Discussion 1 - Random Variables and Conditional Probability

Albert Xue

January 2023

1 Random Variables

Here's a question: What is the value of a (fair) coin toss before it happens?

Well, the toss hasn't gone up yet, so we'd be lying if we said either heads or tails. It would be incorrect to assign the coin toss a value before it happens, because we are uncertain about the outcome.

But we aren't clueless either, because we still have a pretty good description of the process. First off, we know that the result will either be heads or tails, because that's how coin tosses work. Coins don't come up "dolphin"; they either land face-up or face-down. Second, we know that this coin is fair, because I told you so, so heads and tails have equal probability of coming up.[1]

We can capture both this uncertainty and this *a priori* knowledge through the concept of a **random variable**. Random variables are core to statistics; unfortunately, they are (arguably) confusingly named, as they are not "variables" in the classic sense.

Definition 1. *A **random variable** assigns a real value to each outcome of a random circumstance.*

What does this mean? It means that random variables must be distinguished from probabilities and from events, because they are neither.[2] They're actually functions! Random variables are a mapping from events to probabilities; feed in an event, and they'll tell you with what probability that event will occur. More formally, random variables are functions $X : \Omega \to \mathbb{R}$ that map an event space Ω to the real numbers.[3]

Let's return to our coin toss. We assign X to the outcome of the coin toss *before* tossing. We know that in the end, the coin toss will either be heads or tails; this defines our event space, which we call Ω, and more explicitly means that $\Omega \coloneqq \{\text{heads}, \text{tails}\}$. We also know that the probability for each is equal, so in notation, $\mathbb{P}(X = \text{heads}) = \frac{1}{2}$ and $\mathbb{P}(X = \text{tails}) = \frac{1}{2}$.

Then X is a random variable with the following definition, or **distribution**:

$$X = \begin{cases} \text{heads}, & \text{with probability } \frac{1}{2} \\ \text{tails}, & \text{with probability } \frac{1}{2} \end{cases} \tag{1}$$

[1] These are the **assumptions**, the premises that simplify our system so that we can start thinking about it. We could complicate our system arbitrarily. It might be a muddy day, and coins might get stuck in the mud side-up; maybe the wind is blowing so that our coin is no longer fair. Maybe I was lying about the coin being fair. We could do inference on all of these, in theory, but they get complicated very quickly. In general, these are not helpful lines of thoughts, and my head hurts. So for today coins are fair, and they can only land heads or tails.

[2] Although they incorporate both.

[3] The notation $X : \Omega \to \mathbb{R}$ means that X is a function that takes input from one space, denoted Ω, and gives output in the real numbers, denoted $\mathbb{R}$.

X is neither heads nor tails, but a function that tells you two things that the probability of X being heads is $\frac{1}{2}$, same as tails. This is all of our systemic knowledge of coin tosses captured in one variable, and is uniquely powerful for reasoning about coin tosses.

Expanding beyond coin tosses, random variables can be **discrete**, meaning that Ω is finite or countable (like our coin toss example). They can also be **continuous**, which means that Ω is an uncountable set like the real numbers $\mathbb{R}$, but each element in Ω is still assigned a probability through some function (like a Normal distribution).

Random variables also follow **distributions**, and some of these distributions have standard named forms. For example, the distribution of X in Equation 1 is a Bernoulli distribution parameterized by $p = \frac{1}{2}$; equivalently, we might say that X follows a Bernoulli distribution with $p = \frac{1}{2}$. There's nothing insightful about this comment; calling something a Bernoulli distribution with parameter p is simply a shorthand for the following distribution:

$$f(X, p) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}. \tag{2}$$

1.1 Expectation of a Random Variable

When we perform random actions, we expect things. When I walk outside in Los Angeles, I expect it to be sunny. That's because it's sunny in Los Angeles *on average*, although this is clearly not always true.

We formalize this intuition of "on average" through the statistical concept of the **expectation** of a random variable. This can also be called the **mean**, or the **expected value**, of a distribution/random variable, but they all encode the same idea.

Definition 2. *The **expected value** of a random variable X, denoted $\mathbb{E}(X)$, is defined as*

$$\mathbb{E}(X) = \int_{x \in \Omega} x \mathbb{P}(X = x) \tag{3}$$

If X is a discrete random variable, this can also be written as a sum:

$$\mathbb{E}(X) = \sum_{x \in \Omega} x \mathbb{P}(X = x) \tag{4}$$

Note here the difference between the lowercase x and the uppercase X. The lowercase x is a value in the event space Ω; the uppercase X is a random variable, which maps events in Ω (in other words, instantiations of X, or lowercase x's) to their probabilities.

The expectation $\mathbb{E}$ is a linear function, which means that it enjoys the niceties of all linear functions, notably that

$$\mathbb{E}(cX) = c\mathbb{E}(X), \ c \in \mathbb{R}, \tag{5}$$

and

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y), \tag{6}$$

no matter if X and Y are independent or dependent. Further, if X and Y **are** independent,[4] then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y). \tag{7}$$

Please note that Eq. 7 is only valid when X and Y are independent. In general, while expectation has many other nice properties, you should not assume that any operation is valid under expectation. For example, $\mathbb{E}(\frac{1}{X}) \neq \frac{1}{\mathbb{E}(X)}$.

[4] We will explain independence later on.

2 Conditional Probability

Conditional probability is a more intuitive subject, because it's essentially a line of logic we all use in our daily lives. We begin with certain beliefs about the probabilities of given events; given new information, we adjust these beliefs accordingly.

For example, let's say I'm meeting my friend X for lunch. I hate waiting, so I want to know what the probability is that they'll be late. There's always some base probability that they will be late! Maybe there's a ton of traffic that day, or a meeting ran over, or they got bitten by a dog. We can denote this with $\mathbb{P}(X \text{ is late})$.

But also, a lot of my friends are straight up flakes. *Given* that the friend I'm meeting is a known flake, from past experience I know that the probability of them being late is high; If X is a flake and tells me 12 o clock, I'll probably leave at 1.

We introduce new notation to formalize these notions. Previously, the "unconditional probability", or **marginal probability**, that X is late is denoted $\mathbb{P}(X \text{ is late})$. Here, we don't know whether or not X is a flake, and so this quantity is averaged over flake and non-flake friends. When we incorporate the new knowledge that, no, friend X cannot be trusted with daily deadlines, we write $\mathbb{P}(X \text{ is late} \mid X \text{ is a flake})$, where we read this little bar thing $\mid$ as "given". $\mathbb{P}(X \text{ is late})$ and $\mathbb{P}(X \text{ is late})$ are different quantities.

Formally, we write the conditional probability of event A occurring, given that event B has occurred, as $\mathbb{P}(A \mid B)$. It is then true that

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}. \tag{8}$$

We can think of this identity geometrically, in terms of Venn diagrams (see Figure 1). If we're looking for the probability of A *given* B, geometrically we would take the area over B (somewhat salmon-colored) and ask, what proportion of this vaguely salmon-colored area intersects with A? (in this case, the area labeled A, B). And that is exactly what Eq 8 tells us to do! We take A, B and divide it by B; $A \mid B$ is the proportion of B that also contains A.

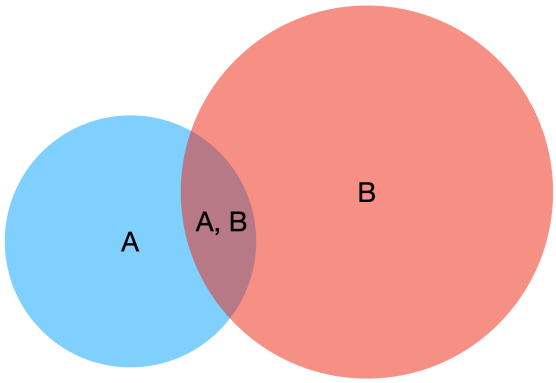

Figure 1: A Venn Diagram.

2.1 Averaging Conditional Probabilities

Theorem 1. *For a partition $B_1, \ldots, B_n$ of Ω,*

$$\mathbb{P}(A) = \mathbb{P}(A \mid B_1)P(B_1) + \cdots + \mathbb{P}(A \mid B_n)\mathbb{P}(B_n). \tag{9}$$

Let's say I give you three coins. The first coin, when tossed, always gives you heads; the second coin, when tossed, always gives you tails. The third coin is just a normal fair coin. What happens when you choose a coin at random from among the three and then toss it?

Each coin has a different conditional probability of obtaining a heads; given that you picked the first coin, for example, the conditional probability of a heads is 100%. But intuitively, we know that the marginal probability of a heads here is still $\frac{1}{2}$. Our intuition is formalized in Theorem 1, which shows us how to average out different conditional probabilities. Under Theorem 1, each coin here represents a separate partition of Ω, and so

$$\begin{aligned}
\mathbb{P}(X \text{ is heads}) &= \mathbb{P}(X \text{ is heads} \mid \text{coin 1 is chosen})\mathbb{P}(\text{coin 1 is chosen}) \\
&\quad + \mathbb{P}(X \text{ is heads} \mid \text{coin 2 is chosen})\mathbb{P}(\text{coin 2 is chosen}) \\
&\quad + \mathbb{P}(X \text{ is heads} \mid \text{coin 3 is chosen})\mathbb{P}(\text{coin 3 is chosen}) \\
&= (1)(\frac{1}{3}) + (0)(\frac{1}{3}) + (\frac{1}{2})(\frac{1}{3}) \\
&= \frac{1}{2}.
\end{aligned} \tag{10}$$

2.2 Independence

Some events just aren't related to each other; the probability of event A may not affect the probability of event B.

Here's a fact, as I sit here in cozy LS 5214: when it rains it leaks from our lab's ceiling.[5] However, no matter how much Harold complains to management, our ceiling is never fixed. Therefore, I could say that *the probability of Harold complaining to management is independent of the probability that our ceiling is fixed.*

We write independence between event A and event B as $A \perp\!\!\!\perp B$. Then we have the following identity:

$$A \perp\!\!\!\perp B \iff \mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B) \tag{11}$$

We write $\iff$ to mean *if and only if*; that means both that if A is independent of B then $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$, and also the other way around.

2.3 Bayes Theorem

Let's say that I am still waiting on my friend for lunch. I know that given that my friend X is a flake, there's some higher conditional probability that they will be Y minutes late (relative to the marginal). But let's say that I'm not sure whether or not my friend is a flake yet.[6] Given that this friend of mine is Y minutes late, can I effectively determine whether or not friend X is a flake?

[5] An empirical truth inferred from two years of working here. Also our windows don't close fully. Join Pimentel Lab!

[6] There's still hope for some of you.

It turns out that Thomas Bayes gave us Bayes Theorem to do just such an inference. Essentially, we are able to invert a conditional probability $\mathbb{P}(A \mid B)$ into a completely different distribution, $\mathbb{P}(B \mid A)$.

Theorem 2. *Bayes Theorem. For event A and event B,*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{12}$$

Bayes Theorem requires understanding of the marginal probability $\mathbb{P}(A)$, but in return allows us to invert the conditional probability $\mathbb{P}(B \mid A)$ into $\mathbb{P}(A \mid B)$. It's worth remembering here, too, that

$$\mathbb{P}(A) = \mathbb{P}(A \mid B_1)\mathbb{P}(B_1) + \cdots + \mathbb{P}(A \mid B_n)\mathbb{P}(B_n) \tag{13}$$

for some partition $B_1, \ldots, B_n$ of Ω.

3 Exercises

Problem 1 (Pitmann 3.2.3): What is the expected number of sixes appearing on three die rolls? What is the expected number of odd numbers?

Problem 2 (Pittmann 1.4.3): Suppose $\mathbb{P}(\text{rain today}) = 40\%$, $\mathbb{P}(\text{rain tomorrow}) = 30\%$, and $\mathbb{P}(\text{rain today and tomorrow}) = 30\%$. Given that it rains today, what is the chance that it will rain tomorrow?

Problem 3 (Pitmann 1.4.9): Three high schools have senior classes of size 100, 400, 500, respectively. Here are two schemes for selecting a student from among the three senior classes:

A: Make a list of all 1000 seniors, and choose a student at random from this list.

B: Pick one school at random, then pick a student at random from the senior class in that school. Show that these two schemes are not probabilistically equivalent.

Here is a third scheme:

C: Pick school i with probability $p_i (p_1 + p_2 + p_3 = 1)$, then pick a student at random from the senior class in that school.

Find the probabilities p_1, p_2, p_3 which make scheme C equivalent to scheme A.

Problem 4 (Pitmann 1.5.3): A manufacturing process produces integrated circuit chips. Over the long run the fraction of bad chips produced by the process is around 20%. Thoroughly testing a chip to determine whether it is good or bad is rather expensive, so a cheap test is tried. All good chips will pass the cheap test, but so will 10% of the bad chips.

1. Given a chip passes the cheap test, what is the probability that it is a good chip?
2. If a company using this manufacturing process sells all chips which pass the cheap test, over the long run what percentage of chips sold will be bad?

Problem 5 (Pitmann 1.5.5): The fraction of persons in a population who have a certain disease is 0.01. A diagnostic test is available to test for the disease. But for a healthy person the chance of being falsely diagnosed as having the disease is 0.05, while for someone with the disease the chance of being falsely diagnosed as healthy is 0.2. Suppose the test is performed on a person selected at random from the population.

1. What is the probability that the test shows a positive result?
2. What is the probability that the person selected at random is one who has the disease but is diagnosed healthy?
3. What is the probability that the person is correctly diagnosed and is healthy?
4. Suppose the test shows a positive result. What is the probability that the person tested actually has the disease?