Conditional probability
18.600 Problem Set 3, due March 1

Welcome to your third 18.600 problem set! Conditional probability is defined by $P(A|B) = P(AB)/P(B)$, which implies

$$P(B)P(A|B) = P(AB) = P(A)P(B|A),$$

and dividing both sides by $P(B)$ gives Bayes' rule:

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

which we may view as either a boring tautology or (after spending a few hours online reading about Bayesian epistemology, Bayesian statistics, etc.) the universal recipe for revising a worldview in response to new information. Bayes' rule relates $P(A)$ (our Bayesian *prior*) to $P(A|B)$ (our Bayesian *posterior* for A, once B is given). If we embrace the idea that our brains have subjective probabilities for *everything* (existence of aliens, next year's interest rates, Sunday's football scores) we can imagine that our minds continually use Bayes' rule to update these numbers. Or least that they would if we were clever enough to process all the data coming our way.

By way of illustration, here's a fanciful example. Imagine that in a certain world, a *normal* person says 10^5 things per year, each of which has a 10^{-5} chance (independently of all others) of being truly horrible. A *truly horrible* person says 10^5 things, each of which has a 10^{-2} chance (independently of all others) of being truly horrible. Ten percent of the people in this world are truly horrible. Suppose we meet someone on the bus and the first thing that person says is truly horrible. Using Bayes' rule, we conclude that this is probably a truly horrible person.

Then we turn on cable news and see an unfamiliar politician saying something truly horrible. Now we're less confident. We don't know how the quote was selected. Perhaps the politician has made 10^5 recorded statements and we are seeing the only truly horrible one. So we make the quote selection mechanism part of our sample space and do a more complex calculation.

The problem of selectively released information appears in many contexts. For example, lawyers select evidence to influence how judges and jurors calculate conditional probability *given* that evidence. If I'm trying to convince you that a number you don't know (but which I know to be 49) is prime, I could give you some selective information about the number without telling you exactly what it is (it's a positive integer, not a multiple of 2 or 3 or 5, less than 50) and if you don't consider my motives, you'll say "It's probably prime."

Note also that legal systems around the world designate various "burdens of proof" including *probable cause*, *reasonable suspicion*, *reasonable doubt*, *beyond a shadow of a doubt*, *clear and convincing evidence*, *some credible evidence*, and *reasonable to believe.* Usually, these terms lack clear meaning as numerical probabilities (does "beyond reasonable doubt" mean with probability at least .95, or at least .99, or something else?) but there is an exception: *preponderance of evidence* generally indicates that a probability is greater than fifty percent, so that something can be said to be “more likely than not.” An interesting question (which I am not qualified to answer) is whether numerical probabilities should be assigned to the other terms as well.

Please stop by my weekly office hours (2-249, Wednesday 3 to 5) for discussion.

A. FROM TEXTBOOK CHAPTER THREE:

1. Problem 43: There are 3 coins in a box. One is a two-headed coin, another is a fair coin, and the third is a biased coin that comes up heads 75 percent of the time. When one of the 3 coins is selected at random and flipped, it shows heads. What is the probability that it was the two-headed coin?

B. A medical practice uses a “rapid influenza diagonistic test” to get a quick (under 30 minute) assessment of whether a patient has the flu. The **sensitivity** of the test (i.e., the fraction of the time it returns a positive result if the patient has the disease) is .5 while the **specificity** (i.e., the fraction of the time it returns a negative result if the person does not have the flu) is .9. In other words, people without the flu are relatively unlikely (10 percent chance) to get a false positive, but people *with* the flu have a larger chance (50 percent) to get a false negative (e.g., because the particular strain of flu isn’t picked up by the test, or virus somehow didn’t make it onto the swab).

Suppose that based on time of year and symptoms (fever, chills, cough, etc.) the doctor thinks *a priori* that the event F that a patient has the flu has probability $P(F) = .8$. Assume further that the doctor believes that the specificity/sensitivity results mentioned above apply to this individual (given what is known), so that if T is the event that the test comes back positive, we have $P(T|F) = .5$ and $P(T|F^c) = .1$. After the doctor administers the test and discovers that the test is negative, what is the doctor’s *a posteriori* estimate of the probability that the patient has the flu? In other words, what is $P(F|T^c)$? What is $P(F|T)$? Give approximate percentages.

Remark: Google *RIDT specificity and sensitivity* to see actual estimates of these values (which vary with the study, the type of test, the flu strains prevalent in a given year, etc.) One practical decision a doctor might make is whether to prescribe an antiviral medication (like Tamiflu) that is thought to reduce symptom duration by about one day on average *if* a person has the flu, and zero days otherwise. (By comparison, a flu vaccine that reduces the risk of a 10-symptom-day flu during a season from 20 percent to 10 percent would also decrease the expected number of symptom days by one. Google *Tamiflu effectivness* and *flu vaccine effectiveness* for actual data on such things, which apparently vary quite a lot from year to year and place to place.) One might guess that if a person has a p chance of having the flu, then taking the drug decreases the expected number of symptom days by p (one day on average if flu is really there, zero otherwise). If p is below some threshold the doctor and patient may conclude that the cost (time it takes to fill prescription, price of drug, small side effect risk) just isn’t worth it. (Another hard-to-measure consideration: how much do vaccines/antivirals decrease the risk of infecting others?)

C. Every customer at Lavinia’s Diner orders one of seven types of beverages and one of ten types of sandwishes. For $1 \leq i \leq 7$ and $1 \leq j \leq 10$, let $p_{i,j}$ denote the probability that a customer selects the ith type of beverage and the jth type of sandwich. Show that sandwich type and beverage type are *independent* if and only if, as a 7 by 10 matrix, $(p_{i,j})$ has rank one (i.e., there is some column of the matrix such that each of the other columns is a constant multiple of that one).

D. Suppose that a fair coin is tossed infinitely many times, independently. Let X_i denote the outcome of the ith coin toss (an element of $\{H, T\}$). Compute the probability that:

1. $X_i = H$ for all $i \geq 0$
2. There exists an infinite arithmetic progression such that $X_i = H$ for all i in that arithmetic progression. In other words, there exist positive integers a and b such $X_i = H$ whenever $i \in \{a, a+b, a+2b, a+3b, a+4b, \ldots\}$. (Hint: use the countably additivity axiom.)
3. The pattern HHTTHHTTTH occurs at some point in the sequence $X_1, X_2, X_3, \ldots$.

E. Suppose that the quantities $P[A|X_1], P[A|X_2], \ldots, P[A|X_k]$ are all equal. Check that $P[X_i|A]$ is proportional to $P[X_i]$. In other words, check that the ratio $P[X_i|A]/P[X_i]$ does not depend on i. (This requires no assumptions about whether the X_i are mutually exclusive.)

Remark: This can be viewed as a mathematical version of Occam's razor. We view A as an "observed" event and each X_i as an event that might "explain" A. What we showed is that if each X_i "explains" A equally well (i.e., $P(A|X_i)$ doesn't depend on i) then the conditional probability of X_i *given* A is proportional to how likely X_i was a *a priori*. For example, suppose A is the event that there are certain noises in my attice, X_1 is the event that there are squirrels there, and X_2 is the event that there are noisy ghosts. I might say that $P(X_1|A) >> P(X_2|A)$ because $P(X_1) >> P(X_2)$. Note that after looking up online definitions of "Occam's razor" you might conclude that it refers to the above tautology *plus* the common sense rule of thumb that $P(X_1) > P(X_2)$ when X_1 is "simpler" than X_2 or "requires fewer assumptions."

F. On Cautious Science Planet, science is done as follows. First, a team of wise and well informed experts concocts a hypothesis. Experience suggests the hypotheses produced this way are correct ninety percent of the time, so we write $P(H) = .9$ where H is the event that the hypothesis is true. Before releasing these hypotheses to the public, scientists do an additional experimental test (such as a clinical trial or a lab study). They decide in advance what constitutes a "positive" outcome to the experiment. Let T be the event that the positive outcome occurs. The test is constructed so that $P(T|H) = .95$ but $P(T|H^c) = .05$. The result is only announced to the public if the test is positive. (Sometimes the test involves checking whether an empirically observed quantity is "statistically significant." The quantity $P(T|H)$ is sometimes called the *power* of the test.)

(a) Compute $P(H|T)$. This tells us what fraction of published findings we expect to be correct.

(b) On Cautious Science Planet, results have to be replicated before they are used in practice. If the first test is positive, a second test is done. Write $\tilde{T}$ for the event that the second test is positive, and assume the second test is like the first test, so that $P(\tilde{T}|HT) = .95$ but $P(\tilde{T}|H^cT) = .05$. Compute the reproducibility rate $P(\tilde{T}|T)$.

(c) Compute $P(H|T\tilde{T})$. This tells us how reliable the replicated results are. (Pretty reliable, it turns out—your answer should be close to 1.)

On Speculative Science Planet, science is done as follows. First creative experts think of a hypothesis that would be rather surprising and interesting if true. These hypotheses are correct only five percent of the time, so we write $P(H) = .05$. Then they conduct a test. This time $P(T|H) = .8$ (lower power) but again $P(T|H^c) = .05$. Using these new parameters:

(d) Compute $P(H|T)$.

(e) Compute the reproducibility rate $P(\tilde{T}|T)$. Assume the second test is like the first test, so that $P(\tilde{T}|HT) = .8$ but $P(\tilde{T}|H^cT) = .05$.

Remark: If you google Nosek reproducibility you can learn about one attempt to systematically reproduce 100 psychology studies, which succeeded a bit less than 40 perent of the time. Note that $P(\tilde{T}|T) \approx .4$ is (for better or worse) closer to Speculative Science Planet than Cautious Science Planet. The possibility that $P(H|T) < 1/2$ for real world science was famously discussed in a paper called *Why Most Published Research Findings Are False* by Ioannidis in 2005. A more recent mass replication attempt (involving just *Science* and *Nature*) allowed scientists to bet on whether a study would be replicated and found that to some extent scientists were good at predicting such things. See `https://www.nature.com/articles/d41586-018-06075-z`.

Questions for thought: What are the pros and cons of the two planets? Is it necessarily bad for $P(\tilde{T}|T)$ and $P(H|T)$ to be low in some contexts (assuming that people know this and don't put too much trust in single studies)? Do we need to do larger and more careful studies? What improvements can be made in fields like medicine, where controlled clinical data is sparse and expensive but life and death decisions have to be made nonetheless? And I do mean expensive. The cost of recruiting and pre-screening a *single* Alzheimer's patient for trial is \$100,000, per this article `https://www.nytimes.com/2018/07/23/health/alzheimers-treatments-trials.html` These questions go well beyond the scope of this course, but we will say a bit more about the tradeoffs involved when we study the central limit theorem.

G. **Doomsday:** Many people think it is likely that intelligent alien civilizations exist *somewhere* (though perhaps so far separated from us in space in time that we will never encounter them). When a species becomes roughly as advanced and intelligent as our own, how long does it typically survive before extinction? A few thousand years? A few millions years? A few billion years? Closely related question: how many members of such a species typically get to exist before it goes extinct?

Let's consider a related problem. Suppose that one factory has produced a million baseball cards in 10,000 batches of 100. Each batch is numbered from 1 to 100. Another factory has produced a million baseball card in 1,000 batches of 1,000, each batch numbered from 1 to 1,000. A third factory produced a million baseball card in 100 batches of 10,000, with each batch numbered from one to 10,000. You chance upon a baseball card from one of these three factories, and *a priori* you think it is equally likely to come from each of the three factories. Then you notice that the number on it is 76.

(a) Given the number you have seen, what is the conditional probability that the card comes from the first factory? The second? The third?

Now consider the following as a variant of the card problem. Suppose that one universe contains 10^{30} intelligent beings, grouped into civilizations of size 10^{12} each. Another universe contains 10^{30} intelligent beings, grouped into civiliations of size 10^{15} each. A final universe contains 10^{30} intelligent beings, grouped into civilizations of size 10^{18} each. You pick a random one of these 3×10^{30} beings and learn that before this being was born, exactly $141,452,234,521$ other beings were born in its civilization.

(b) What is the conditional probability that the being comes from the first universe?

Remark: The *doomsday argument* (google it) is that it is relatively likely that human civilization will disappear within thousands of years — as opposed to lasting millions of years — for the following reason: *if* advanced civilizations typically lasted for millions of years (with perhaps 10 billion beings born per century), then it would seem *coincidental* for us to find ourselves among the first few thousand. People disagree on what to make of this argument (what the Bayesian prior on civilization length should be, what to do with all the other information we have about our world, what measure to put on the set of alternative universes, etc.) But maybe we should at least consider the *possibility* of near-term human extinction, and whether preparing for apocalyptic scenarios (giant asteroids, incurable plagues, nuclear war, climate disaster, resource depletion, the next ice age, etc.) might improve our chance of surviving a few thousand (or million or billion) more years.