

Conditional probability

18.600 Problem Set 3, due September 26

Welcome to your third 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.

A. Let's think about the conditional probabilities associated to different possible carnival games.

1. A game involves tossing a fair coin 10 times. You win if all 10 are heads. What is the *conditional probability* that all 10 are heads *given* that the *first* 9 are heads? What is the *conditional probability* that all 10 are heads *given* that *at least* 9 (not necessarily consecutive) are heads?
2. At the next carnival booth over, there is an urn with 100 balls: 90 white and 10 black. You reach in and choose 10 at random, with all of the $\binom{100}{10}$ subsets being equally likely. Let E_k be the event that exactly k of your balls are black (and hence $10 - k$ are white). Compute $P(E_k)$ as a function of k for $k \in \{0, 1, \dots, 10\}$. Use wolframalpha to give a table of values. You can use notation like `N[Table[(10 choose k) (90 choose 10-k)/(100 choose 10), {k,0,10}]]`
3. Let A_k be the event that *at least* k balls are black. Compute the conditional probability $P(A_{10}|A_9)$ both with and without wolframalpha. Give an exact rational number. Explain intuitively why this differs from Part 1, and why it is similar to the Powerball question.

B. FROM TEXTBOOK CHAPTER THREE:

1. 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 $P(\cup_i B_i)$

C. On Interrogation Planet, there are 730 suspects, and it is known that exactly one of them is guilty of a crime. It is also known that any time you ask a guilty person a question, that person will give a "suspicious-sounding" answer with probability .9 and a "normal-sounding" answer with probability .1. Similarly, any time you ask an innocent person a question, that person will give a suspicious-sounding answer with probability .1 and a normal-sounding answer with probability .9. (And these probabilities apply *regardless* of how the suspect has answered questions in the past; in other words, once a person's guilt or innocence is fixed, that person's answers are *independent* from one question to the next.)

Interrogators pick a suspect at random (all 730 people being equally likely) and ask that person nine questions. The first three answers sound normal but the next six answers all sound suspicious. The interrogators say "Wow, six suspicious answers in a row. Only a one in a million chance we'd see that from an innocent person. This person is obviously guilty." But you want to do some more thinking. Given the answers thus far, compute the conditional probability that the suspect is guilty. Give an exact numerical answer.

D. In population genetics, one may [model](#) the [matrilineal descendant tree](#) in a population beginning with a single female (often called “Eve”).¹ Fix p_0, p_1, p_2, \dots with $\sum_{j=0}^{\infty} p_j = 1$. Assume that for each $j \geq 0$, the probability that Eve has exactly j daughters is p_j . Write E_k for the event that Eve has a “ k th daughter” (i.e., that Eve has at least k daughters). Then $P(E_k) = a_k$ where $a_k := \sum_{j=k}^{\infty} p_j$.

1. Let $R = \sum_{j=0}^{\infty} j p_j$ be the expected number of daughters Eve will have. Check that $R = \sum_1^{\infty} a_k$.

More generally, let E_{k_1, k_2, \dots, k_m} denote the event that Eve has a k_1 th daughter *and* that daughter has a k_2 th daughter *and* so on until the m th generation m (i.e., distance m from Eve in the family tree).

Assume that *given* that a female is born, her probability of having at least k children is the same as Eve’s was—in other words, assume that $P(E_{k_1, k_2, \dots, k_{m-1}, k_m} | E_{k_1, k_2, \dots, k_{m-1}}) = a_{k_m}$.

2. Use the above equation and induction on m to show that $P(E_{k_1, k_2, \dots, k_m}) = a_{k_1} a_{k_2} \cdots a_{k_m}$.
3. Show that Eve expects to have R^2 granddaughters and R^3 greatgranddaughters—and generally R^m descendants in generation m .
4. Conclude that if $R < 1$ then the matrilineal descendant tree is almost surely finite, and the expected number of females in the tree (counting Eve herself) is $1 + R + R^2 + \dots = 1/(1 - R)$.
5. Imagine that after her own birth, Eve lives an integer number of years, and the probability that she lives *at least* k years is b_k (where b_k decreases to zero as $k \rightarrow \infty$ since Eve is mortal). Assume that *given* that she lives at least k years, she expects to have d_k daughters (and t_k children total) during her k th year. Argue that her overall expected number of daughters is $R = \sum_{k=1}^{\infty} b_k d_k$.

Remark: Each year, demographers observe age-specific death/fertility rates and work out what R and $T = \sum t_k$ would be if these age-specific rates were the same every year. The resulting R value is called the [net reproduction rate](#) while T is called the [total fertility rate](#) (it includes male *and* female births and does not account for deaths). [Replacement-level fertility](#) is $R = 1$, which corresponds to $T \approx 2.1$ (in developed countries where early death is less common) or $T \approx 2.3$ (globally). In times/places with R well above 1, leaders may try to reduce fertility. In times/places with R well below 1, leaders may try to boost fertility. The global R value is currently close to 1 (good news for sustainability) but *fertility inequality* between nations is high, with [almost a factor of eight](#) separating e.g. DR Congo and South Korea. Migration from high to low fertility countries is politically limited ([net between \$\pm 3\%\$ /year for most countries](#)) and high fertility has [boosted Africa recently](#). [UN forecasters expect](#) regional fertility gaps to narrow tremendously over this century. ([Stein’s law?](#)) But [forecasters admit great uncertainty](#), despite some aspects of demographic forecasting being easier than (say) stock market forecasting—e.g. to bound the number of Japanese-born 50-year-olds in 2075, just count the Japanese births in 2025.

Remark: If there are K founders and the total future population (including founders) is $N \gg K$, then there must be $N - K$ future births, so the *average* female has $(N - K)/N \approx 1$ daughters. So *if* the number of descendants is large but finite, then the $R > 1$ and $R < 1$ times/places *have to* balance out.

¹The same math applies to [patrilineal descendant trees](#). The foundational work by [Galton and Watson](#) in this area considered the spread (and possible extinction) of paternally inherited aristocratic surnames in Great Britain in the 1870’s.

E. 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:

1. the conditional probability the first toss is H *given* that exactly 7 of the first 10 tosses are H.
2. the probability that the pattern THHTTT appears at least once in the sequence X_1, X_2, X_3, \dots
3. the probability that TTT appears in the sequence X_1, X_2, X_3, \dots before HH appears.
4. the conditional probability that the first two tosses are both heads *given* that exactly 5 of the first 10 tosses are heads. Is this number greater or smaller than $1/4$?
5. the probability that *every* finite-length pattern appears *infinitely many times* in the sequence X_1, X_2, X_3, \dots

F. 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.”

G. 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 \hat{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 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. 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>. Another study found that even when using the same data set and asking the same question, researchers may arrive at [very different conclusions](#) because they formalize the question and analyze the data in different ways.

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.

H. 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 10 million baseball cards in 100,000 batches of 100. Each batch is numbered from 1 to 100. Another factory has produced 10 million baseball card in 10,000 batches of 1,000, each batch numbered from 1 to 1,000. A third factory produced a 10 million baseball card in 1000 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 87.

- (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^{50} intelligent beings, grouped into civilizations of size 10^{12} each. Another universe contains 10^{50} intelligent beings, grouped into civilizations of size 10^{16} each. A final universe contains 10^{50} intelligent beings, grouped into civilizations of size 10^{20} each. You pick a random one of these 3×10^{50} 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](#) is that it is relatively likely that the number of future humans will not be too large, for the following reason: *if* advanced civilizations typically lasted for tens of millions of years (with perhaps 10^8 beings born per year—and over 10^{15} beings total), then it would seem very *coincidental* for us to find ourselves among the first [10¹¹ or so](#). 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.) By some accounts, [typical mammalian species tend to last for millions of years](#). Should one *a priori* expect the duration of a species like ours to be longer (we are uniquely resourceful) or shorter (we have the potential to be uniquely self-destructive)? In any case, the doomsday argument may at least help us consider the *possibility* of near-term human extinction, and whether preparing for apocalyptic scenarios (asteroids, plagues, nuclear war, climate disaster, supervolcanos, resource depletion, persistent low fertility, the next ice age, etc.) might improve our chance of surviving long enough for another ten billion (hundred billion? trillion? quadrillion?) more humans to enjoy this world.