Chris Gregg
CS109

Section #2
April 15th, 2026

Section 2: Core Probability and Random Variables

1 Prickly Dictators and Plausible Deniability

You are trying to estimate the fraction of people in a country who support an authoritarian regime. Directly asking "do you support the dictator?" is unsafe and respondents will not respond truthfully. Luckily, you've taken CS109, so, you use the following survey instead:

1. Is your birthday in Jan–Jun?
2. Is the last digit of your ID number even?
3. Do you enjoy surprises?
4. Would you rather have a salad over a soup?
5. Does your first name have an even number of letters?
6. Do you support the dictator?

Instead of collecting answers to each question, you only collect **the total number of "yes" answers** from each person. Because question 6 is sensitive, reporting only the total number of "yes" answers gives respondents plausible deniability. Even then, using our probability muscles, we can still estimate how often people answered "yes" to question 6. Assume the following:

- For each question 1–5, a person answers "yes" with probability 0.5.
- The answers to questions 1–5 are independent of each other. The answer to question 6 is independent of the answers to questions 1–5.
- Let p be the probability that a person answers "yes" to question 6 (supporting the dictator).

a. What is the probability that a randomly chosen person answers "yes" to exactly 4 questions?

Let X be the number of "yes" answers to questions 1–5. Since each of these is "yes" with probability 0.5 and the questions are independent,

$$X \sim \text{Bin}(5, 0.5).$$

Let Z be the indicator random variable for answering "yes" to question 6, so

$$Z \sim \text{Bern}(p).$$

The total number of "yes" answers is

$$Y = X + Z.$$

We want $P(Y = 4)$. This can happen in exactly two ways:

- $X = 4$ and $Z = 0$

- $X = 3$ and $Z = 1$

Using independence of X and Z,

$$\begin{aligned} P(Y = 4) &= P(X = 4, Z = 0) + P(X = 3, Z = 1) \\ &= P(X = 4)P(Z = 0) + P(X = 3)P(Z = 1) \\ &= P(X = 4)(1 - p) + P(X = 3)p. \end{aligned}$$

Now compute the binomial probabilities:

$$P(X = 4) = \binom{5}{4}(0.5)^4(0.5)^1 = \binom{5}{4}(0.5)^5 = \frac{5}{32},$$

$$P(X = 3) = \binom{5}{3}(0.5)^3(0.5)^2 = \binom{5}{3}(0.5)^5 = \frac{10}{32}.$$

Plugging in:

$$\begin{aligned} P(Y = 4) &= \frac{5}{32}(1 - p) + \frac{10}{32}p \\ &= \frac{5}{32} + \frac{5}{32}p \\ &= \frac{5}{32}(1 + p). \end{aligned}$$

b. Suppose you survey 10,000 people, and let 2,500 be the number of respondents who report a total of 4 “yes” answers. Give an estimate for p, the probability that a person answers ”yes” to supporting the dictator.

From part (a), the probability of reporting 4 “yes” answers is

$$P(Y = 4) = \frac{5}{32}(1 + p).$$

With many surveys, the empirical fraction should be close to this probability, so we set

$$P(Y = 4) \approx \frac{2{,}500}{10{,}000}$$

Solving for p:

$$\begin{aligned} 0.25 &= \frac{5}{32}(1+p) \\ 0.25 \cdot \frac{32}{5} &= 1+p \\ \frac{32}{5} \cdot \frac{1}{4} &= 1+p \\ \frac{8}{5} &= 1+p \\ p &= \frac{8}{5} - 1 \\ p &= \frac{3}{5} \\ p &= 0.6. \end{aligned}$$

So, based on this sample, we estimate $\widehat{p} = 0.6$.

c. (Optional Challenge) Instead of focusing only on the event that a randomly chosen person answers "yes" to exactly 4 of the 6 questions, we can estimate p using the **average** number of "yes" answers.

Let Y be the total number of "yes" answers out of 6. Find $\mathbb{E}[Y]$ in terms of p, and give an estimate for p.

Recall that $Y = X + Z$, where:

- X is the number of "yes" answers to questions 1–5, and $X \sim \text{Bin}(5, 0.5)$
- Z is the indicator for answering "yes" to question 6, so $Z \sim \text{Bern}(p)$

By linearity of expectation,

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[X+Z] \\ &= \mathbb{E}[X] + \mathbb{E}[Z] \\ &= 5(0.5) + p \\ &= 2.5 + p. \end{aligned}$$

Now suppose we survey n people, and let $\overline{Y}$ be the average total yes-count across the sample survey data. We expect $\overline{Y} \approx \mathbb{E}[Y]$ (the sample mean is a good estimate of $\mathbb{E}[Y]$. We will cover this later, but feel free to use just $\mathbb{E}[Y]$), then we can set

$$\overline{Y} \approx 2.5 + p.$$

Solving for p gives the estimate

$$\widehat{p} = \overline{Y} - 2.5.$$

d. (Optional Challenge) Suppose the dictator learned that a particular respondent reported a total of 4 "yes" answers out of 6. Can the dictator find out whether this person answered "yes" to question 6?

While the dictator does not know for certain whether or not this person definitely answered "yes" to the question, he could reconstruct a probability that a person did with the data too.

Recall the setup: $Y = X + Z$, where $X \sim \text{Bin}(5, 0.5)$ is the number of "yes" answers to questions 1–5, and $Z \sim \text{Bern}(p)$ is the indicator for answering "yes" to question 6.
If $Y = 4$, there are exactly two possibilities:

- $Z = 1$ and $X = 3$
- $Z = 0$ and $X = 4$

Using Bayes' rule and independence of X and Z,

$$\begin{aligned} P(Z = 1 \mid Y = 4) &= \frac{P(Y = 4 \mid Z = 1)P(Z = 1)}{P(Y = 4)} \\ &= \frac{P(X = 3)\, p}{P(X = 3)\, p + P(X = 4)\,(1 - p)}. \end{aligned}$$

From $X \sim \text{Bin}(5, 0.5)$,

$$P(X = 3) = \binom{5}{3}(0.5)^5 = \frac{10}{32}, \qquad P(X = 4) = \binom{5}{4}(0.5)^5 = \frac{5}{32}.$$

Plugging in,

$$\begin{aligned} P(Z = 1 \mid Y = 4) &= \frac{\left(\frac{10}{32}\right) p}{\left(\frac{10}{32}\right) p + \left(\frac{5}{32}\right)(1 - p)} \\ &= \frac{10p}{10p + 5(1 - p)} \\ &= \frac{10p}{5 + 5p} \\ &= \frac{2p}{1 + p}. \end{aligned}$$

So, for $p = 0.6$, the probability that the particular respondent likes the dictator is 75%, which may or may not be enough for the prickly dictator to make up his mind about someone. The power of probability can go both ways... Whoopsie.

2 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. (Challenge) 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?

Solution 1: Let X be your profit from this scheme. If Betcoin crashes on day 1, you lose all your money so $X = -100$. Otherwise, you make \$100 from selling on day 1, and play the game again. When you play the game again, you expect to make an additional profit of $E[X]$. Thus, $X = 100 + E[X]$.
Putting all this together in the formula for expectation:

$$\begin{aligned} E[X] &= -100 \cdot P(\text{crash}) + (100 + E[X]) \cdot P(\text{no crash}) \\ E[X] &= -100 \cdot 0.4 + (100 + E[X]) \cdot 0.6 \\ E[X] &= 20 + 0.6E[X] \\ E[X] &= 50. \end{aligned}$$

Solution 2: 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.

3 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

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

4 Review (Optional): 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, and we already calculated one of the necessary terms on the right-hand side in part a.

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

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 $P(L)$.

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