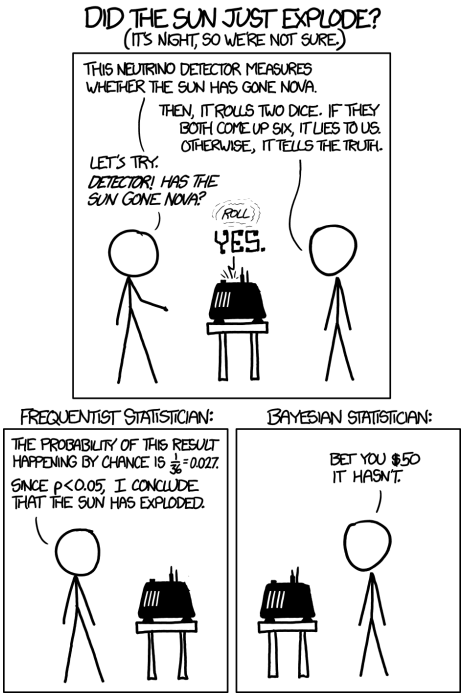
Chapter 6.1 - Uncertainty

Rather than the standard quote to start things off, I think this XKCD comic does a much better job introducing the theme of this section.



If you wanted to criticize the entire field of statistics you would only have to focus on the complete lack of unity in our methods. Almost every other science has some degree of unity in their fundamental theorems and philosophies: physics considers one unified model of relativity, biology embraces the theory of evolution, chemists generally accept the existence of sub-atomic particles, and engineers overwhelmingly agree that money is very nice to have.

Statisticians, however, can't agree on much at all. The worst part is that our philosophical arguments are so persistent and escalatory that the other sciences can't keep up with the drama. If you asked any tenured professor outside of statistics what they believed the hottest debate in statistics was they would likely say: "Whether *Frequentist* or *Bayesian* statistics is better!". That'd be the correct answer if this textbook was written in the year 2000.

These days that discussion has fizzled out to a steady statement of: "Both methods are valid and asymptotically equivalent". The validity of that statement is worth a far greater dialogue— somewhere else besides here. That said, it's quite difficult to discuss the subsequent passages of this book without explaining the theoretical basis of uncertainty quantification.

Since all statistics stems from mathematics and all mathematics begins with philosophy we have to engage in a philosophical discussion.

In my endless naiveté I believed that the preparation of this chapter would be painless.

Differing Perspectives

> A doctor is stuck in a room with 3 patients who have been exposed to a **perfectly lethal** (0% survival) and **perfectly transmissible** (any exposure leads to inoculation) virus that only occurs in 0.15% of the population. The doctor tests all 3 patients for the pathogen and finds that one of them is *positive* while the other two are *negative*. If a patient is sick the test will **always** be positive but if the patient is healthy the test will be negative 95% of the time. The doctor has 4 doses of anti-viral available to them that will **always cure** the illness and prevent death, **given that the pathogen is present**. If the pathogen **is not present** the anti-viral will **always kill** the recipient. Who should the doctor administer the anti-viral to?

The methods detailed in this book so far would lead us to a simple solution; we administer the anti-viral to the positive patient since the probability of it killing them is only 5%. While this assumes that a 5% chance is an acceptable level of *risk*— we can confidently make that assumption for most people facing this decision. In this case the additional information isn't necessary to solve the problem. The perspective that we don't need to consider the additional information to solve the problem can be considered *frequentist* in nature.

Frequentist: Methods using the perspective that the probability of an event is defined as the relative frequency of that event's occurence across *infinite* trials.

Frequentist methods are so commonly taught that many consider them to be the default state of statistics. While the methods we've covered up until this point haven't been *inherently* frequentist, they are the fundamental building blocks for the frequentist framework. There are two general reasons these methods are so prevalent:

1. Frequentist methods are (typically) analytically simple.
2. The theories behind frequentist methods are very accessible.

These two reasons result in frequentism being **much easier** to teach than other statistical methodologies; if something is easy to teach it tends to be taught *a lot*. But this idea of accepting only the data we're presented and using some defined method of assigning probabilities *should* make us uncomfortable in this scenario. If the stakes were lower we might be okay with accepting our known methods but there **must** be some way of considering all of this additional information. The perspective that *previously* or *prior* known information can be incorporated into the assignment of probability is the basis for *Bayesian* methods.

Bayesian: Methods using the perspective that the probability of an event is the result of comparing the likelihood and prior knowledge of the event with the observed evidence.

Bayesian methods are the product the philosopher (among other professions/titles) Thomas Bayes. Bayes developed a philosophical and mathematical framework for assigning probabilities using propositional logic rather than strictly quantitative techniques. We've talked about these methods before and the conclusion of that (brief) discussion was that they were hopelessly complex. We'll need to accept that this is still true despite how simple the implementation is about to look.

Bayes' Rule

To solve this problem using bayesian techniques we only need **Bayes' rule**:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Which can be simplified to "plain English" as:

$$\text{Posterior} \ = \ \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

The rule feels like an extended definition of conditional probability, which it is, but there's context we need to apply to the equation in order to understand how it *differs* from a basic definition. For instance the **likelihood** can be multiple things: the likelihood function of a distribution (the primary case for Bayesian modeling), the likelihood of event B occurring upon observation of A, the degree with which we *believe* the event to be probable. The **prior** is almost always an assumption we place on the warrant (often the data generating process) and the **evidence** is some observation of reality. We sometimes refer to the evidence as the **marginal** since the equation uses conditional probability assignments and this is the standard vocabulary in conditional probability for individual event probabilities. The **posterior** is the result of interest which makes this equation *implicit*.

The marginal is the most problematic component of Bayes' Rule since it's often the only one besides the posterior that's *unknown*. While it might seem confusing to refer to observed evidence as unknown but this is exactly what produces the "magic" of Bayesian methods. We propose assumptions about the process we seek to understand as a *substitute* for pure observation of data. In order to proceed with calculating the equation by using those assumptions we have to solve for the marginal using the **law of total probability**:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

When we fully expand these formulas we can see the rationale behind the law:

$$P(B) = \frac{P(B \cap A)}{P(A)} \times P(A) + \frac{P(B \cap A^c)}{P(A^c)} \times P(A^c)$$

$$P(B) = P(B \cap A) + P(B \cap A^c)$$

This is true when we consider the definition:

$$P(A) + P(A^c) = P(S)$$

Since we're intersecting A and everything that **isn't** A the equation is analogous to:

$$P(B) = P(B \cap S)$$

Which is generally stating "B is equivalent to all of the spaces in the sample space where B is". As this is the definition of an event in a sample space we've done a lot of back flips and cartwheels to define B without ever observing B independently.

Plugging this into Bayes' rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

With this we can stage our attack on this problem! We're interesting in discovering the probability that the patient is sick *given that* they tested positive.

$$\begin{aligned}
\text{Let} \quad & D^+ \equiv \text{Disease Positive} \\
& D^- \equiv \text{Disease Negative} \\
& T^+ \equiv \text{Test Positive} \\
& T^- \equiv \text{Test Negative} \\
& \\
& P(D^+) = 0.0015 \\
& P(D^-) = 0.9985 \\
& P(T^+|D^+) = 1.00 \\
& P(T^-|D^-) = 0.95 \\
& P(T^+|D^-) = 0.05
\end{aligned}$$

$$P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)}$$

$$= \frac{1 \times 0.0015}{1 \times 0.0015 + 0.05 \times 0.9985}$$

$$= 0.02917$$

By this logic there's a 97.083% chance that if we administer the anti-viral to the positive patient it will **kill** them. Why is this? What we've done is apply the additional context of the tests probability of false positives *and* the rarity of the virus. Even though the chance of a false positive is very low the chance of being positive to begin with is *even lower*. Bayes' rule allows us to apply this prior knowledge to inferential process and get a wildly different result from frequentist methods.

Errors in Estimation

It's at this point that many of us might be worried there's an error in the text, possibly screaming at the inconsistency or smugly believing they've cracked the riddle. The original problem stated two related facts right away:

1. All of the patients have been exposed to the virus.
2. The virus is perfectly transmissible, any exposure leads to inoculation.

Our inference from both methods were *completely wrong*. Everyone is sick so everyone should get the anti-viral, including the doctor. This is more than a simple check of reading comprehension it's a chance to look at the reality of statistics;

> "All models are wrong, but some are useful" - George Box

Everything we do in statistics is "wrong" regardless of the method. Thus, statisticians aren't so much concerned with the accuracy of an estimate but rather the degree with which it is wrong. Depending on the statistical methods you're using you'll quantify that "wrongness" (which we refer to as uncertainty) in different ways.

If all 3 patients are sick but only 1 tested positive then the test must be capable of producing false negatives. The frequentist perspective of uncertainty would have us *test the test* to see if it's truly 100% sensitive for true positives. This is the technique we'll be covering throughout chapters 6 and 7, we check the uncertainty associated with the methods we've used rather than the estimates they produce.

The reason I've introduced Bayesian methods is because the Bayesian perspective of uncertainty is the one student's often *confuse* to be frequentist. Bayesian methods quantify the uncertainty of the *proposition* rather than the method. If we were to calculate the uncertainty for each method:

- Frequentist uncertainty would be a "confidence interval", interpreted as the probability that the method (the test for the pathogen) is going to produce the correct result.
- Bayesian uncertainty would be a "credible interval", interpreted as the probability that the estimate (the proposition that the patient is disease positive) is the truth.

Both of these are scientifically valid and in many advanced statistics courses students go through the exercise of proving that they *converge* to the same results as the experiments are repeated *ad infinitum*. For our purposes we should recognize that the *approach* we take when modeling processes defines how our *uncertainty* is interpreted. If we produce scientific inference from flawed understanding of statistical inference we're destined to create horrendous errors (and thus, bad science).

Making Assumptions

This problem highlights another important feature of all statistical inference, and thus all science. No matter what approach we take to solving this problem there is no *strictly correct* answer without accepting some feature of the problem as **truth**. In order to proceed with quantifying the uncertainty in the test we need to operate under the assumption that the virus is perfectly transmissible. If we wish to challenge the statement that exposure guarantees inoculation then we have to assume that the test is truly 100% sensitive for true positives and 95% specific for true negatives. Even if we were so allergic to assuming anything about the virus and the test we would still have to operate under *one of the remaining assumptions*. Were all 3 patients actually exposed? Is the anti-viral guaranteed to cure when someone's sick and kill when they're healthy?

The debate between statistical methods generally boils down to a bizarre argument between "subjectivity" and "objectivity". Yet, in order to perform any worthwhile science we must always accept some degree of *risk*. It would be a tragedy for the remainder of scientific progress to consist of repeatedly proving everything we've ever discovered or concluded to be correct. While we should strive to be as correct as possible there is no science without accepting the chance of being blatantly incorrect.

> "The subjectivist (i.e. Bayesian) states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science."
> - I.J. Good

Statistics is the result of using logic and mathematics to assist with scientific inference. Logic, mathematics, and science as a whole are the endeavors of humans. As humans are inherently imperfect, so are these endeavors.