

Conditional probability

18.600 Problem Set 3, due February 26

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 information 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 television 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, which might also factor in a belief that politicians are more (or less) likely *a priori* to be truly horrible than bus riders.

Lawyers select evidence to influence how judges and jurors calculate conditional probability *given* that evidence. Listeners try to take this into account. They assess the probability that — given that a person is innocent — there is still *some* collection of true facts such that conditioned on those facts the person looks guilty. Real life Bayesian reasoning is complicated.

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

A. FROM TEXTBOOK CHAPTER THREE:

1. Problem 26: Suppose that 5 percent of men and .25 percent of women are color blind. A color-blind person is chosen at random. What is the probability of this person being male? Assume that there are an equal number of males and females. What if the population consisted of twice as many males as females?
2. 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?
3. Problem 47: An urn contains 5 white and 10 black balls. A fair die is rolled and that number of balls is randomly chosen from the urn. What is the probability that all of the balls selected are white? What is the conditional probability that the die landed on 3 if all the balls selected are white?
4. Theoretical Exercise 24: A round-robin tournament of n contestants is a tournament in which each of the $\binom{n}{2}$ pairs of contestants play each other exactly once, with the outcome of any play being that one of the contestants wins and the other loses. For a fixed integer k , $k < n$, a question of interest is whether it is possible that the tournament outcome is such that, for every set of k players, there is a player who beat each member of that set. Show that if

$$\binom{n}{k} \left[1 - \left(\frac{1}{2}\right)^k\right]^{n-k} < 1$$

then such an outcome is possible. *Hint:* Suppose that the results of the games are independent and that each game is equally likely to be won by either contestant. Number the $\binom{n}{k}$ sets of k contestants, and let B_i denote the event that no contestant beat all of the k players in the i th set. Then use Boole's inequality to bound $P(\cup_i B_i)$.

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

1. $X_i = H$ for all positive integers i .

2. The pattern HHTTHHTT occurs at some point in the sequence X_1, X_2, X_3, \dots

C. Two unfair dice are tossed. Let $p_{i,j}$, for i and j in $\{1, 2, 3, 4, 5, 6\}$, denote the probability that the first die comes up i and the second j . Suppose that for any i and j in $\{1, 2, 3, 4, 5, 6\}$ the event that the first die comes up i is independent of the event that the second die comes up j . Show that this independence implies that, as a 6 by 6 matrix, $p_{i,j}$ has rank one (i.e., show that there is some column of the matrix such that each of the other five column vectors is a constant multiple of that one).

D. Suppose that the quantities $P[A|X_1], P[A|X_2], \dots, P[A|X_k]$ 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 priori*. For example, suppose A is the event that there are certain noises in my attic, 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 $P(X_1|A) \gg P(X_2|A)$ because $P(X_1) \gg P(X_2)$. 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."

E. 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 actually 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 percent 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.

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? 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.