Notes for Tutorial 3

6.042 - February 21 and 24, 2003

The goal this week is not to teach new mathematical techniques, but rather to see some implications of the methods we've learned in daily life and computer science. There are four topics:

- Birthday paradox
- The witness.
- Disease testing.
- Digital fingerprinting.
- Forging digital signatures.

1 Birthday Paradox

This is overflow from Thursday's lecture.

Suppose that there are 23 people in a room. What is the probability that two of them share the same birthday? One might expect the probability to be rather low; after all, there are 365 different days in a year to choose from. Suprisingly, however, the probability that some two of the people share a birthday is over 1/2.

Let's generalize the problem by making the specific numbers into variables. In particular, let m be the number of people in the room, and let N be the number of days in a year. Furthermore, let's make some simplifying assumptions:

- No one was born on February 29 (leap day).
- For each person, all N birthdays are equally likely.
- The birthdates of the people are mutually independent; that is, the probability that one person was born on a given day is $\frac{1}{N}$ regardless of the dates on which the other people were born.

Number the people from 1 to m. Let D_i be the event that the first i people have distinct birthdays. Then $D_1 = 1$ trivially. In general, the total probability law says:

$$\Pr\{D_i\} = \Pr\{D_i \mid D_{i-1}\} \cdot \Pr\{D_{i-1}\} + \Pr\{D_i \mid \overline{D_{i-1}}\} \cdot \Pr\{\overline{D_{i-1}}\}$$

There is no way for the first i people to have different birthdays if the first $i-1$ people do not. Therefore, $\Pr\{D_i \mid \overline{D_{i-1}}\} = 0$, and so we have a simpler equation:

$$\Pr\{D_i\} \quad = \quad \Pr\{D_i \mid D_{i-1}\} \cdot \Pr\{D_{i-1}\}$$

Now let's tackle the term $\Pr\{D_i \mid D_{i-1}\}$. This is the probability that the first i people have different birthdays, given that the first $i-1$ people have different birthdays. This event happens unless the birthday of the i-th person happens to fall on one of the $i-1$ birthdays already taken. Therefore, the probability of this event is $1 - \frac{i-i}{N}$, according to our assumptions. Substituting this into the equation above, we find:

$$\Pr\{D_i\} \quad = \quad \left(1 - \frac{i-i}{N}\right) \cdot \Pr\{D_{i-1}\}$$

We can use this equation to compute the probability that m people have different birthdays:

$$\begin{aligned}
\Pr\{D_m\} &= \left(1 - \frac{m-1}{N}\right) \cdot \Pr\{D_{m-1}\} \\
&= \left(1 - \frac{m-1}{N}\right) \cdot \left(1 - \frac{m-2}{N}\right) \Pr\{D_{m-2}\} \\
\cdots \\
&= \left(1 - \frac{m-1}{N}\right) \cdot \left(1 - \frac{m-2}{N}\right) \cdots \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{0}{N}\right) \\
&= \prod_{i=0}^{m-1} \left(1 - \frac{i}{N}\right)
\end{aligned}$$

Now we can answer our original question by plugging numbers into the formula above. Assuming a year with 365 days, the probability that 22 people all have different birthdays is 0.524. The probability that 23 people all have different birthdays is 0.493; in other words, it is more likely than not that some two people have the same birthday!

The product used above is difficult to work with. We can get an approximate answer that is easier to use.

Fact 1 *For every real number x, the following inequality holds:*

$$1 + x \quad \leq \quad e^x$$

When x is small, this inequality is nearly tight; that is, the left side is very nearly equal to the right side. Combining this fact with our solution above, we find:

$$\begin{aligned}
\Pr\{D_m\} &= \prod_{i=0}^{m-1}\left(1 - \frac{i}{N}\right) \\
&\leq \prod_{i=0}^{m-1} e^{-i/N} \\
&= e^{-(0/N+1/N+2/N+\ldots+(m-1)/N)} \\
&= e^{-\frac{m(m-1)}{2N}}
\end{aligned}$$

In the last step uses the summation formula:

$$1 + 2 + \ldots + (m-1) = \frac{m(m-1)}{2}$$

If $m = \sqrt{N+1}$, then $\Pr\{D_m\}$ is less than $e^{-1/2}$, which is about 61%. We can restate this conclusion as a rule of thumb:

Fact 2 *If years have N days, then you need about $\sqrt{N}$ people for it to be likely that some two people share a birthday.*

This rule has many applications, as we shall see.

2 The Witness

This is a made-up story about crime and race, but one not far from reality.

There is a robbery. The thief gets away, but there is a witness who says that the criminal was black. Subsequent tests suggest that this is a very good witness; in recreations of the event, he correctly identifies the race of the simulated thief 95% of the time. What is the probability that the actual thief was black?

[There is not enough information to answer this question, but I suggest you ask it anyway. The temptation is to answer 95%. But if they reject that answer, you might ask why.]

We can shed more light on this question by casting it in terms of probability. Let B be the event that the thief was black. Let "B" be the event that the witness says the thief was black.

Since the witness was 95% accurate in tests, we have:

$$\Pr\{\text{“}B\text{”} \mid B\} \quad = \quad 0.95$$

This is *not* the answer to the question asked! What we asked for was $\Pr\{B \mid \text{“}B\text{”}\}$! [Perhaps give an example to remind them that $\Pr\{A \mid B\}$ and $\Pr\{B \mid A\}$ can be very different!]

Here is some additional information. The thief is someone selected uniformly at random from the population of the city, which is 95% white and 5% black. Now what is the probability that the thief was black?

[Work out a tree diagram. On the first level, split on race of the thief. On the second, split on what the witness says. Label edges B, W”, “B”, and “W”.]

We get:

$$\begin{aligned}
\Pr\{B \mid \text{“}B\text{”}\} &= \frac{\Pr\{B \cap \text{“}B\text{”}\}}{\Pr\{\text{“}B\text{”}\}} \\
&= \frac{1}{2}
\end{aligned}$$

The race of the thief is no more determined that the toss of a coin!

Let's consider the reverse situation. If the witness says that the thief is white, what is the probability that the thief is actually white?

$$\begin{aligned}
\Pr\{W \mid \text{“}W\text{”}\} &= \frac{\Pr\{W \cap \text{“}W\text{”}\}}{\Pr\{\text{“}W\text{”}\}} \\
&= \frac{361}{362}
\end{aligned}$$

If the witness says that the criminal was white, there's a greater than 99% chance that he actually is! What is going on?

[Let students work out some intuition for this strange phenomenon.]

Let's make the effect even more dramatic. Suppose that our witness is only 90% accurate. After all, he got only a quick glimpse at twilight, he was terrified, etc. [Perhaps mention the search for a white van in the sniper case, which turned out to be a sky-blue Chevy Caprice.] Now what is the probability that the thief was black, given that the witness says he was?

[You might want to create a new tree diagram, since it is small and students can just smudge out the old numbers in their notes as you can at the board.]

$$\begin{aligned}\Pr\{B \mid \text{“}B\text{”}\} &= \frac{\Pr\{B \cap \text{“}B\text{”}\}}{\Pr\{\text{“}B\text{”}\}} \\ &= \frac{9}{28}\end{aligned}$$

There are better than 2-to-1 odds that the thief was not black, even though the witness says he was!

3 Disease Testing

There is a growing capability to diagnose rare diseases in children with high confidence long before they are even born. Potential parents who use such diagnostic technology may be confronted with terrible choices about what to do next. What do you do if a 99% accurate test says that your child will suffer from a horrible disease? Probability doesn't answer the hard moral questions, but it leads to some startling conclusions about the limitations of such testing.

Suppose that each person carries a certain, rare disease with probability 1 in 1000. Two doctors claim to be able to test a child for the disease while it is still in the womb.

Doctor A was educated at Harvard Medical School and employs the latest high-tech genetic test. If a fetus has the disease, then the test comes back positive 99% of the time. If the patient does not have the disease, then the test comes back negative 97% of the time. What is the probability that Doctor A correctly diagnoses a patient?

$$\begin{aligned}\Pr(A \text{ is correct}) &= \Pr(\text{is diseased} \cap \text{test positive}) \\ &\quad + \Pr(\text{not diseased} \cap \text{test negative}) \\ &= \Pr(\text{is diseased}) \cdot \Pr(\text{test positive} \mid \text{is diseased}) \\ &\quad + \Pr(\text{not diseased}) \cdot \Pr(\text{test negative} \mid \text{not diseased}) \\ &= 0.001 \cdot 0.99 + 0.999 \cdot 0.97 \\ &\approx 0.970\end{aligned}$$

Doctor B murmurs strange chants and burns incense. With probability 1 in 1000, Doctor B says that a fetus has the disease, regardless of whether or not the patient actually does. What is the probability that Doctor B correctly disgnoses a patient?

$$
\begin{aligned}
\Pr(B \text{ is correct}) &= \Pr(\text{is diseased} \cap \text{test positive}) \\
&\quad + \Pr(\text{not diseased} \cap \text{test negative}) \\
&= \Pr(\text{is diseased}) \cdot \Pr(\text{test positive} \mid \text{is diseased}) \\
&\quad + \Pr(\text{not diseased}) \cdot \Pr(\text{test negative} \mid \text{not diseased}) \\
&= 0.001 \cdot 0.001 + 0.999 \cdot 0.999 \\
&\approx 0.998
\end{aligned}
$$

Which doctor makes more errors? And how often, compared to the other? Doctor A makes about 15 times more errors than Doctor B.

What are the implications of this calculation?

4 Digital Fingerprinting

The following problem is based on a lawsuit against Akamai, a recent MIT startup company.

[I'll try to gather colorful details on the case and add them to the notes later today. The math is fine, though.]

First, we have to understand the notion of a "digital fingerprinting function". This is a function that maps a block of data of any size— a text file, an image, whatever— down to a short "digital fingerprint" consisting of just n bits.

$$F : \text{file} \mapsto \{0, 1\}^n$$

Maybe n is 50 or 100 or 500 or something. A fingerprinting function has the additional property that it is very difficult to find two messages x and y such that $F(x) = F(y)$. In other words, a digital fingerprint acts a lot like a real fingerprint. If two files have the same digital fingerprint, they are almost surely the same file.

Sometimes Akamai wanted to check whether an image on a computer in Mongolia was the *right* image. To do this, they couldn't compare a picture A at their headquarters in Cambridge pixel-by-pixel against the image M in Mongolia– that would take too much bandwidth. So they would get a *fingerprint* for the picture in Mongolia and a fingerprint for the image in Cambridge and compare those:

$$F(M) \quad = \quad F(A)?$$

If the fingerprints are different, then the files are certainly different. If the fingerprints are the same, then the files are very likely to be the same (since it is difficult to find x and y such that $F(x) = F(y)$.)

In fact, Akamai would maintain a fingerprint with every file so they could easily keep track of all the stuff they have.

Akamai was using a fingerprinting function called MD5. Now MD5 produces a 128-bit fingerprint. But Akamai was actually using only 56 of those bits. Since a cryptographic hash function's behavior is essentially random there is only a $1/2^{56}$ chance that two different files get the same fingerprint.

Unfortunately, a rival company called Digital Island held a patent on the use of fingerprinting functions for this purpose.

In order to win the lawsuit, the Akamai people had a weird strategy. They had to prove that their function was really *not* very good; that is, it was so bad that it couldn't be considered a fingerprinting function because it was rather *easy* to find files x and y such that $F(x) = F(y)$. So their problem was to come up with pairs of different images that had exactly the same fingerprint.

Without knowing the details of MD5, how would you go about this?

[The birthday paradox shows that you only need about 2^{28} images before you find a match. This is only about 256 million, which is within computational range. Of course, Akamai didn't have to find 256 million different images. They could just take two and twiddle 28 low-order bits in each in all possible ways.]

[Storage is a consideration in this problem. Even if you could consider 2^{56} different images, you probably wouldn't have enough disk space to store all their tags, so you wouldn't know if you'd gotten the same tag twice.]

[This probem may provide a good opportunity to talk about the range of computability. E.g., how many operations can a computer do in a year? How about all the computers in the world working together? Can you do 2^{64} calculations? Can you do 2^{128}?]

5 Forging Digital Signatures

I'll flesh out details here later. Here is the main idea.

Someone proposed the following system for digital signatures. Given a document, compute a fingerprint. Then digitally sign the fingerprint. That way, if anyone modifies the document, the fingerprint will be wrong. If anyone modifies the fingerprint, the signature will be wrong. This simplifies and accelerates the signing process, since you need only sign a short fingerprint rather than an entire message.

This problem is vulnerable to a birthday attack.

Suppose that the fingerprinting function generates an n-bit fingerprint.

Suppose Mary wants John to sign a bad contract.

If she generates two contracts– one good and one bad– with *identical* fingerprints, she can give Jon the good one to sign, then swap in the bad one. There is no way to detect the forgery!

To do this, she makes up $2^{n/2}$ bad contracts with different fingerprints and $2^{n/2}$ good contracts with different fingerprints. They differ in tiny ways... placement of commas, British vs. American spellings, ordering of adjectives, etc.

(Note that she getting this many contracts with different fingerprints is easy, despite the birthday paradox. With overwhelming probability, each new contract she generates has a unique fingerprint. If one doesn't, she discards it.)