Chris Piech
CS109

Section #2
Jan 23rd, 2025

Section 2: Core Probability and Random Variables

1 Conditional Probabilities: Missing Not at Random

Preamble: We have three big tools for manipulating conditional probabilities:

- Chain Rule: $P(EF) = P(E|F)P(F)$
- Law of Total Probability: $P(E) = P(EF) + P(EF^C) = P(E|F)P(F) + P(E|F^C)P(F^C)$
- Bayes Rule: $P(E|F) = \frac{P(F|E)P(E)}{P(F)} = \frac{P(F|E)P(E)}{P(F|E)P(E)+P(F|E^C)P(E^C)}$

We're going to practice inferring which formula is best for solving a problem.

Problem: You recently tried out a new collaborative note-taking tool in class and want to know if students like it. You email all 100 people in class, asking them to reply saying if they liked it or not.

User Response	**Count**
Responded that they liked your tool	40
Responded that they didn't like your tool	45
Did not respond	15

Let L be the event that a person liked your tool. Let R be the event that a person responded. We are interested in estimating $P(L)$; however, that is hard, given that 15 people did not respond.

a. What is the probability that a user liked your tool and that they responded to the email $P(L \text{ and } R)$?

   Out of 100 people surveyed, 40 of them fall into the category of people who replied that they liked the tool, so $P(L \text{ and } R) = \frac{40}{100}$. This comes from the formula $P(E) = \frac{|E|}{|S|}$.
   A common pitfall is to say $\frac{40}{85}$ instead, but that would imply shrinking our sample space to only the people who replied, which means $\frac{40}{85} = P(L|R)$, not $P(L \text{ and } R)$.

b. Which formula would you use to calculate $P(L)$? Consider that people who like your tool are in one of two (mutually exclusive) groups: those that replied, and those that did not.

The law of total probability. It breaks down $P(L)$ into two parts, the part which intersects with R and the part that intersects with R^C.

$$P(L) = P(L \text{ and } R) + P(L \text{ and } R^C)$$

c. You estimate that the probability that someone did not respond, given that they liked the tool, is $P(R^C|L) = \frac{1}{5}$. Calculate $P(L)$.

$$\begin{aligned}
P(L) &= P(L \text{ and } R) + P(L \text{ and } R^C) && \text{Law of Total Probability} \\
&= \frac{40}{100} + P(L \text{ and } R^C) && \text{From part a} \\
&= \frac{40}{100} + P(R^C|L)P(L) && \text{Chain rule} \\
P(L) - P(R^C|L)P(L) &= \frac{40}{100} && \text{Solving for } P(L) \\
P(L) \cdot [1 - P(R^C|L)] &= \frac{40}{100} \\
P(L) \cdot \frac{4}{5} &= \frac{40}{100} \\
P(L) &= \frac{40}{100} \cdot \frac{5}{4} \\
P(L) &= \frac{1}{2}
\end{aligned}$$

2 Independent Infants

Each child in a daycare has a 0.2 probability of having disease A, and has an independent 0.4 probability of having disease B. A child is sick if they have either disease A or disease B.

a. What is the probability that a child is sick?

Let A and B be the events that a child has disease A and disease B, respectively. A child is healthy if they have neither disease A nor disease B. So,

$$\begin{aligned}
P(\text{sick}) &= 1 - P(\text{healthy}) \\
&= 1 - P(A^C, B^C) \\
&= 1 - P(A^C)P(B^C) && \text{Definition of Independence} \\
&= 1 - (1 - P(A))(1 - P(B)) \\
&= 1 - (0.8)(0.6) \\
&= 1 - (0.48) \\
&= 0.52
\end{aligned}$$

Alternate solution: A child is considered sick if they have disease A, disease B, or both. Thus, we can use Inclusion-Exclusion:

$$\begin{aligned}
P(\text{sick}) &= P(A) + P(B) - P(AB) \\
&= P(A) + P(B) - (P(A) \cdot P(B)) && \text{Definition of Independence} \\
&= (0.2) + (0.4) - (0.4 \cdot 0.2) \\
&= (0.2) + (0.4) - (0.08) \\
&= 0.52
\end{aligned}$$

Note that a sick child can have both diseases because disease A and disease B are independent events, which by definition means thy not mutually exclusive. Independent events cannot be mutually exclusive, and mutually exclusive events cannot be independent. For example, when events A and B are independent, knowing that event A has happened does not change our belief in B (i.e. $P(B|A) = P(B)$). On the other hand, when events A and B are mutually exclusive, knowing that event A happened does change our belief in B, because we know B didn't happen.

b. If there are 10 children in a daycare, what is the probability that 3 or more are sick?

Let Y be the number of children that are sick. We can write this as $Y \sim \text{Bin}(10, P(\text{sick}))$.

Thus, we have

$$
\begin{aligned}
P(Y \geq 3) &= 1 - P(Y < 3) \\
&= 1 - \sum_{k=0}^{2} \binom{10}{k} (0.52)^k (1 - 0.52)^{10-k}.
\end{aligned}
$$

3 Betcoin

Preamble: When a random variable fits neatly into a family we've seen before (e.g. Binomial), we get its expectation for free. When it does not, we have to use the definition of expectation.

Problem: Your friend tells you about the hottest new investment craze: Betcoin! The cryptocurrency is so popular that its value doubles every day. That means that if you invested \$100 today, it would be worth \$200 tomorrow, \$400 the day after tomorrow, etc.

However, market research says that Betcoin's value is likely to crash any day now. Each day, it has a 40% chance of crashing – and if it crashes, you'll lose all of the money you invest!

a. Imagine you decide to invest \$100 today. What is the expected value of your investment tomorrow? What does it mean for this number to be greater or less than \$100?

Let X_1 be the value of our \$100 Betcoin investment after 1 day. X_1 has two possible values: 0 (if Betcoin crashes) and 200 (if it doesn't crash). $P(X_1 = 0) = 0.4$ and $P(X_1 = 200) = 0.6$. Using the definition of expectation:

$$
\begin{aligned}
E[X_1] &= 0 \cdot P(X_1 = 0) + 200 \cdot P(X_1 = 200) \\
&= 200 \cdot 0.6 = 120
\end{aligned}
$$

Since the expectation is greater than 100, we are more likely to earn money than lose money if we invest for a single day only.

b. (Extra Practice) You devise a scheme: each day that Betcoin doesn't crash and your \$100 investment doubles in value, you sell the \$100 surplus, leaving \$100 still invested (to potentially be doubled again the next day). You repeat this process daily, selling \$100 and leaving \$100 still invested, until Betcoin crashes and you lose the invested \$100. Generally, if Betcoin crashes on day i, you'd earn \$100 for $i - 1$ days, then lose \$100 the last day, giving you a net profit of $100(i - 2)$. What is your expected profit with this scheme?

Let D be what day Betcoin crashes. D could have any value from 1 to infinity. $P(D = 1) = 0.4$ from the problem statement; then for each additional day that Betcoin doesn't crash, we multiply by the complement, 0.6. In other words, $D \sim \text{Geo}(p = 0.4)$. So $P(D = i) = 0.4 \cdot 0.6^{i-1}$ (PMF of the Geometric).

Let X be your profit from this scheme. If Betcoin crashes on day 1, you lose all your money; so when $D = 1$, $X = -100$. If $D = 2$, you make \$100 from selling on day 1, but then lose your initial investment ($X = 0$). If $D = 3$, $X = 100$; if $D = 4$, $X = 200$; and so on. Technically, this could go on forever, though the probability of Betcoin still not crashing gets smaller and smaller.
Putting all this together in the formula for expectation:

$$\begin{aligned} E[X] &= \sum_{\text{all } x} x \cdot P(X = x) \\ &= \sum_{i=1}^{\infty} 100(i-2) \cdot P(D = i) \\ &= \sum_{i=1}^{\infty} 100(i-2) \cdot 0.4 \cdot 0.6^{i-1} \end{aligned}$$

On an exam, you'd get almost all the points if you stopped here. But we actually can break down this sum to get a final number, and it will help you to learn to recognize when we can do this. Let's factor out the 100 from the sum, then split into two sums by distributing the $i - 2$:

$$\begin{aligned} \sum_{i=1}^{\infty} 100(i-2) \cdot 0.4 \cdot 0.6^{i-1} &= 100 \sum_{i=1}^{\infty} i \cdot 0.4 \cdot 0.6^{i-1} - 200 \sum_{i=1}^{\infty} 0.4 \cdot 0.6^{i-1} \\ &= 100 \cdot 2.5 - 200 \cdot 1 \\ E[X] &= 50 \end{aligned}$$

Notice that the left sum matches the expectation of a geometric random variable! Knowing $D \sim \text{Geo}(p = 0.4)$, it's $E[D] = 2.5$.

The right sum equals 1 because it is the sum of $P(D = i)$ for all possible i (for any random variable, the sum of the PMF over all possible values is 1.)

So this strategy would result in a small profit, half your original investment, if the assumptions of the problem hold true (which they usually don't in the real world): that Betcoin exactly doubles in value each day (which is extreme), that the probability of a crash is the same daily, and that each day is independent of all others.