

Notes for Recitation 20

1 Bayes' Rule

Bayes' Rule says that if A and B are events with nonzero probabilities, then:

$$\Pr(A | B) \cdot \Pr(B) = \Pr(B | A) \cdot \Pr(A)$$

- a. Prove Bayes' Rule.

Solution. We reason as follows:

$$\begin{aligned}\Pr(A \cap B) &= \Pr(A \cap B) \\ \frac{\Pr(A \cap B)}{\Pr(B)} \cdot \Pr(B) &= \frac{\Pr(A \cap B)}{\Pr(A)} \cdot \Pr(A) \\ \Pr(A | B) \cdot \Pr(B) &= \Pr(B | A) \cdot \Pr(A)\end{aligned}$$

In the first step, we rewrite both sides using the facts that $\Pr(A)$ and $\Pr(B)$ are nonzero. The second step uses the definition of conditional probability. ■

- b. A weatherman walks to work each day. Some days it rains:

$$\Pr(\text{rains}) = 0.30$$

Sometimes the weatherman brings his umbrella. Usually this is because he predicts rain, but he also sometimes carries it to ward off bright sunshine.

$$\Pr(\text{carries umbrella}) = 0.40$$

As a weatherman, he usually doesn't get caught out in a storm without protection:

$$\Pr(\text{carries umbrella} | \text{rains}) = 0.80$$

Suppose you see the weatherman walking to work, carrying an umbrella. What is the probability of rain? Use Bayes' Rule.

Solution.

$$\begin{aligned}\Pr(\text{rains} | \text{carries umbrella}) &= \Pr(\text{carries umbrella} | \text{rains}) \cdot \frac{\Pr(\text{rains})}{\Pr(\text{carries umbrella})} \\ &= 0.80 \cdot \frac{0.30}{0.40} \\ &= 0.60\end{aligned}$$

We've turned around cause and effect! Risk of rain has the effect of making the weatherman carry his umbrella. Yet we've shown that if he carries his umbrella, it is pretty likely to rain! ■

2 DNA Profiles

Suppose that we create a national database of DNA profiles. Let's make some (overly) simplistic assumptions:

- Each person can be classified into one of 20 billion different “DNA types”. (For example, you might be type #13,646,572,661 and the person next to you might be type #2,785,466,098.) Let $T(x)$ denote the type of person x .
 - Each DNA type is equally probable.
 - The DNA types of Americans are mutually independent.
- a. A congressman argues that there are only about 300 million Americans, so even if a profile for every American were stored in the database, the probability of even one coincidental match would be very small.

Recall from lecture that if there are N days in a year and m people in a room, then the probability that no two people in the room have the same birthday is about $e^{-m^2/(2N)}$. Using this fact, what is the probability that two people's DNA profiles would match if every person's profile were stored in the database?

Solution. The probability of a match $P(\text{match}) = 1 - P(\text{nonmatch})$. By the birthday principle, this is:

$$P(\text{match}) = 1 - e^{-\frac{(3 \cdot 10^8)^2}{2 \cdot 2 \cdot 10^{10}}} = 1 - e^{-\frac{9 \cdot 10^6}{4}} \approx 1$$

■

- b. After this database is implemented, some DNA is found at a crime scene. The DNA is sequenced and a person with matching DNA is found through the database and accused of the crime. At the trial the defense attorney argues that, by the birthday principle, the probability that there are multiple people whose DNA is identical is a virtual certainty, and so the jury cannot conclude beyond a reasonable doubt that the defendant is the criminal.

What is the flaw in this argument? Under what circumstances could you conclude based on DNA evidence alone that there is no doubt that the defendant committed the crime? (assuming no errors in the DNA tests, a comprehensive database, etc. etc.)

Solution. The birthday principle can be used to compute the probability that there are two people in the database who have the same DNA profile; this is analogous to the probability that there are two people in the room with the same birthday. It cannot be used directly to compute the probability that there is another person in the database with the same DNA profile as the criminal; this is analogous to the probability that there is a person in the room with the same birthday as a *particular person*.

So in this case, we are interested in the probability that the defendant is the criminal, given that the defendant DNA matches the crime scene profile. This could be computed if we knew how many people in the database share that profile: If there is only one person in the database with that profile, then you have your perpetrator. If there are n such people, then the probability that a person drawn at random is the perpetrator is $1/n$, which is less than $1/2$ for any $n > 1$ and thus not beyond a reasonable doubt.

We can also compute the probability that the defendant is guilty even if we didn't know the number of the people in the database who shared that profile by summing over all the possibilities. We'll look at how to describe this in a later lecture.

Note: Bear in mind that this is just a 6.042 problem and not a reasonable model for the way that DNA evidence is actually used in forensics; genomes are not independent, just for starters. ■

3 The Immortals

There were n Immortal Warriors born into our world, but in the end *there can be only one*. The Immortals' original plan was to stalk the world for centuries, dueling one another with ancient swords in dramatic landscapes until only one survivor remained. However, after a thought-provoking discussion of probabilistic independence, they opt to give the following protocol a try:

1. The Immortals forge a coin that comes up heads with probability p .
2. Each Immortal flips the coin once.
3. If *exactly one* Immortal flips heads, then he or she is declared The One. Otherwise, the protocol is declared a failure, and they all go back to hacking each other up with swords.
 - a. One of the Immortals (the Kurgan from the Russian steppe) argues that as n grows large, the probability that this protocol succeeds must tend to zero. Another (McLeod from the Scottish highlands) argues that this need not be the case, provided p is chosen *very carefully*. What does your intuition tell you?

Solution. Your intuition tells you that it is not to be trusted and that there *can be only one* way to solve this problem: do the math. ■

- b. What is the probability that the experiment succeeds as a function of p and n ?

Solution. The sample space consists of all possible results of n coin flips, which we can represent by the set $\{H, T\}^n$. Let E be the event that the experiment successfully

selects The One. Then E consists of the n outcomes which contain a single head. In general, the probability of an outcome with h heads and $n - h$ tails is:

$$p^h(1 - p)^{n-h}$$

Summing the probabilities of the n outcomes in E gives the probability that the procedure succeeds:

$$\Pr(E) = np(1 - p)^{n-1}$$

■

- c. How should p , the bias of the coin, be chosen in order to maximize the probability that the experiment succeeds? (You're going to have to compute a derivative!)

Solution. We compute the derivative of the success probability:

$$\frac{d}{dp} np(1 - p)^{n-1} = n(1 - p)^{n-1} - np(n - 1)(1 - p)^{n-2}$$

Now we set the right side equal to zero to find the best probability p :

$$\begin{aligned} n(1 - p)^{n-1} &= np(n - 1)(1 - p)^{n-2} \\ (1 - p) &= p(n - 1) \\ p &= 1/n \end{aligned}$$

This answer makes sense, since we want the coin to come up heads exactly 1 time in n . ■

- d. What is the probability of success if p is chosen in this way? What quantity does this approach when n , the number of Immortal Warriors, grows large?

Solution. Setting $p = 1/n$ in the formula for the probability that the experiment succeeds gives:

$$\Pr(E) = \left(1 - \frac{1}{n}\right)^{n-1}$$

In the limit, this tends to $1/e$. McLeod is right. ■