the 2010 U.S. census, we learn that the racial (self-reported) make-up of the U.S. population (then at 303 million) was 72.4% "white" and 12.6% "black" self-reported. In this sense, a statistic is something you calculate as a description. Do you recall the distinction between "statistic" and "statistics" discussed in Chapter 2?

> **Note:**
>
> This confusion is not restricted to the province of the of beginners. For example, I stumbled upon another imputation of the "Lies, Damned Lies, and Statistics" in a header of a published article (Baker et al 2014) in which the authors argued that more than 50% of papers published over a two-year period on experimental autoimmune encephalomyelitis (EAE) in rodents applied the wrong statistical procedures. The disagreement in this case had to do with **data types**; the outcome variable for EAE should be **ordinal**, but as many as half the authors reportedly (according to Baker et al) proceeded to calculate means and

conduct parametric statistical tests. Medians and not means are appropriate descriptive statistics for ordinal data types. Data types and descriptive statistics were covered in Chapter 3: Exploring data.

Secondly, two studies essentially about the same topic, yet reaching seemingly different conclusions, may differ in the assumptions employed. It should be obvious that if different assumptions are used, researchers may reach different outcomes.

Finally, how we communicate statistics can be misleading. For example, use of percentages in particular can be confusing, especially in communication of the chance that some event may happen to us (e.g., incidence of disease, or number of new cases in a specific time period, compared to prevalence of a disease, or number of cases of a disease in a specific period of time). On the one hand, percentages seem easy. A percentage is simply a proportion multiplied by 100%, and takes any value between 0% and 100%.

When a product says that it kills 99.99% of all germs on contact, do you feel better? Here's a cartoon to consider as you think about that statement.

Figure 6.1.4: xkcd comic strip, from https://imgs.xkcd.com/comics/hand_sanitizer.png

Of course numbers cannot be used to justify simultaneously mutually exclusive conclusions, but being able to recognize careless (or deliberate) miscommunication with numbers, well, this needs to be part of your skill set. As you read this next section, I ask that you consider:

- are the correct statistical descriptors in use?
- what assumptions are being made?
- does the reporting of percentages lead to clear conclusions?

Some concluding thoughts about "lies and statistics"

Statistics is tricky because there are assumptions to be made. And you have to be clear in your thinking.

If the assumptions hold true, then we aren't lying, and Twain had it wrong.

But if we disagree on the assumptions, then we will necessarily have to disagree on the conclusions drawn from the calculated numbers (the statistics). Risk analysis in particular, but statistics in general, is a tricky business because many assumptions need to be made, and we won't necessarily have all of the relevant information available to make sure our assumptions are truthful. But it is the assumptions that matter: if we agree with the assumptions that are made, then we have confidence in the conclusions drawn from the statistics.

In a typical statistics course, we would spend a bunch of time on probability. We will here as well, but in the context of risk analysis and in the other contexts, in a less than formal presentation on the subject of probability. For example, in talking about inference, the testing of null hypotheses and estimating the probability that the null hypothesis is correct, I will say things like, "Imagine we repeated the experiment a million times — how many times by chance would we think a correct null hypothesis would nonetheless be rejected?"

There's no real substitute for a formal course in probability theory, and you should be aware that this foundation is pretty important if you go forward with biostatistics and epidemiology. For now, I will simply refer you to chapters 1, 2 and 3 of a really nice online book on probability from one of the masters, Richard Jeffrey (1926 – 2002; click here to go to Wikipedia). Much of what I will present to you follows from similar discussions.

My aim is to teach you what you need about probability theory by the doing. In the next couple of days we will deal with an aspect of risk analysis, namely a consideration of CONDITIONAL PROBABILITY, and Baye's Theorem that will help you evaluate claims such as the one made for airline safety. Risk analysis is tricky, but it is not a subject above and beyond our abilities; by applying some of the rules of statistical reasoning, we can check claims based on statistics. A healthy degree of skepticism is part of becoming a scientist. Do try this at home!

Some examples to consider: what is the relative risk for the following scenarios?

1. Drug testing at the workplace: risk of a worker who does not use illegal drugs registering positive (false positives in drug testing);

2. Positive HIV from blood sample from USA male with no associated risks (e.g., intravenous drug user), false positives in HIV testing; false positives with mammography;

3. Benefits versus risks of taking a statin drug (drug that reduces serum cholesterol levels) to a person with no history of heart disease;

4. Is it safer to travel by car or by airliner? We'll break this problem down in the next section.

What we are looking for is the probability of an occurrence of a particular event, e.g., that a person who does not use illegal drugs may nonetheless test positive; we are looking for a way to make rational decisions and understanding probability is the foundation.

Questions

1. What do you make of the claim (joke) that "*There are lies and then there are statistics*?"
2. For the various proportions listed, can these also be considered to be rates?
3. Distinguish between empirical and theoretical probability; use examples.
4. CODIS STR D5S818 and CSF1PO are both located on the q-arm of chromosome 5, and are therefore linked and not independent (the recombination frequency is about 0.25). The common allele for D5S818 is 11 at 40.84% and for CSF1PO the allele is 12 at 34.16%. Given that the person has allele 11 at D5S818 (genotype 11,11), what are the chances that they also have allele 12 at CSF1PO (genotype 12,12)?
