Math 3332: Probability and Inference

Unit II: Discrete Random Variables

Mikhail Lavrov (mlavrov@kennesaw.edu)

Spring 2021

What we know so far

So far, when talking about random experiments, we've had:

- A **sample space** consisting of all the **outcomes** that can occur.
- Sets of outcomes called **events**, with a **probability** for each event.

Events let us talk about some true/false property of the outcome: something that either will or will not happen.

But sometimes, we want to ask about numerical properties. For example, we roll 10 dice, and want to ask questions about:

- The sum of all the dice.
- The number of ⚄'s we rolled.
- And so on.

Random variables

Informally, a random variable is some number associated with the outcome of an experiment.

For example, we might say "Let $\mathbf{X}$ be the sum of the dice" or "Let $\mathbf{Y}$ be the number of ⚄'s rolled".

We can combine these random variables to get new ones: for example $\mathbf{X} - 5\mathbf{Y}$ is a third random variable (it tracks the sum of just the dice that didn't come up ⚄).

Formally, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{X} - 5\mathbf{Y}$, and all other random variables are **functions** from the sample space to the real numbers.

When we talk about $\Pr[\mathbf{X} = 20]$, "$\mathbf{X} = 20$" is shorthand for the event consisting of all outcomes s such that $\mathbf{X}(s) = 20$.

The range of a random variable

Here is an example: we flip 3 coins, so our sample space is $S = \{\text{HHH}, \text{HHT}, \ldots, \text{TTT}\}$. Let $\mathbf{X}$ be the number of tails. This is a function $\mathbf{X} : S \to \mathbb{R}$ given by:

$$\begin{aligned}
\mathbf{X}(\text{HHH}) &= 0 \\
\mathbf{X}(\text{HHT}) &= \mathbf{X}(\text{HTH}) = \mathbf{X}(\text{THH}) = 1 \\
\mathbf{X}(\text{HTT}) &= \mathbf{X}(\text{THT}) = \mathbf{X}(\text{TTH}) = 2 \\
\mathbf{X}(\text{TTT}) &= 3
\end{aligned}$$

The **range** of a random variable is the set of values it can take on. The range of $\mathbf{X}$ is $R_{\mathbf{X}} = \{0, 1, 2, 3\}$. All our random variables are functions $S \to \mathbb{R}$, but it's also okay to think of $\mathbf{X}$ as a function $S \to R_{\mathbf{X}}$.

A random variable is **discrete** if its range is either finite, or countably infinite.

Probability mass functions

All the properties of a discrete random variable $\mathbf{X}$ that we'll need to know are summarized by knowing $\Pr[\mathbf{X} = x]$ for all $x \in R_{\mathbf{X}}$. (If $x \notin R_{\mathbf{X}}$, then $\Pr[\mathbf{X} = x] = 0$.)

The **probability mass function** of $\mathbf{X}$ is the function $P_{\mathbf{X}}$ that records this information: $P_{\mathbf{X}}(x) = \Pr[\mathbf{X} = x]$. It's okay to think of $P_{\mathbf{X}}$ as a function $R_{\mathbf{X}} \to [0, 1]$ or as a function $\mathbb{R} \to [0, 1]$ which just happens to be 0 everywhere outside $R_{\mathbf{X}}$.

In our example,

$$P_{\mathbf{X}}(x) = \begin{cases} \frac{1}{8} & x = 0 \text{ or } x = 3 \\ \frac{3}{8} & x = 1 \text{ or } x = 2 \\ 0 & \text{otherwise.} \end{cases}$$

Probability distributions

Two different random variables can have the same range and PMF. When they do, we say that they have the same **distribution**.

For example, suppose that in the same experiment, $\mathbf{Y}$ counts the number of heads (and $\mathbf{X}$, as before, counts the number of tails).

- $\mathbf{X}$ and $\mathbf{Y}$ are not the same random variable. As functions, they always disagree: for example, $\mathbf{X}(\mathsf{TTH}) = 2$ but $\mathbf{Y}(\mathsf{TTH}) = 1$.
- However, $R_{\mathbf{X}} = R_{\mathbf{Y}} = \{0, 1, 2, 3\}$ and the PMFs $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$ are the same. So $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.

Or suppose we roll 3 dice, and $\mathbf{Z}$ counts the number of odd values rolled. This is **very much** not the same random variable: it's a different experiment, even! But $\mathbf{Z}$ has the same distribution as $\mathbf{X}$ and $\mathbf{Y}$.

What's the point?

Why do we introduce random variables, probability distributions, and so on? They are a higher-level “convenience”.

- A **random variable** lets us consider many related events all at once.

  It is easier to talk about properties of $\mathbf{X}$ than to deal with the four events “there are no tails”, “there are 2 tails”, and so on.

- A **probability distribution** lets us talk about the properties of a random variable that don't depend on the “flavor text”.

  If we understand $\mathbf{X}$, then we can apply that same understanding to $\mathbf{Y}$ and $\mathbf{Z}$.

Setting up a PMF

Suppose we take a bag with 4 blue and 6 red marbles in it. We draw 3 marbles and let $\mathbf{X}$ be the number of blue marbles in our sample.

Then $R_{\mathbf{X}} = \{0, 1, 2, 3\}$ and we can figure out $P_{\mathbf{X}}$ by solving a familiar probability problem. For example,

$$P_{\mathbf{X}}(0) = \Pr[\mathbf{X} = 0] = \Pr[\{\text{●●●}\}] = \frac{\binom{4}{3}}{\binom{10}{3}} = \frac{1}{30}.$$

Once we also find

$$P_{\mathbf{X}}(1) = \frac{3}{10} \quad P_{\mathbf{X}}(2) = \frac{1}{2} \quad P_{\mathbf{X}}(3) = \frac{1}{6}$$

we are ready to answer all questions about $\mathbf{X}$.

Using a PMF

Let $\mathbf{X}$ be a random variable with $R_{\mathbf{X}} = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ and

$$P_{\mathbf{X}}(x) = \begin{cases} 0.4 & x = 0 \\ 0.3 & x = \frac{1}{3} \\ 0.2 & x = \frac{2}{3} \\ 0.1 & x = 1 \end{cases}$$

- $\Pr[\mathbf{X} \leq \frac{1}{2}] = P_{\mathbf{X}}(0) + P_{\mathbf{X}}(\frac{1}{3}) = 0.4 + 0.3 = 0.7.$
- $\Pr[\mathbf{X} = 0 \mid \mathbf{X} \leq \frac{1}{2}] = \frac{\Pr[\mathbf{X}=0 \text{ and } \mathbf{X}\leq\frac{1}{2}]}{\Pr[\mathbf{X}\leq\frac{1}{2}]} = \frac{0.4}{0.4+0.3} = \frac{4}{7}.$
- $\Pr[3\mathbf{X} \text{ is even}] = P_{\mathbf{X}}(0) + P_{\mathbf{X}}(\frac{2}{3}) = 0.4 + 0.2 = 0.6.$

From one PMF to another

Let $\mathbf{X}$ be a random variable with $R_{\mathbf{X}} = \{-2, -1, 0, 1, 2\}$ and $P_{\mathbf{X}}(k) = \frac{1}{5}$ for each $k \in R_{\mathbf{X}}$.

- Let $\mathbf{Y} = \mathbf{X} + 2$. Then $R_{\mathbf{Y}} = \{0, 1, 2, 3, 4\}$ and $P_{\mathbf{Y}}(k) = \frac{1}{5}$ for each $k \in R_{\mathbf{Y}}$.
- Let $\mathbf{Z} = \mathbf{X}^2$. Then $R_{\mathbf{Z}} = \{0, 1, 4\}$ and

$$P_{\mathbf{Z}}(k) = \begin{cases} \frac{1}{5} & k = 0 \\ \frac{2}{5} & k = 1 \\ \frac{2}{5} & k = 4 \end{cases}$$

Note that to answer these questions, we don't need to know the random experiment that was used to define $\mathbf{X}$.

A peek at joint distributions

Suppose we flip a coin 6 times. Let $\mathbf{X}$ be the number of tails in the first 3 flips, $\mathbf{Y}$ be the number of heads in the first 3 flips, $\mathbf{Z}$ be the number of tails in the last 3 flips.

All three random variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ have the same distribution. If we find $\Pr[\mathbf{X} = 1 \mid \mathbf{X} > 0]$, then this tells us $\Pr[\mathbf{Y} = 1 \mid \mathbf{Y} > 0]$.

When we ask questions about two of these at a time, things change. For example:

- $\mathbf{X} + \mathbf{Y}$ is always 3.
- $\mathbf{X} + \mathbf{Z}$ counts the number of tails in all 6 flips: it is a random variable with range $\{0, 1, 2, 3, 4, 5, 6\}$.

To answer such questions, we need to know how $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are related.

Named distributions

When a particular probability distribution shows up over and over, we give that distribution a name.

For example, consider a random experiment with repeated Bernoulli trials: n independent experiments, each of which is a “success” with probability p, and a “failure” otherwise.

We have already encountered the binomial formula, which tells us that

$$\Pr[\text{exactly } k \text{ successes}] = \binom{n}{k} p^k (1-p)^{n-k}.$$

This formula is really telling us the PMF of the random variable that counts the number of successes.

The binomial distribution

Suppose that a random variable $\mathbf{X}$ has range $R_{\mathbf{X}} = \{0, 1, 2, \ldots, n\}$ and that for each $k \in R_{\mathbf{X}}$,

$$P_{\mathbf{X}}(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We say $\mathbf{X}$ has the **binomial distribution with parameters n and p.**

This has a shorthand: $\mathbf{X} \sim$ *Binomial*(n, p). This is how we point to a random variable and say: "Look! We can understand this quantity using the binomial formula!"

Even though we have an idea for what kind of random experiment causes this behavior, we say $\mathbf{X} \sim$ *Binomial*(n, p) if $P_{\mathbf{X}}$ has the right form, no matter what the underlying random experiment is doing.

Examples

In general, suppose that $A_1, A_2, \ldots, A_n$ are independent events, with $\Pr[A_i] = p$ for all i, and $\mathbf{X}$ counts the number of these events that occur. Then $\mathbf{X} \sim$ *Binomial*(n, p).

- Let $\mathbf{X}$ count the number of heads in a sequence of n fair coinflips. Then $\mathbf{X} \sim$ *Binomial*$(n, \frac{1}{2})$.
- Let $\mathbf{Y}$ count the number of 6's in a sequence of n fair die rolls. Then $\mathbf{Y} \sim$ *Binomial*$(n, \frac{1}{6})$.
- Let $\mathbf{Z}$ be the position of a particle that starts at 0 on the number line and takes n steps of $+1$ or -1 with equal probability. Then $\mathbf{Z}$ is not binomial. . .

  . . . but the number of $+1$ steps is $\mathbf{W} \sim$ *Binomial*$(n, \frac{1}{2})$, and $\mathbf{Z}$ can be written as $\mathbf{W} - (n - \mathbf{W}) = 2\mathbf{W} - n$.

The Bernoulli distribution

We say that $\mathbf{X}$ has the **Bernoulli distribution** with parameter p if $R_{\mathbf{X}} = \{0, 1\}$ with $P_{\mathbf{X}}(1) = p$ and $P_{\mathbf{X}}(0) = 1 - p$.

Shorthand notation: $\mathbf{X} \sim$ *Bernoulli*(p).

These are not very interesting on their own. Later, we will use these as building blocks to compute things about more complicated distributions.

Sometimes, we use these to turn events into random variables. Given an event A, we can define $\mathbf{X}$ to be 1 if A happens, and 0 if it does not. (Formally, $\mathbf{X}(s) = 1$ if $s \in A$ and $\mathbf{X}(s) = 0$ if $s \notin A$.)

Then $\mathbf{X} \sim$ *Bernoulli*(p) with $p = \Pr[A]$. We also call $\mathbf{X}$ the **indicator random variable** of A.

Rolling until we roll a 6

Suppose that we have a single fair die, and we roll it until we roll a ⚅ for the first time. Let $\mathbf{X}$ be the number of rolls. What can we say about $\mathbf{X}$?

Its range $R_{\mathbf{X}}$ is infinite: $R_{\mathbf{X}} = \mathbb{N} = \{1, 2, 3, \dots\}$. We have:

- $P_{\mathbf{X}}(1) = \Pr[\mathbf{X} = 1] = \frac{1}{6}$: this is the probability we roll ⚅ on the first try.
- $P_{\mathbf{X}}(2) = \Pr[\mathbf{X} = 2] = \frac{5}{6} \cdot \frac{1}{6}$: we have a $\frac{5}{6}$ chance of **not** rolling ⚅ on the first try, and a $\frac{1}{6}$ chance of rolling ⚅ on the second try.
- $P_{\mathbf{X}}(3) = \Pr[\mathbf{X} = 3] = (\frac{5}{6})^2 \cdot \frac{1}{6}$. We must avoid ⚅ on the first two rolls, but then get ⚅ on the third roll.
- In general, $P_{\mathbf{X}}(k) = (\frac{5}{6})^{k-1} \cdot \frac{1}{6}$.

The geometric distribution

We say that $\mathbf{X}$ has the **geometric distribution** with parameter p if $R_{\mathbf{X}} = \mathbb{N} = \{1, 2, 3, \dots\}$ and for each $k \in \mathbb{N}$,

$$P_{\mathbf{X}}(k) = p(1-p)^{k-1}.$$

Shorthand notation: $\mathbf{X} \sim \textit{Geometric}(p)$.

The dice-rolling random variable on the previous slide had the $\textit{Geometric}(\frac{1}{6})$ distribution.

If $A_1, A_2, A_3, \dots$ are independent events with $\Pr[A_i] = p$ for all i, and $\mathbf{X}$ is the first value of k for which event A_k occurs, then we get $\mathbf{X} \sim \textit{Geometric}(p)$.

Properties of the geometric distribution

To check that $P_{\mathbf{X}}(k) = p(1-p)^{k-1}$ is a valid PMF (with $R_{\mathbf{X}} = \mathbb{N}$), we can check that

$$\sum_{i=1}^{\infty} p(1-p)^{i-1} = 1.$$

This is a geometric series that starts at $a = p$, and has ratio $r = 1 - p$, so it converges to $\frac{a}{1-r} = \frac{p}{1-(1-p)} = 1$.

Similarly, we can compute $\Pr[\mathbf{X} > k]$ using a geometric series:

$$\Pr[\mathbf{X} > k] = \sum_{i=k+1}^{\infty} P_{\mathbf{X}}(i) = \sum_{i=k+1}^{\infty} p(1-p)^{i-1} = \frac{p(1-p)^k}{1-(1-p)} = (1-p)^k.$$

This has an intuitive meaning: $\mathbf{X} > k$ if the first k trials all fail.

Memorylessness

Let's look at a conditional probability problem. If $\mathbf{X} \sim$ *Geometric*(p), what is $\Pr[\mathbf{X} = \ell \mid \mathbf{X} > k]$? *(Let's assume $\ell > k$.)*

By definition,

$$\Pr[\mathbf{X} = \ell \mid \mathbf{X} > k] = \frac{\Pr[\mathbf{X} = \ell \text{ and } \mathbf{X} > k]}{\Pr[\mathbf{X} > k]} = \frac{\Pr[\mathbf{X} = \ell]}{\Pr[\mathbf{X} > k]} = \frac{p(1-p)^{\ell-1}}{(1-p)^k}$$

which simplifies to $p(1-p)^{\ell-k-1}$. In other words,

$$\Pr[\mathbf{X} = \ell \mid \mathbf{X} > k] = \Pr[\mathbf{X} = \ell - k].$$

This also has an intuitive meaning: when we are given that the first k trials fail, $\mathbf{X} = \ell$ iff it takes $\ell - k$ more trials to succeed.

Expected value

Given a random variable $\mathbf{X}$, its **expected value** $\mathbb{E}[\mathbf{X}]$ is “what $\mathbf{X}$ is on average”.

It is defined by the formula

$$\mathbb{E}[\mathbf{X}] = \sum_{x \in R_{\mathbf{X}}} x \cdot P_{\mathbf{X}}(x).$$

The idea: we are taking an average of the possible values in $R_{\mathbf{X}}$, each value taken proportionally to how often it appears.

In the special case that all values of $\mathbf{X}$ are equally likely, this becomes an ordinary average.

A few examples

Roll a 6-sided die; let $\mathbf{X}$ be the value of the outcome. Then

$$\mathbb{E}[\mathbf{X}] = \frac{1+2+3+4+5+6}{6} = 3.5.$$

Flip a fair coin 3 times; let $\mathbf{Y}$ be the number of heads we see. Then

$$\mathbb{E}[\mathbf{Y}] = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1.5.$$

Draw a card; let $\mathbf{Z} = 1$ if we draw Q♠, and $\mathbf{Z} = 0$ if we don't. Then

$$\mathbb{E}[\mathbf{Z}] = 0 \cdot \frac{51}{52} + 1 \cdot \frac{1}{52} = \frac{1}{52}.$$

Betting games

"Give me a quarter and guess the card I drew. If you get it right, I'll pay you \$10."

Is this game worth it?

Let $\mathbf{X}$ be the amount of money you get. Then $R_{\mathbf{X}} = \{-0.25, +9.75\}$ with

$$P_{\mathbf{X}}(-0.25) = \frac{51}{52}, \qquad P_{\mathbf{X}}(9.75) = \frac{1}{52}.$$

The expected value is

$$\mathbb{E}[\mathbf{X}] = -0.25 \cdot \frac{51}{52} + 9.75 \cdot \frac{1}{52} \approx -0.058.$$

On average, you lose about 6 cents every time you play this game.

Linearity of expectation

Linearity of expectation tells us that when we add random variables, expected values add. When we scale random variables, expected values scale. Formally:

- $\mathbb{E}[\mathbf{X} + \mathbf{Y}] = \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{Y}]$.
- $\mathbb{E}[\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n] = \mathbb{E}[\mathbf{X}_1] + \mathbb{E}[\mathbf{X}_2] + \cdots + \mathbb{E}[\mathbf{X}_n]$.
- $\mathbb{E}[a\mathbf{X} + b] = a\mathbb{E}[\mathbf{X}] + b$ (for fixed $a, b \in \mathbb{R}$).

For example, suppose we roll two dice. Let $\mathbf{X}$ be the value of the first and $\mathbf{Y}$ be the value of the second. Then

$$\mathbb{E}[\mathbf{X} + \mathbf{Y}] = \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{Y}] = 3.5 + 3.5 = 7$$

rather than having to do $2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + \cdots + 12 \cdot \frac{1}{36}$.

Saving work using linearity of expectation

Flip 5 coins; let $\mathbf{X}$ be the number of heads. Then

$$\mathbb{E}[\mathbf{X}] = 0 \cdot \frac{1}{32} + 1 \cdot \frac{5}{32} + 2 \cdot \frac{10}{32} + 3 \cdot \frac{10}{32} + 4 \cdot \frac{5}{32} + 5 \cdot \frac{1}{32}$$

which is tedious to compute. Here are two shortcuts.

- Let $\mathbf{Y}$ be the number of tails in those same 5 coinflips. Then $\mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{X} + \mathbf{Y}] = 5$.

  But $\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{Y}]$. So $2\mathbb{E}[\mathbf{X}] = 5$, or $\mathbb{E}[\mathbf{X}] = 2.5$.

- Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_5$ be 1 if the $1^{\text{st}}, 2^{\text{nd}}, \ldots, 5^{\text{th}}$ coin lands heads.

  $$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{X}_1] + \mathbb{E}[\mathbf{X}_2] + \mathbb{E}[\mathbf{X}_3] + \mathbb{E}[\mathbf{X}_4] + \mathbb{E}[\mathbf{X}_5]$$

  which simplifies to $\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2.5$.

Bernoulli distribution

Let A be an event. Let $\mathbf{X}$ be the indicator random variable of A: $\mathbf{X} = 1$ when A happens and $\mathbf{X} = 0$ when A doesn't happen. Then

$$\mathbb{E}[\mathbf{X}] = 0 \cdot P_{\mathbf{X}}(0) + 1 \cdot P_{\mathbf{X}}(1) = P_{\mathbf{X}}(1) = \Pr[\mathbf{X} = 1] = \Pr[A].$$

In other words, if $\mathbf{X} \sim$ *Bernoulli*(p), then $\mathbb{E}[\mathbf{X}] = p$.

A common strategy for dealing with complicated expected values:

1. Write your random variable as a sum of indicator random variables.
2. Compute their expected values (that's just probability).
3. By linearity of expectation, we can just add these together.

Binomial distribution

Recall: if we have n independent events $A_1, A_2, \ldots, A_n$, each with probability p, and $\mathbf{X}$ counts the number of them that occur, then $\mathbf{X} \sim$ *Binomial*(n, p).

What is $\mathbb{E}[\mathbf{X}]$? By the definition:

$$\mathbb{E}[\mathbf{X}] = \sum_{k=0}^{n} k \cdot P_{\mathbf{X}}(k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = ???$$

By linearity of expectation: write $\mathbf{X} = \sum_{i=1}^{n} \mathbf{X}_i$, where $\mathbf{X}_i = 1$ if A_i occurs, and $\mathbf{X}_i = 0$ otherwise. Then

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} \mathbb{E}[\mathbf{X}_i] = \sum_{i=1}^{n} p = np.$$

Geometric distribution

Recall: if we do many independent trials with a probability p of success on each one, and $\mathbf{X}$ is the number of trials it takes to get a success, then $\mathbf{X} \sim$ *Geometric*(p). What is $\mathbb{E}[\mathbf{X}]$?

- Let $\mathbf{X}_i = 1$ if $\mathbf{X} \geq i$: if the first $i-1$ trials all fail. Then

$$\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 + \dots$$

- By linearity of expectation **and cheating slightly**, we have

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{X}_1] + \mathbb{E}[\mathbf{X}_2] + \mathbb{E}[\mathbf{X}_3] + \dots$$

- We have $\mathbb{E}[\mathbf{X}_i] = \Pr[\mathbf{X}_i = 1] = \Pr[\mathbf{X} \geq i] = (1-p)^{i-1}$. So

$$\mathbb{E}[\mathbf{X}] = 1 + (1-p) + (1-p)^2 + \dots = \frac{1}{1-(1-p)} = \frac{1}{p}.$$

Examples

- You roll 300 dice. What is the expected number of ⚅'s?

  The number of ⚅'s has the *Binomial*$(300, \frac{1}{6})$ distribution, so its expected value is $300 \cdot \frac{1}{6} = 50$.

- You roll dice until you roll a ⚅. How many rolls does this take?

  The number of trials until a success has the *Geometric*$(\frac{1}{6})$ distribution, so its expected value is 6.

  Note that this doesn't **guarantee** that you'll have rolled a ⚅ after 6 rolls! Actually, there is a probability of $(1 - \frac{1}{6})^6 \approx 0.335$ that it'll take longer than that. . .

Another linearity of expectation example

You have 100 chickens sitting in a circle. At the same time, each chicken pecks one of the two adjacent chickens at random: either its left neighbor or its right neighbor.

What is the expected number of chickens that remain unpecked?

To solve this problem, we write the number of unpecked chickens as a sum

$$\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_{100}$$

where $\mathbf{X}_i = 1$ if the i^{th} chicken was unpecked, and 0 otherwise.

We have $\mathbb{E}[\mathbf{X}_i] = \Pr[\mathbf{X}_i = 1] = (\frac{1}{2})^2 = \frac{1}{4}$: for a chicken to remain unpecked, both neighbors have to peck in the other direction.

Therefore the expected number of unpecked chickens is $100 \cdot \frac{1}{4} = 25$.

The Pascal distribution

We do many independent trials, each of which succeeds with probability p.

What we know so far:

- The number of the trial on which we first succeed has the *Geometric*(p) distribution.
- This is k with probability $p(1-p)^{k-1}$, and its expected value is $\frac{1}{p}$.

The Pascal distribution is a generalization.

We have $\mathbf{X} \sim$ *Pascal*(m, p) if $\mathbf{X}$ is the number of the trial on which we get the m^{th} success.

Basic facts

Let $\mathbf{X} \sim \mathit{Pascal}(m, p)$. What is the range of $\mathbf{X}$?

$$R_{\mathbf{X}} = \{m, m+1, m+2, m+3, \dots\}$$

because the earliest trial which can have the m^{th} success is the m^{th} trial, but it can happen arbitrarily late if we have bad luck.

What is $P_{\mathbf{X}}(k) = \Pr[\mathbf{X} = k]$? This requires:

- The first $k-1$ trials to have exactly $m-1$ successes, which happens with probability $\binom{k-1}{m-1} p^{m-1} (1-p)^{k-m}$.
- The k^{th} trial to succeed, which happens with probability p.

Therefore

$$P_{\mathbf{X}}(k) = \binom{k-1}{m-1} p^{m} (1-p)^{k-m}.$$

Expected value

Let $\mathbf{X} \sim$ *Pascal*(m, p). What is $\mathbb{E}[\mathbf{X}]$? This is hard to compute from the definition, but we can use linearity of expectation. Let:

- $\mathbf{X}_1$ be the number of trials until we get the first success;
- $\mathbf{X}_2$ be the number of trials between the first and second success;
- and so on, with $\mathbf{X}_m$ the number of trials between the $(m-1)^{\text{th}}$ and m^{th} success.

Each $\mathbf{X}_i$ has the *Geometric*(p) distribution, so $\mathbb{E}[\mathbf{X}_i] = \frac{1}{p}$ for each i.

The sum $\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_m$ just gives $\mathbf{X}$.

Therefore $\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{X}_1] + \mathbb{E}[\mathbf{X}_2] + \cdots + \mathbb{E}[\mathbf{X}_m] = \frac{m}{p}$.

The hypergeometric distribution

Suppose we have a bag with b blue marbles and r red marbles. We draw k marbles from the bag, without replacement. (Assume $k \leq b + r$.)

Let $\mathbf{X}$ be the number of blue marbles in our sample. We call the distribution of $\mathbf{X}$ the **hypergeometric distribution** with parameters b, r, k.

Shorthand: $\mathbf{X} \sim \mathit{Hypergeometric}(b, r, k)$.

Some other examples of the hypergeometric distribution:

- Number of ♠ you draw, when you draw k cards from a deck.
- Result of polling k people on who they will vote for, in terms of the vote counts b and r in the overall population.

Basic facts

What is the range $R_{\mathbf{X}}$? (A bit tricky.)

- The least possible value of $\mathbf{X}$ is $\max\{0, k-r\}$.
- The greatest possible value of $\mathbf{X}$ is $\min\{b, k\}$.

For $x \in R_{\mathbf{X}}$, what is $P_{\mathbf{X}}(x) = \Pr[\mathbf{X} = x]$?

There are $\binom{b+r}{k}$ total ways to draw k of the $b+r$ marbles. The number of ways to draw x blue marbles and $k-x$ red marbles is $\binom{b}{x}\binom{r}{k-x}$. Therefore

$$P_{\mathbf{X}}(x) = \frac{\binom{b}{x}\binom{r}{k-x}}{\binom{b+r}{k}}.$$

Expected value

Let $\mathbf{X} \sim$ *Hypergeometric*(b, r, k). Once again, we will use linearity of expectation to find $\mathbb{E}[\mathbf{X}]$.

Let $\mathbf{X}_i = 1$ if the i^{th} marble we draw is blue, so that

$$\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_k.$$

Then $\mathbb{E}[\mathbf{X}_i] = \Pr[\mathbf{X}_i = 1]$, the probability we draw a blue marble.

We have $\mathbb{E}[\mathbf{X}_1] = \frac{b}{b+r}$ going by the totals. But actually, $\mathbb{E}[\mathbf{X}_i] = \frac{b}{b+r}$ for all i: it doesn't matter which marble in our sample is the i^{th} one. Therefore

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{X}_1] + \mathbb{E}[\mathbf{X}_2] + \cdots + \mathbb{E}[\mathbf{X}_k] = \frac{bk}{b+r}.$$

Four for the price of one!

Suppose that out of KSU's 1000 academic staff, 600 prefer coffee and 400 prefer tea. We take a random sample of 50 of them.

	Prefer coffee	Prefer tea	Total
Random sample	$\mathbf{X}$	$50 - \mathbf{X}$	50
Not sampled	$600 - \mathbf{X}$	$\mathbf{X} + 350$	950
Total	600	400	1000

There are $\mathbf{X} \sim \mathit{Hypergeometric}(600, 400, 50)$ coffee drinkers in our sample. We can solve for the other three cells... and they are also hypergeometric: e.g., $600 - \mathbf{X} \sim \mathit{Hypergeometric}(600, 400, 950)$.

Each cell has two descriptions: e.g., $\mathbf{X} \sim \mathit{Hypergeometric}(50, 950, 600)$. "Out of 600 coffee drinkers, how many are in our random sample?"

The Poisson distribution

The Poisson distribution is the hardest to describe, because it comes out of a limiting process.

Suppose we are doing n independent trials, where n is very large. Each trial succeeds with probability p, where p is very small.

- If we know n and p, then the number of successes is just *Binomial*(n, p).
- But suppose we do not know n or p; from experience, we know the product $\lambda = np$, the average number of successes.

  Rather than take a guess at a very large n, and look at *Binomial*$(n, \frac{\lambda}{n})$, we take the limit as $n \to \infty$ of this distribution.

Examples of the Poisson distribution

Here are some examples of things we model with the Poisson distribution.

- The number of Google searches done in a given minute.
- The number of emails you get each day.
- The number of raindrops that land on your window each second.
- The number of traffic accidents in the US on a given week.

In combinatorics, it is the most common behavior to expect from a limit of random variables with a fixed mean.

Basic facts

We write $\mathbf{X} \sim \mathit{Poisson}(\lambda)$ if $\mathbf{X}$ has the Poisson distribution with **rate** λ. (We think $\mathbb{E}[\mathbf{X}] = \lambda$, since *Binomial*$(n, \frac{\lambda}{n})$ has mean λ for all n.)

For all $k \geq 0$, we have $\Pr[\mathbf{X} = k] = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$.

- $\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} \approx \frac{n^k}{k!}$. (Formally, $\lim_{n\to\infty} \frac{\binom{n}{k}}{n^k/k!} = 1$.)
- This cancels with $(\frac{\lambda}{n})^k$ to give $\frac{\lambda^k}{k!}$.
- $(1 - \frac{\lambda}{n})^{n-k} \approx (1 - \frac{\lambda}{n})^n$, which converges to $e^{-\lambda}$ as $n \to \infty$.

This gives us the actual PMF:

$$P_{\mathbf{X}}(k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Expected value

A key fact when dealing with the Poisson distribution:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

This tells us that our distribution is legit: if $\mathbf{X} \sim \mathit{Poisson}(\lambda)$, then

$$\sum_{k=0}^{\infty} P_{\mathbf{X}}(k) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

It also confirms that the expected value is what we expected:

$$\sum_{k=0}^{\infty} k P_{\mathbf{X}}(k) = \sum_{k=0}^{\infty} k \cdot e^{-\lambda} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda.$$

Application: the hat swap problem

A party has n guests, each with a hat. We take their hats and randomly redistribute them.

Question 1. What is the expected number of people that get their own hat back?

Let $\mathbf{X}_i = 1$ if the i^{th} guest gets their own hat back, and $\mathbf{X}_i = 0$ otherwise. Then:

- The number of guests that get their own hat back is $\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n$.
- $\mathbb{E}[\mathbf{X}_i] = \frac{1}{n}$: the guest could get n hats, 1 of which is correct.

Therefore the expected value is $\mathbb{E}[\mathbf{X}_1 + \cdots + \mathbf{X}_n] = \frac{1}{n} + \cdots + \frac{1}{n} = 1$.

Application: the hat swap problem

Question 2. What is the probability that at least one guest gets their own hat back?

The exact solution is tricky and requires lots of the inclusion-exclusion principle. (Section 2.1.5, problem 7 in the textbook.)

We can get an approximate probability for large n by guessing that the distribution is approximately Poisson. (Reasoning: the events that the guests get their own hats back aren't quite independent, but they're close. There are many of these events, and they're individually unlikely.)

The correct Poisson approximation to use is *Poisson*(1), because the expected value is 1. If $\mathbf{X} \sim$ *Poisson*(1), then

$$\Pr[\mathbf{X} \geq 1] = 1 - P_{\mathbf{X}}(0) = 1 - e^{-1}\frac{1^0}{0!} = 1 - \frac{1}{e}.$$

Functions of a random variable

If $\mathbf{X}$ is a random variable, and g is a function $\mathbb{R} \to \mathbb{R}$ (or even just $R_{\mathbf{X}} \to \mathbb{R}$), then $g(\mathbf{X})$ is another random variable.

Some examples:

- Gambling. If all you do is flip a coin 5 times and count the number of heads, then $\mathbf{X} \sim \mathit{Binomial}(5, \frac{1}{2})$.

  If you win $\$1$ when the coin lands heads and lose $\$1$ when the coin lands tails, your winnings are $\mathbf{X} - (5 - \mathbf{X}) = 2\mathbf{X} - 5$.

- Say you invite 10 people to a round-robin chess tournament, and each accepts with probability $\frac{2}{3}$. Then $\mathbf{X} \sim \mathit{Binomial}(10, \frac{2}{3})$ is the number of people at the tournament.

  The **number of games played** is $\binom{\mathbf{X}}{2} = \frac{\mathbf{X}(\mathbf{X}-1)}{2}$.

Injective functions

When it comes to understanding $g\mathbf{X})$, the key factor is whether g is injective.

(A function $g : R_{\mathbf{X}} \to \mathbb{R}$ is **injective** if for any $x_1, x_2 \in R_{\mathbf{X}}$ with $x_1 \neq x_2$, we have $g(x_1) \neq f(x_2)$.)

Injective functions just "relabel" $R_{\mathbf{X}}$. In the gambling example:

- $R_{\mathbf{X}} = \{0, 1, 2, 3, 4, 5\}$ and $P_{\mathbf{X}}$ has values $\frac{1}{32}, \frac{5}{32}, \frac{10}{32}, \frac{10}{32}, \frac{5}{32}, \frac{1}{32}$ on these 6 numbers.
- If $\mathbf{Y} = 2\mathbf{X} - 5$, then $R_{\mathbf{Y}} = \{-5, -3, -1, 1, 3, 5\}$ and $P_{\mathbf{Y}}$ has values $\frac{1}{32}, \frac{5}{32}, \frac{10}{32}, \frac{10}{32}, \frac{5}{32}, \frac{1}{32}$ on these 6 numbers.

But $\mathbf{Z} = |\mathbf{Y}|$ has $R_{\mathbf{Z}} = \{1, 3, 5\}$ and its PMF is given by $P_{\mathbf{Z}}(1) = \frac{10}{16}$, $P_{\mathbf{Z}}(3) = \frac{5}{16}$, and $P_{\mathbf{Z}}(5) = \frac{1}{16}$.

The general idea

In general, when $\mathbf{Y} = g(\mathbf{X})$, each value $y \in R_{\mathbf{Y}}$ corresponds to an entire set of values in $R_{\mathbf{X}}$. To find $P_{\mathbf{Y}}(y)$, sum over that set:

$$P_{\mathbf{Y}}(y) = \Pr[\mathbf{Y} = y] = \Pr[g(\mathbf{X}) = y] = \sum_{x:g(x)=y} P_{\mathbf{X}}(x).$$

For example: let $\mathbf{X} \sim$ *Geometric*$(\frac{1}{2})$, and let $\mathbf{Y} = \mathbf{X} \bmod 3$ (the remainder when $\mathbf{X}$ is divided by 3). $\mathbf{Y}$ has range $R_{\mathbf{Y}} = \{0, 1, 2\}$.

To find something like $P_{\mathbf{Y}}(1)$, first find the corresponding subset of $R_{\mathbf{X}}$: it is $\{1, 4, 7, 10, 13, \dots\}$. Then add up the values of $P_{\mathbf{X}}(k) = \frac{1}{2^k}$ over this set:

$$\frac{1}{2^1} + \frac{1}{2^4} + \frac{1}{2^7} + \frac{1}{2^{10}} + \dots = \frac{1/2}{1 - 1/2^3} = \frac{4}{7}.$$

Expected value of a function

Our main motivation in looking at functions of random variables is finding expected values $\mathbb{E}[g(\mathbf{X})]$. There are three main approaches:

1. When we are dealing with a linear function $g(x) = ax + b$, linearity of expectation says that $\mathbb{E}[a\mathbf{X} + b] = a\mathbb{E}[\mathbf{X}] + b$, and we just have to find $\mathbb{E}[\mathbf{X}]$.
2. We can always figure out the distribution of the random variable $g(\mathbf{X})$, and then apply its definition.
3. A technique called LOTUS lets us "pretend that g is injective" and use the PMF of $\mathbf{X}$, rather than $g(\mathbf{X})$, for computing $\mathbb{E}[g(\mathbf{X})]$.

Expected value of $g(\mathbf{X})$

Warning: In general, $\mathbb{E}[g(\mathbf{X})]$ and $g(\mathbb{E}[\mathbf{X}])$ are very different!

Example: suppose $\mathbf{X} \sim$ *Geometric*$(\frac{1}{2})$ (e.g. $\mathbf{X}$ might count the number of coin tosses until the coin lands heads). Take $g(n) = 2^n$.

- We have already seen that $\mathbb{E}[\mathbf{X}] = \frac{1}{1/2} = 2$; $2^{\mathbb{E}[\mathbf{X}]} = 4$.
- $2^{\mathbf{X}}$ is 2^k with probability $(\frac{1}{2})^k$. Therefore

$$\mathbb{E}[2^{\mathbf{X}}] = \sum_{k=1}^{\infty} 2^k \cdot \left(\frac{1}{2}\right)^k = \sum_{k=1}^{\infty} 1 = \infty.$$

There is one exception: we do have $\mathbb{E}[a\mathbf{X} + b] = a\mathbb{E}[\mathbf{X}] + b$.

(This is one part of linearity of expectation).

A typical problem

Let $\mathbf{X}$ be the outcome of rolling a 6-sided die: $\mathbf{X}$ is equally likely to be $1, 2, 3, 4, 5, 6$. We've already seen that $\mathbb{E}[\mathbf{X}] = 3.5$.

Let's find $\mathbb{E}[(\mathbf{X} - 3.5)^2]$. We will see what this computation tells us in the next lecture. Using the definition of expected value:

1. First, with $\mathbf{Y} = (\mathbf{X} - 3.5)^2$, we find the distribution of $\mathbf{Y}$.

   The values $\mathbf{X} = 1, 6$ give $\mathbf{Y} = 2.5^2$, while $\mathbf{X} = 2, 5$ both give $\mathbf{Y} = 1.5^2$ and $\mathbf{X} = 3, 4$ both give $\mathbf{Y} = 0.5^2$. The range $R_{\mathbf{Y}}$ is $\{0.25, 2.25, 6.25\}$, with $P_{\mathbf{Y}}(0.25) = P_{\mathbf{Y}}(2.25) = P_{\mathbf{Y}}(6.25) = \frac{1}{3}$.

2. We compute the final answer from the definition of $\mathbb{E}[\mathbf{Y}]$:

$$\mathbb{E}[\mathbf{Y}] = 0.25 \cdot \frac{1}{3} + 2.25 \cdot \frac{1}{3} + 6.25 \cdot \frac{1}{3} = 2.91666\ldots$$

Law of the Unconscious Statistician

Another approach for computing $\mathbb{E}[g(\mathbf{X})]$ is the following rule:

$$\mathbb{E}[g(\mathbf{X})] = \sum_{x \in R_{\mathbf{X}}} g(x) \cdot P_{\mathbf{X}}(x).$$

Why is this different from the definition? Consider an earlier example: $\mathbf{X} \sim$ *Geometric*$(\frac{1}{2})$ and $\mathbf{Y} = \mathbf{X} \bmod 3$.

- The definition says we first compute $P_{\mathbf{Y}}(0) = \frac{1}{7}$, $P_{\mathbf{Y}}(1) = \frac{4}{7}$, and $P_{\mathbf{Y}}(2) = \frac{2}{7}$. Then,

$$\mathbb{E}[\mathbf{Y}] = 0 \cdot \frac{1}{7} + 1 \cdot \frac{4}{7} + 2 \cdot \frac{2}{7} = \frac{8}{7}.$$

- LOTUS says we can take the infinite sum

$$1 \cdot \frac{1}{2^1} + 2 \cdot \frac{1}{2^2} + 0 \cdot \frac{1}{2^3} + 1 \cdot \frac{1}{2^4} + 2 \cdot \frac{1}{2^5} + 0 \cdot \frac{1}{2^6} + 1 \cdot \frac{1}{2^7} + \dots$$

Motivation for variance

Sometimes two random variables have the same mean, but one is more tightly clustered around it than the other.

- The expected value of rolling a 6-sided die is 3.5. If we roll a 4-sided die and add 1, the expected value is 3.5, but only values $\{2, 3, 4, 5\}$ are possible.
- If we flip 100 coins and let $\mathbf{X}$ the number of heads, then $\mathbf{X} \sim \textit{Binomial}(100, \frac{1}{2})$ and $\mathbb{E}[\mathbf{X}] = 50.$

  Choosing a uniformly random number from $\{0, 1, 2, \ldots, 100\}$ also has expected value 50, but extreme values are more likely.

We would like to quantify how "spread out" a distribution is to capture these differences.

Variance

Let $\mu = \mathbb{E}[\mathbf{X}]$ (Greek letter “m” for “mean”) for short. If we want to measure how far $\mathbf{X}$ typically is from μ:

- The first thing we might try is $\mathbb{E}[|\mathbf{X} - \mu|]$. This would do what we want, but turns out to be hard to work with.
- The second thing we try is $\mathbb{E}[(\mathbf{X} - \mu)^2]$. We call this quantity $\mathrm{Var}[\mathbf{X}]$, the **variance** of $\mathbf{X}$.

We can also compute the variance differently:

$$\mathbb{E}[(\mathbf{X} - \mu)^2] = \mathbb{E}[\mathbf{X}^2 - 2\mathbf{X}\mu + \mu^2] = \mathbb{E}[\mathbf{X}^2] - 2\mu\mathbb{E}[\mathbf{X}] + \mu^2$$

and since $\mu = \mathbb{E}[\mathbf{X}]$, this simplifies to $\mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2$.

Some simple examples

- Let $\mathbf{X} \sim$ *Bernoulli*(p): $P_{\mathbf{X}}(1) = p$ and $P_{\mathbf{X}}(0) = 1 - p$.

  Then $\mathbb{E}[\mathbf{X}] = p$ and actually $\mathbf{X}^2 = \mathbf{X}$ so $\mathbb{E}[\mathbf{X}^2] = p$ as well.

  This gives us $\mathrm{Var}[\mathbf{X}] = p - p^2 = p(1 - p)$.
- Let $\mathbf{Y}$ be the result of rolling a 6-sided die. Then $\mathbb{E}[\mathbf{Y}] = 3.5$ and we have already computed $\mathrm{Var}[\mathbf{Y}] = \mathbb{E}[(\mathbf{Y} - 3.5)^2] \approx 2.91667$.
- Let $\mathbf{Z}$ be 1 or 6 with equal probability; we also have $\mathbb{E}[\mathbf{Z}] = 3.5$. However, $(\mathbf{Z} - 3.5)^2$ is **always** 6.25, so $\mathrm{Var}[\mathbf{Z}] = 6.25$.

Compare the last two variables: they have the same expected value, but $\mathbf{Z}$ is more “spread out” and correspondingly has higher variance.

Variance and standard deviation

A few properties of variance:

- $\mathrm{Var}[\mathbf{X}+b] = \mathrm{Var}[\mathbf{X}]$, since $\mathbf{X}+b$ has mean $\mathbb{E}[\mathbf{X}]+b$, and then we take the squared difference from **that** mean.
- $\mathrm{Var}[a\mathbf{X}] = a^2\mathrm{Var}[\mathbf{X}]$. We can check this when $\mathbb{E}[\mathbf{X}]=0$: then $\mathbb{E}[(a\mathbf{X})^2] = \mathbb{E}[a^2\mathbf{X}^2] = a^2\mathbb{E}[\mathbf{X}^2]$, so $\mathrm{Var}[a\mathbf{X}] = a^2\mathrm{Var}[\mathbf{X}]$.

  By the previous bullet point, it's true for any other $\mathbb{E}[\mathbf{X}]$ as well.

This second bullet point motivates defining the **standard deviation**: $\mathrm{SD}[\mathbf{X}] = \sqrt{\mathrm{Var}[\mathbf{X}]}$, so that $\mathrm{SD}[a\mathbf{X}] = |a|\cdot\mathrm{SD}[\mathbf{X}]$.

We often write σ (Greek letter "s") for $\mathrm{SD}[\mathbf{X}]$, and $\mathrm{Var}[\mathbf{X}] = \sigma^2$.

Some examples

Here are some distributions with the same mean and different standard deviations, to help your intuition.

- Suppose that $\mathbf{X}$ takes on each value in the range $\{0, 1, \ldots, 100\}$ with the same probability $\frac{1}{101}$.

  This has mean $\mathbb{E}[\mathbf{X}] = 50$ and $\mathrm{SD}[\mathbf{X}] = 5\sqrt{34} \approx 29.15$.

- Suppose that $\mathbf{Y} \sim \textit{Binomial}(100, \frac{1}{2})$.

  This has mean $\mathbb{E}[\mathbf{Y}] = 50$ and $\mathrm{SD}[\mathbf{Y}] = 5$.

- Suppose that $\mathbf{Z}$ is 0 or 100 with probability $\frac{1}{2}$ each.

  This has mean $\mathbb{E}[\mathbf{Z}] = 50$ and $\mathrm{SD}[\mathbf{Y}] = 50$.

Quantifying what variance means

Very vaguely: it should be unlikely that $\mathbf{X}$ is many standard deviations away from its mean.

How do we put numbers on this vague principle?

- Later, we will learn about the normal distribution. If $\mathbf{X}$ is approximately normally distributed (many things are), we have $|\mathbf{X} - \mu| > k\sigma$ with probability less than $e^{-k^2/2}$.

  The probability is less than 0.3% when $k = 3$, and falls quickly.

- **Chebyshev's inequality** is true for any random variable. It says that $|\mathbf{X} - \mu| > k\sigma$ with probability less than $\frac{1}{k^2}$.

  This is the worst case: most random variables are more clustered than that.

Proving Chebyshev's inequality

Theorem. Suppose $\mathbb{E}[\mathbf{X}] = 0$ and $\mathrm{Var}[\mathbf{X}] = 1$. Then

$$\Pr\Big[|\mathbf{X}| \geq k\Big] \leq \frac{1}{k^2}.$$

Proof. When $\mathbb{E}[\mathbf{X}] = 0$, $\mathrm{Var}[\mathbf{X}]$ is just $\mathbb{E}[\mathbf{X}^2]$. Also, we can write $|\mathbf{X}| \geq k$ as $\mathbf{X}^2 \geq k^2$. Everything is about $\mathbf{Y} = \mathbf{X}^2$ now!

Let's throw in a third random variable $\mathbf{Z}$, which is 0 when $0 \leq \mathbf{Y} < k^2$ and k^2 when $\mathbf{Y} \geq k^2$. We always have $\mathbf{Z} \leq \mathbf{Y}$, so $\mathbb{E}[\mathbf{Z}] \leq \mathbb{E}[\mathbf{Y}] = 1$.

But $\mathbb{E}[\mathbf{Z}] = k^2 \cdot \Pr[\mathbf{Z} = k^2] = k^2 \cdot \Pr[\mathbf{Y} \leq k^2]$. Therefore $k^2 \cdot \Pr[\mathbf{Y} \leq k^2] \leq 1$, which gives us $\Pr[\mathbf{Y} \leq k^2] \leq \frac{1}{k^2}$. □

Standardizing a random variable

Suppose a random variable $\mathbf{X}$ has $\mathbb{E}[\mathbf{X}] = \mu$ and $\mathrm{Var}[\mathbf{X}] = \sigma^2$. Then

$$\mathbf{Z} = \frac{\mathbf{X} - \mu}{\sigma}$$

has $\mathbb{E}[\mathbf{Z}] = \frac{\mu - \mu}{\sigma} = 0$ and $\mathrm{Var}[\mathbf{Z}] = (\frac{1}{\sigma})^2 \sigma^2 = 1$.

$\mathbf{Z}$ is called the "standardization" of $\mathbf{X}$. It's useful for many things; today, it's useful for proving Chebyshev's inequality for $\mathbf{X}$.

By what we proved on the previous slide, $\Pr[|\mathbf{Z}| \geq k] \leq \frac{1}{k^2}$, or

$$\Pr\left[\left|\frac{\mathbf{X} - \mu}{\sigma}\right| \geq k\right] \leq \frac{1}{k^2} \implies \Pr\left[|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}.$$

Variance of a sum

Let $\mathbf{X}, \mathbf{Y}$ be two random variables. What is $\mathrm{Var}[\mathbf{X} + \mathbf{Y}]$?

By one formula, it is $\mathbb{E}[(\mathbf{X} + \mathbf{Y})^2] - \mathbb{E}[\mathbf{X} + \mathbf{Y}]^2$. We can simplify:

- $\mathbb{E}[(\mathbf{X} + \mathbf{Y})^2]$ to $\mathbb{E}[\mathbf{X}^2] + 2\mathbb{E}[\mathbf{X}\mathbf{Y}] + \mathbb{E}[\mathbf{Y}^2]$.
- $\mathbb{E}[\mathbf{X} + \mathbf{Y}]^2$ to $\mathbb{E}[\mathbf{X}]^2 + 2\mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}] + \mathbb{E}[\mathbf{Y}]^2$.

If the middle terms didn't exist, then the difference would be $\mathrm{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2$ plus $\mathrm{Var}[\mathbf{Y}] = \mathbb{E}[\mathbf{Y}^2] - \mathbb{E}[\mathbf{Y}]^2$.

Unfortunately, we also get this weird difference $\mathbb{E}[\mathbf{X}\mathbf{Y}] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}]$ which we call $\mathrm{Cov}[\mathbf{X}, \mathbf{Y}]$. Our final formula is:

$$\mathrm{Var}[\mathbf{X} + \mathbf{Y}] = \mathrm{Var}[\mathbf{X}] + 2\mathrm{Cov}[\mathbf{X}, \mathbf{Y}] + \mathrm{Var}[\mathbf{Y}].$$

Covariance

We define the **covariance** of two random variables $\mathbf{X}$ and $\mathbf{Y}$ to be

$$\mathrm{Cov}[\mathbf{X},\mathbf{Y}] = \mathbb{E}[\mathbf{X}\mathbf{Y}] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])]$$

(two equivalent formulas).

The **main** thing to know about covariance is that when $\mathrm{Cov}[\mathbf{X},\mathbf{Y}] = 0$, the formula simplifies to $\mathrm{Var}[\mathbf{X}+\mathbf{Y}] = \mathrm{Var}[\mathbf{X}] + \mathrm{Var}[\mathbf{Y}]$.

In this case, $\mathbf{X}$ and $\mathbf{Y}$ are called **uncorrelated** random variables.

We can also get some sense of the relationship between $\mathbf{X}$ and $\mathbf{Y}$ from $\mathrm{Cov}[\mathbf{X},\mathbf{Y}]$. If $\mathrm{Cov}[\mathbf{X},\mathbf{Y}] > 0$, larger $\mathbf{X}$ tends to correspond to larger $\mathbf{Y}$; if $\mathrm{Cov}[\mathbf{X},\mathbf{Y}] < 0$, the reverse is true.

Examples

Suppose we flip 10 coins. Let $\mathbf{X}$ be the number of **heads** in the **first** five flips; let $\mathbf{Y}$ be the number of **tails** in the **last** five flips.

- $\mathrm{Cov}[\mathbf{X}, \mathbf{Y}] = 0$.

  $\mathbf{X}$ and $\mathbf{Y}$ are **uncorrelated**: knowing how large or small $\mathbf{X}$ is tells us nothing about $\mathbf{Y}$.

- Let $\mathbf{Z}$ be the total number of heads. Then $\mathrm{Cov}[\mathbf{X}, \mathbf{Z}] = 1.25$.

  $\mathbf{X}$ and $\mathbf{Z}$ are **positively correlated**: if $\mathbf{X}$ is large, $\mathbf{Z}$ is more likely to be large.

- For the same $\mathbf{Z}$, $\mathrm{Cov}[\mathbf{Y}, \mathbf{Z}] = -1.25$.

  $\mathbf{Y}$ and $\mathbf{Z}$ are **negatively correlated**: if $\mathbf{Y}$ is large, $\mathbf{Z}$ is more likely to be small.

Variance of an n-term sum

For two random variables, we have

$$\mathrm{Var}[\mathbf{X}+\mathbf{Y}] = \mathrm{Var}[\mathbf{X}] + 2\mathrm{Cov}[\mathbf{X},\mathbf{Y}] + \mathrm{Var}[\mathbf{Y}].$$

What about $\mathrm{Var}[\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n]$?

We get a very similar thing when we expand $\mathbb{E}[(\mathbf{X}_1 + \cdots + \mathbf{X}_n)^2]$ and $(\mathbb{E}[\mathbf{X}_1] + \cdots + \mathbb{E}[\mathbf{X}_n])^2$. The resulting formula is

$$\sum_{i=1}^{n} \mathrm{Var}[\mathbf{X}_i] + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{Cov}[\mathbf{X}_i, \mathbf{X}_j].$$

If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are uncorrelated, the first sum is all that's left. Positive covariances increase it; negative covariances decrease it.

Independence

We say that $\mathbf{X}$ and $\mathbf{Y}$ are **independent** random variables if, for any $x \in R_{\mathbf{X}}$ and $y \in R_{\mathbf{Y}}$,

$$\Pr[\mathbf{X} = x \text{ and } \mathbf{Y} = y] = \Pr[\mathbf{X} = x] \cdot \Pr[\mathbf{Y} = y]$$

(That is, the events "$\mathbf{X} = x$" and "$\mathbf{Y} = y$" are independent.)

If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then we can simplify $\mathbb{E}[\mathbf{XY}]$ to $\mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}]$ with some algebra, and therefore $\mathbf{X}$ and $\mathbf{Y}$ are also uncorrelated: $\mathrm{Cov}[\mathbf{X}, \mathbf{Y}] = 0$. The reverse is not true!

Example: if $\mathbf{X}$ is chosen uniformly from $\{-1, 0, 1\}$, then $\mathbf{X}$ and $\mathbf{X}^2$ are uncorrelated, but very much not independent.

Independent random variables are uncorrelated

Claim. If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $\mathrm{Cov}[\mathbf{X}, \mathbf{Y}] = 0$.

Proof. We'll see where this double sum comes from later:

$$\mathbb{E}[\mathbf{XY}] = \sum_{x \in R_{\mathbf{X}}} \sum_{y \in R_{\mathbf{Y}}} x \cdot y \cdot \Pr[\mathbf{X} = x \text{ and } \mathbf{Y} = y].$$

By using independence, we get

$$\mathbb{E}[\mathbf{XY}] = \sum_{x \in R_{\mathbf{X}}} \sum_{y \in R_{\mathbf{Y}}} x \cdot y \cdot \Pr[\mathbf{X} = x] \cdot \Pr[\mathbf{Y} = y].$$

We can factor this into a part depending only on $\mathbf{X}$ and only on $\mathbf{Y}$:

$$\mathbb{E}[\mathbf{XY}] = \left(\sum_{x \in R_{\mathbf{X}}} x \cdot \Pr[\mathbf{X} = x] \right) \cdot \left(\sum_{y \in R_{\mathbf{Y}}} y \cdot \Pr[\mathbf{Y} = y] \right).$$

Therefore $\mathbb{E}[\mathbf{XY}] = \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}]$ and $\mathrm{Cov}[\mathbf{X}, \mathbf{Y}] = 0$. □

Some examples

Let $\mathbf{X}, \mathbf{Y}$ be the outcomes of two separate die rolls.

- $\mathbf{X}$ and $\mathbf{Y}$ are not just uncorrelated but also independent.
- $\mathbf{X} + \mathbf{Y}$ and $\mathbf{X} - \mathbf{Y}$ are still uncorrelated; we can check that $\operatorname{Cov}[\mathbf{X} + \mathbf{Y}, \mathbf{X} - \mathbf{Y}] = 0$. But they're not independent!
- Let $\mathbf{Z} = 1$ if $\mathbf{X} + \mathbf{Y}$ is odd and $\mathbf{Z} = 0$ if $\mathbf{X} + \mathbf{Y}$ is even. Then $\mathbf{X}$ and $\mathbf{Z}$ are independent: e.g.,

$$\Pr[\mathbf{X} = 4 \text{ and } \mathbf{Z} = 1] = \Pr[\mathbf{X} = 4 \text{ and } \mathbf{Y} \text{ is odd}] = \frac{1}{6} \cdot \frac{1}{2}.$$

- $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are **pairwise** but not **mutually** independent:

$$\Pr[\mathbf{X} = 4 \text{ and } \mathbf{Y} = 4 \text{ and } \mathbf{Z} = 1] = 0 \neq \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{2}.$$

Consequences of independence

If $\mathbf{X}$ and $\mathbf{Y}$ are independent (for example, two die rolls), then:

- Any probability about $\mathbf{X}$ and $\mathbf{Y}$ factors, when that makes sense.

  Example: $\Pr[\mathbf{X} > 2 \text{ and } \mathbf{Y} \neq 6] = \Pr[\mathbf{X} > 2] \cdot \Pr[\mathbf{Y} \neq 6] = \frac{4}{6} \cdot \frac{5}{6}$.

- Conditional probabilities simplify, as with independent events.

  Example: $\Pr[\mathbf{X} = 2 \mid \mathbf{Y} > 5] = \Pr[\mathbf{X} = 2] = \frac{1}{6}$.

- Expected values of products simplify.

  Example: $\mathbb{E}[\mathbf{XY}] = \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}] = 3.5^2$.

- Other joint probabilities may still be tricky.

  Example: $\Pr[\mathbf{X} < \mathbf{Y}]$ doesn't factor into a product of two probabilities, because that wouldn't make any sense.

General joint distributions

If we have two random variables, not necessarily independent, we'd need a grid of probabilities to describe them. For example:

	$\mathbf{Y} = 1$	$\mathbf{Y} = 2$	$\mathbf{Y} = 3$	$\mathbf{Y} = 4$
$\mathbf{X} = 0$	$1/3$	$1/6$	$1/12$	$1/12$
$\mathbf{X} = 10$	$1/6$	$1/12$	$1/24$	$1/24$

- What is the PMF of $\mathbf{X}$? Take the sum of the first row to get $\Pr[\mathbf{X} = 0]$; take the sum of the second row to get $\Pr[\mathbf{X} = 10]$. This gives $P_{\mathbf{X}}(0) = \frac{2}{3}$ and $P_{\mathbf{X}}(10) = \frac{1}{3}$.
- What is the PMF of $\mathbf{Y}$? Take the sums of the columns to get $P_{\mathbf{Y}}(1) = \frac{1}{2}$, $P_{\mathbf{Y}}(2) = \frac{1}{4}$, $P_{\mathbf{Y}}(3) = P_{\mathbf{Y}}(4) = \frac{1}{8}$.
- Are $\mathbf{X}$ and $\mathbf{Y}$ independent? Yes! Check the definition, or that every column is a multiple of the first, or the same for the rows.

Variance of a binomial

Suppose that $\mathbf{X} \sim \textit{Binomial}(n, p)$: we do n trials, with probability p of success on each trial, and let $\mathbf{X}$ be the number of successes.

We've figured out already that $\mathbb{E}[\mathbf{X}] = np$. What is $\mathrm{Var}[\mathbf{X}]$?

Once again, we can write $\mathbf{X}$ as a sum $\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n$ where $\mathbf{X}_i = 1$ if the i^{th} trial succeeds and $\mathbf{X}_i = 0$ if it fails.

Crucially, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent! Therefore

$$\mathrm{Var}[\mathbf{X}] = \mathrm{Var}[\mathbf{X}_1] + \mathrm{Var}[\mathbf{X}_2] + \cdots + \mathrm{Var}[\mathbf{X}_n].$$

We can work out $\mathrm{Var}[\mathbf{X}_i] = \mathbb{E}[\mathbf{X}_i^2] - \mathbb{E}[\mathbf{X}_i]^2 = p - p^2 = p(1-p)$. This gives us $\mathrm{Var}[\mathbf{X}] = np(1-p)$.

Variance of a binomial: what does this tell us?

Suppose that $\mathbf{X} \sim$ *Binomial*(n, p); we know $\mathbb{E}[\mathbf{X}] = np$ and $\mathrm{Var}[\mathbf{X}] = np(1-p)$. What does that mean?

More useful to look at standard deviation: $\mathrm{SD}[\mathbf{X}] = \sqrt{np(1-p)}$.

This tells us that $\mathbf{X}$ deviates from its mean by “around several $\sqrt{n}$’s”.

- From Chebyshev’s inequality (very weak): $\mathbf{X}$ is within k standard deviations of $\mathbb{E}[\mathbf{X}]$ with probability $1 - \frac{1}{k^2}$.
- For binomial random variables where n is large and p is not too close to 0 or 1, much more is true.

  For example, $\mathbf{X}$ is within 3 standard deviations of $\mathbb{E}[\mathbf{X}]$ with probability > 0.997.

Variance of other sums

In general, suppose that we have a “binomial-like” sum:

- $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n$;
- Individually, each $\mathbf{X}_i \sim$ *Bernoulli*(p);
- The $\mathbf{X}_i$ are **not necessarily** independent.

No matter what, we'll get $\mathbb{E}[\mathbf{X}] = np$ by linearity of expectation.

To find the variance, we also have to work out $\operatorname{Cov}[\mathbf{X}_i, \mathbf{X}_j]$ for each pair $i \neq j$.

If $\mathbf{X}_i$ is the indicator variable of an event A_i, then

$$\operatorname{Cov}[\mathbf{X}_i, \mathbf{X}_j] = \mathbb{E}[\mathbf{X}_i\mathbf{X}_j] - \mathbb{E}[\mathbf{X}_i]\mathbb{E}[\mathbf{X}_j] = \Pr[A_i \cap A_j] - \Pr[A_i]\Pr[A_j].$$

Example: hat swap variance

A party has n guests, each with a hat. We take their hats and randomly redistribute them. If $\mathbf{X}_i$ is the indicator variable of event $A_i =$ "the i^{th} guest gets their own hat back", then

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_n$$

is the number of people that get their own hat. We already saw $\Pr[A_i] = \frac{1}{n}$ for all i.

For $i \neq j$, $\Pr[A_i \cap A_j] = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$. Therefore $\operatorname{Cov}[\mathbf{X}_i, \mathbf{X}_j]$ is $\frac{1}{n(n-1)} - \frac{1}{n} \cdot \frac{1}{n} = \frac{1}{n^2(n-1)}$. Next, use the formula

$$\operatorname{Var}[\mathbf{X}] = \sum_{i=1}^{n} \operatorname{Var}[\mathbf{X}_i] + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \operatorname{Cov}[\mathbf{X}_i, \mathbf{X}_j].$$

Hat swap variance

In the formula

$$\operatorname{Var}[\mathbf{X}] = \sum_{i=1}^{n} \operatorname{Var}[\mathbf{X}_i] + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \operatorname{Cov}[\mathbf{X}_i, \mathbf{X}_j]$$

the first sum is just the variance of a binomial: $np(1-p)$. In our case, $p = \frac{1}{n}$ and so this term is $1 - \frac{1}{n}$.

The second sum has $n(n-1)$ $\operatorname{Cov}[\mathbf{X}_i, \mathbf{X}_j]$ terms, and we computed that each one is $\frac{1}{n^2(n-1)}$. Together, they give $\frac{1}{n}$.

So the total variance in the hat swap problem is exactly 1.

Binomial and hypergeometric comparison

Suppose that there are b blue and r red marbles in a bag; we draw k marbles. Let $\mathbf{X}$ be the number of blue marbles.

- If we draw **without** replacement, $\mathbf{X} \sim$ *Hypergeometric*(b, r, k). This is the classic setting for a hypergeometric random variable.
- If we draw **with** replacement, $\mathbf{X} \sim$ *Binomial*$(k, \frac{b}{b+r})$. Only the initial fraction of marbles that are blue matters!
- When b and r are very large compared to k, these are approximately equal.

Question: what sort of behavior should we expect from $\mathrm{Var}[\mathbf{X}]$, in the hypergeometric case?

An initial estimate

Let $\mathbf{X}_i = 1$ if the i^{th} marble is blue, and 0 otherwise. Then $\mathbf{X}$, the number of blue marbles drawn, is $\mathbf{X}_1 + \cdots + \mathbf{X}_k$.

Whether we're sampling with or without replacement, each $\mathbf{X}_i \sim \mathit{Bernoulli}(\frac{b}{b+r})$, so $\mathrm{Var}[\mathbf{X}_i] = \frac{br}{(b+r)^2}$.

- Sampling with replacement: $\mathrm{Cov}[\mathbf{X}_i, \mathbf{X}_j] = 0$. This is the binomial case we already saw, and leads to $\mathrm{Var}[\mathbf{X}] = \frac{brk}{(b+r)^2}$.
- Sampling without replacement: $\mathrm{Cov}[\mathbf{X}_i, \mathbf{X}_j] < 0$. Why?

  Because if $\mathbf{X}_i = 1$, that makes $\mathbf{X}_j = 1$ slightly less likely.

When sampling without replacement, $\mathrm{Var}[\mathbf{X}] < \frac{brk}{(b+r)^2}$.

Covariance of two marbles

To get the exact formula for $\operatorname{Var}[\mathbf{X}]$, we should figure out $\operatorname{Cov}[\mathbf{X}_i, \mathbf{X}_j]$. This is $\Pr[A_i \cap A_j] - \Pr[A_i] \cdot \Pr[A_j]$, where A_i is the event that the i^{th} marble is blue.

- We have $\Pr[A_i] = \Pr[A_j] = \frac{b}{b+r}$: the probability that if you draw one marble, it is blue.
- $\Pr[A_i \cap A_j]$ is the probability that if you draw two marbles, they're both blue. This is
$$\Pr[A_i \cap A_j] = \frac{\binom{b}{2}}{\binom{b+r}{2}} = \frac{b(b-1)}{(b+r)(b+r-1)}.$$

Putting these together, we get $\operatorname{Cov}[\mathbf{X}_i, \mathbf{X}_j] = -\frac{br}{(b+r)^2(b+r-1)}$.

Putting it all together

We have $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_k$; now apply the formula

$$\mathrm{Var}[\mathbf{X}] = \sum_{i=1}^{k} \mathrm{Var}[\mathbf{X}_i] + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} \mathrm{Cov}[\mathbf{X}_i, \mathbf{X}_j].$$

The first term is the “binomial component of the variance”: it is $\frac{brk}{(b+r)^2}$.

The second term has $k(k-1)$ terms that are $-\frac{br}{(b+r)^2(b+r-1)}$ each.

The algebra is messy, but one nice expression for the result is

$$\mathrm{Var}[\mathbf{X}] = \frac{brk}{(b+r)^2} \cdot \frac{b+r-k}{b+r-1}.$$

Definition

The **probability generating function** of a random variable $\mathbf{X}$ is

$$G_{\mathbf{X}}(z) = \mathbb{E}[z^{\mathbf{X}}].$$

When the range of $\mathbf{X}$ is a subset of $\{0, 1, 2, 3, \ldots\}$:

$$G_{\mathbf{X}}(z) = \sum_{k=0}^{\infty} z^k \cdot P_{\mathbf{X}}(z).$$

This collects all the values of $P_{\mathbf{X}}$ into one function.

The textbook defines $M_{\mathbf{X}}(s) = \mathbb{E}[e^{s\mathbf{X}}]$, the **moment generating function**, *instead; this has many of the same properties, and is off by the substitution $z = e^s$, but the PGF will be easier to deal with for us.*

Some PGFs of distributions we know

Suppose $\mathbf{X} \sim \textit{Geometric}(p)$. Then

$$G_{\mathbf{X}}(z) = \sum_{k=1}^{\infty} z^k \cdot p(1-p)^{k-1} = \frac{pz}{1-(1-p)z}$$

by applying the geometric series formula: $a + ar + ar^2 + \cdots = \frac{a}{1-r}$.

Suppose $\mathbf{Y} \sim \textit{Poisson}(\lambda)$. Then

$$G_{\mathbf{Y}}(z) = \sum_{k=0}^{\infty} z^k \cdot e^{-\lambda}\frac{\lambda^k}{k!} = e^{\lambda(z-1)}$$

by applying the formula $1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = e^x$.

Combining PGFs

Fact. Suppose $\mathbf{X}$ and $\mathbf{Y}$ are independent. Then

$$G_{\mathbf{X}+\mathbf{Y}}(z) = G_{\mathbf{X}}(z) \cdot G_{\mathbf{Y}}(z).$$

Proof. If $\mathbf{X}$ and $\mathbf{Y}$ are independent, so are $z^{\mathbf{X}}$ and $z^{\mathbf{Y}}$, therefore $\mathbb{E}[z^{\mathbf{X}+\mathbf{Y}}] = \mathbb{E}[z^{\mathbf{X}} \cdot z^{\mathbf{Y}}] = \mathbb{E}[z^{\mathbf{X}}] \cdot \mathbb{E}[z^{\mathbf{Y}}]$. □

This gives us a quick way to find the PGF of a binomial, which is the sum of independent Bernoulli random variables.

For a *Bernoulli*(p) random variable, the PGF is $z^0 \cdot (1-p) + z^1 \cdot p$.

Therefore if $\mathbf{X} \sim$ *Binomial*(n, p), we get

$$G_{\mathbf{X}}(z) = (1 - p + pz)^n .$$

Getting out probabilities

For finite problems, we can expand $G_{\mathbf{X}}(z)$ and use it to find probabilities.

Example. Let $\mathbf{X}_1 \sim \textit{Binomial}(3, \frac{1}{2})$ and $\mathbf{X}_2 \sim \textit{Binomial}(2, \frac{1}{3})$ be independent, and let $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$. Then

$$G_{\mathbf{X}}(z) = G_{\mathbf{X}_1}(z) \cdot G_{\mathbf{X}_2}(z) = \left(\frac{1}{2} + \frac{1}{2}z\right)^3 \left(\frac{2}{3} + \frac{1}{3}z\right)^2 .$$

Expanding this, we get

$$G_{\mathbf{X}}(z) = \frac{1}{18} + \frac{2}{9}z + \frac{25}{72}z^2 + \frac{19}{72}z^3 + \frac{7}{72}z^4 + \frac{1}{72}z^5$$

so for example $\Pr[\mathbf{X} = 3] = \frac{19}{72}$.

Alternatives to expansion

If the range of $\mathbf{X}$ is infinite, or if $G_{\mathbf{X}}(z)$ is just too hard to expand, we have other options:

- We can take derivatives of $G_{\mathbf{X}}(z)$. That's because the coefficient of $\frac{z^k}{k!}$ in the Taylor series of $G_{\mathbf{X}}(z)$ is the k^{th} derivative $G_{\mathbf{X}}^{(k)}(0)$. In other words,
$$\Pr[\mathbf{X} = k] = \frac{G_{\mathbf{X}}^{(k)}(0)}{k!}.$$
- Computers are great at extracting coefficients; you can use WolframAlpha, for instance.

What are moments?

In general, the expected value of powers of a random variable are called its **moments**.

There are many variants, which we can convert between:

- The **raw moments** of $\mathbf{X}$ are $\mathbb{E}[\mathbf{X}], \mathbb{E}[\mathbf{X}^2], \mathbb{E}[\mathbf{X}^3], \mathbb{E}[\mathbf{X}^4], \ldots$.
- The **central moments** are $\mathbb{E}[(\mathbf{X}-\mu)^2], \mathbb{E}[(\mathbf{X}-\mu)^3], \ldots$.
- The **factorial moments** are $\mathbb{E}[\mathbf{X}(\mathbf{X}-1)]$, $\mathbb{E}[\mathbf{X}(\mathbf{X}-1)(\mathbf{X}-2)]$, $\mathbb{E}[\mathbf{X}(\mathbf{X}-1)(\mathbf{X}-2)(\mathbf{X}-3)], \ldots$.

These measure increasingly hard-to-interpret properties of $\mathbf{X}$. For example, $\mathbb{E}[(\mathbf{X}-\mu)^3]$ is the **skewness** which measures how asymmetric a distribution is.

Moments and the PGF

Once we define $G_{\mathbf{X}}(z) = \mathbb{E}[z^{\mathbf{X}}]$, we get:

- $G'_{\mathbf{X}}(z) = \mathbb{E}[\mathbf{X}z^{\mathbf{X}-1}]$ so $G'_{\mathbf{X}}(1) = \mathbb{E}[\mathbf{X}]$.
- $G''_{\mathbf{X}}(z) = \mathbb{E}[\mathbf{X}(\mathbf{X}-1)z^{\mathbf{X}-2}]$ so $G''_{\mathbf{X}}(1) = \mathbb{E}[\mathbf{X}(\mathbf{X}-1)]$.
- $G^{(k)}_{\mathbf{X}}(1)$ is the k^{th} factorial moment of $\mathbf{X}$.

(Technical note: sometimes, to avoid convergence issues in a formula we get, we need to take a limit $\lim_{z\to 1^-}$ instead of setting $z = 1$.)

We get $\operatorname{Var}[\mathbf{X}] = G''_{\mathbf{X}}(1) + G'_{\mathbf{X}}(1) - [G'_{\mathbf{X}}(1)]^2$, because

$$\operatorname{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2 = \mathbb{E}[\mathbf{X}(\mathbf{X}-1)] + \mathbb{E}[\mathbf{X}] - \mathbb{E}[\mathbf{X}]^2.$$

The geometric distribution

We saw earlier that if $\mathbf{X} \sim \textit{Geometric}(p)$, then $G_{\mathbf{X}}(z) = \frac{pz}{1-z(1-p)}$.

Taking derivatives:

$$G'_{\mathbf{X}}(z) = \frac{p}{(1-z(1-p))^2}$$

$$G''_{\mathbf{X}}(z) = \frac{2p(1-p)}{(1-z(1-p))^3}$$

Therefore $\mathbb{E}[\mathbf{X}] = \frac{p}{(1-(1-p))^2} = \frac{1}{p}$ and $\mathbb{E}[\mathbf{X}(\mathbf{X}-1)] = \frac{2(1-p)}{p^2}$.

If we combine these, we get $\mathrm{Var}[\mathbf{X}] = \frac{1-p}{p^2}$.

The Poisson distribution

If $\mathbf{X} \sim \mathit{Poisson}(\lambda)$, then $G_{\mathbf{X}}(z) = e^{\lambda(z-1)}$.

The derivative of e^{cx} is ce^{cx}, so we get:

$$G'_{\mathbf{X}}(z) = \lambda e^{\lambda(z-1)}$$

$$G''_{\mathbf{X}}(z) = \lambda^2 e^{\lambda(z-1)}$$

$$G^{(k)}_{\mathbf{X}}(z) = \lambda^k e^{\lambda(z-1)}.$$

From here, $\mathbb{E}[\mathbf{X}] = \lambda$, $\mathbb{E}[\mathbf{X}(\mathbf{X}-1)] = \lambda^2$, and $\mathrm{Var}[\mathbf{X}] = \lambda$.

Summary of distributions

Here is a summary of the mean, variance, and PGF of the distributions we know:

Distribution	Mean	Variance	PGF
Bernoulli(p)	p	$p(1-p)$	$1-p+pz$
Binomial(n,p)	np	$np(1-p)$	$(1-p+pz)^n$
Geometric(p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pz}{1-z(1-p)}$
Poisson(λ)	λ	λ	$e^{\lambda(z-1)}$
Hypergeometric(b,r,k)	$\frac{bk}{b+r}$	$\frac{brk}{(b+r)^2} \cdot \frac{b+r-k}{b+r-1}$	skip this
Pascal(m,p)	$\frac{m}{p}$	$m \cdot \frac{1-p}{p^2}$	$\left(\frac{pz}{1-z(1-p)}\right)^m$

Markov's inequality

Theorem (Markov's inequality). If $\mathbf{X}$ is a nonnegative random variable, then $\Pr[\mathbf{X} \geq k] \leq \frac{\mathbb{E}[\mathbf{X}]}{k}$.

Intuitively: if $\mathbf{X}$ is larger than k with probability more than $\frac{x}{k}$, that in itself is enough to make $\mathbb{E}[\mathbf{X}]$ larger than x, even if $\mathbf{X}$ is 0 the rest of the time.

Proof. Let $\mathbf{Y} = 0$ if $\mathbf{X} < k$, and let $\mathbf{Y} = k$ if $\mathbf{X} \geq k$. This satisfies $\mathbf{Y} \leq \mathbf{X}$, so $\mathbb{E}[\mathbf{Y}] \leq \mathbb{E}[\mathbf{X}]$.

By definition, $\mathbb{E}[\mathbf{Y}] = k \cdot \Pr[\mathbf{Y} = k]$. Therefore $\mathbb{E}[\mathbf{X}] \geq k \cdot \Pr[\mathbf{Y} = k]$, or $\mathbb{E}[\mathbf{X}] \geq k \cdot \Pr[\mathbf{X} \geq k]$.

This can be rearranged into Markov's inequality. □

Binomial tail probabilities

Markov's inequality has lots of useful consequences. Our proof of Chebyshev's inequality is essentially applying Markov's inequality to the random variable $(\mathbf{X} - \mu)^2$ instead of $\mathbf{X}$.

Our goal today will be showing that a random variable $\mathbf{X} \sim$ *Binomial*(n, p) is almost always close to its mean np, with better accuracy than Chebyshev's inequality.

To do this, we'll apply Markov's inequality to another function of $\mathbf{X}$: we'll apply it to the function $z^{\mathbf{X}}$ for some carefully chosen $z > 0$.

This will relate $\Pr[\mathbf{X} \geq k]$ to the probability generating function $G_{\mathbf{X}}(z) = \mathbb{E}[z^{\mathbf{X}}]$.

An example

Suppose $\mathbf{X} \sim$ *Binomial*$(n, \frac{1}{2})$. We've seen that the probability generating function of $\mathbf{X}$ is $G_{\mathbf{X}}(z) = \mathbb{E}[z^{\mathbf{X}}] = (\frac{1}{2} + \frac{1}{2}z)^n$.

One special case: $\mathbb{E}[(\frac{27}{8})^{\mathbf{X}}] = (\frac{1}{2} + \frac{27/8}{2})^n = (\frac{35}{16})^n$.

What is $\Pr[\mathbf{X} \geq \frac{2}{3}n]$? Well, it's the same as $\Pr[(\frac{27}{8})^{\mathbf{X}} \geq (\frac{27}{8})^{2n/3}]$, or $\Pr[(\frac{27}{8})^{\mathbf{X}} \geq (\frac{9}{4})^n]$.

By Markov's inequality:

$$\Pr[\mathbf{X} \geq \tfrac{2}{3}n] = \Pr[(\tfrac{27}{8})^{\mathbf{X}} \geq (\tfrac{9}{4})^n] \leq \frac{(\frac{35}{16})^n}{(\frac{9}{4})^n}.$$

This simplifies to $(\frac{35}{36})^n$, which tends to 0 as $n \to \infty$.

The bound we get

As a result: if $\mathbf{X} \sim \mathit{Binomial}(n, \frac{1}{2})$, then

$$\Pr\left[\frac{1}{3}n \leq \mathbf{X} \leq \frac{2}{3}n\right] \geq 1 - 2\left(\frac{35}{36}\right)^n.$$

(To bound $\Pr[\mathbf{X} \leq \frac{1}{3}n]$, do the same thing to $n - \mathbf{X}$.)

This is a much better guarantee than Chebyshev's inequality would give us here. The range from $\frac{1}{3}n$ to $\frac{2}{3}n$ is within $\frac{1}{3}\sqrt{n}$ standard deviations of the mean, so we'd get a bound of $1 - \frac{9}{n}$.

I chose $z = \frac{27}{8} = (\frac{2}{3})^3$ so that the value $\frac{35}{36}$ would come out reasonably nice, and also less than 1.

It would be smarter to pick the value of z that gives the **best** bound for the range we're looking at. (This takes some work.)

Chernoff-type bounds

Bounds of this type are traditionally called “Chernoff-type bounds”, though none of them were proven by Chernoff. Here is one due to Hoeffding:

Hoeffding’s inequality. Let $\mathbf{X} \sim \textit{Binomial}(n, p)$. Then

$$\Pr[\mathbf{X} \geq \mathbb{E}[\mathbf{X}] + t] \leq e^{-2t^2/n}.$$
$$\Pr[\mathbf{X} \leq \mathbb{E}[\mathbf{X}] - t] \leq e^{-2t^2/n}.$$

More generally, this is true whenever $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n$, where $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are mutually independent and always between 0 and 1.

Motivation

One intuition for what “probability” and “expected value” mean is that:

- The probability of an event is the approximate fraction of times it will happen, if we do many trials.
- The expected value of a random variable is the approximate average value it will have, if we do many trials.

Are we guaranteed that sufficiently many trials will actually get us close to the “real probability” or “real expected value”?

This is what laws of large numbers try to tell us.

The weak law of large numbers

Suppose $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots$ is a sequence of **i.i.d. random variables** with a finite expected value $\mathbb{E}[\mathbf{X}_i] = \mu$ and a finite variance $\mathrm{Var}[\mathbf{X}_i] = \sigma^2$.

Let $\overline{\mathbf{X}}_n = \frac{\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n}{n}$. Then for any $\epsilon > 0$,

$$\lim_{n \to \infty} \Pr[|\overline{\mathbf{X}}_n - \mu| \geq \epsilon] = 0.$$

In other words: for any margin of error, if we do enough trials, the probability that our estimated average $\overline{\mathbf{X}}_n$ differs from the "real" average μ by more than the margin of error tends to 0.

To estimate probabilities in this way, take $\mathbf{X}_i \sim$ *Bernoulli*(p). Then $\overline{\mathbf{X}}_n$ is an estimate of p based on what fraction of the first n trials were successes.

No expected value

Suppose each $\mathbf{X}_i = (-4)^{\mathbf{Y}_i}$, where $\mathbf{Y}_i \sim$ *Geometric*$(\frac{1}{2})$. Then $\mathbb{E}[\mathbf{X}_i]$ is undefined: the infinite sum oscillates between large positive and negative values.

Question: What will happen when we look at $\overline{\mathbf{X}}_n = \frac{\mathbf{X}_1 + \cdots + \mathbf{X}_n}{n}$?

Answer: $\overline{\mathbf{X}}_n$ will usually be determined by the most extreme value of $\mathbf{X}_i$ seen so far.

(A bit more detail: usually, when we see a “record-breaking” value $\mathbf{Y}_i = k$, there will not be enough $\mathbf{Y}$-values of $1, 2, \ldots, k-1$ to balance out the $(-4)^k$ we got from $\mathbf{X}_i$. So $\mathbf{X}_i$ will swing the average to a large positive number if k is even, or a large negative number if k is odd.)

This is typical for poorly-behaved random variables.

The strong law of large numbers

The strong law of large numbers, again, applies to a sequence of i.i.d. random variables $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots$, with $\overline{\mathbf{X}}_n = \frac{\mathbf{X}_1+\mathbf{X}_2+\cdots+\mathbf{X}_n}{n}$.

Assuming that the $\mathbf{X}_i$ have mean μ, the strong law of large numbers says that the sequence $\overline{\mathbf{X}}_1, \overline{\mathbf{X}}_2, \overline{\mathbf{X}}_3, \dots$ **is a convergent sequence** with probability 1, and its limit is μ.

Why is this stronger than the weak law? Suppose we're estimating μ by looking at the averages $\overline{\mathbf{X}}_n$ over time.

- Weak law: for any margin of error, we will be within that margin with higher and higher probability as time goes on (tending to 1).
- Strong law: for any margin of error, we will eventually **stay** within that margin forever.

A few words about proofs

The weak law of large numbers can be shown by applying Chebyshev's inequality to $\overline{\mathbf{X}}_n = \frac{\mathbf{X}_1+\mathbf{X}_2+\cdots+\mathbf{X}_n}{n}$. This has expected value μ, but it variance is the decreasing function $\frac{\sigma^2}{n}$.

Chebyshev's inequality is not enough to prove the strong law of large numbers. This requires **exponentially decreasing** probabilities of $\overline{\mathbf{X}}_n$ being outside a range $[\mu - \epsilon, \mu + \epsilon]$.

(Then, even the sum of these probabilities over all large n will be small.)

For the special case of estimating probabilities, Hoeffding's inequality from earlier today can be used to prove the strong law of large numbers.

The finite case

Suppose that $\mathbf{X}$ is 0 or 1 with probability $\frac{1}{2}$ each, and $\mathbf{Y}$ is chosen from $\{1, 2, 3, 4\}$ with probability $\frac{1}{4}$ each. What is $\Pr[\mathbf{X} = \mathbf{Y}]$?

This is not enough information! There are many possibilities.

Example 1: $\mathbf{X}$ and $\mathbf{Y}$ could be independent: each value from $\{0, 1\} \times \{1, 2, 3, 4\}$ is the value of $(\mathbf{X}, \mathbf{Y})$ with probability $\frac{1}{8}$.

Example 2: The probabilities could be given by the following table:

	$\mathbf{Y} = 1$	$\mathbf{Y} = 2$	$\mathbf{Y} = 3$	$\mathbf{Y} = 4$
$\mathbf{X} = 0$	$1/4$		$1/4$	
$\mathbf{X} = 1$		$1/4$		$1/4$

Then, $\Pr[\mathbf{X} = \mathbf{Y}] = 0$.

There are infinitely many other examples.

Marginal and joint distributions

With two random variables in the picture, the distributions of $\mathbf{X}$ and $\mathbf{Y}$ are called **marginal distributions**. They let us fill in the margins of the table below:

	$\mathbf{Y} = 1$	$\mathbf{Y} = 2$	$\mathbf{Y} = 3$	$\mathbf{Y} = 4$	Total
$\mathbf{X} = 0$					$1/2$
$\mathbf{X} = 1$					$1/2$
Total	$1/4$	$1/4$	$1/4$	$1/4$	1

To know the relationship between $\mathbf{X}$ and $\mathbf{Y}$, we need to know the entries in the middle: their **joint distribution**. This is commonly given by the **joint PMF**:

$$P_{\mathbf{XY}}(a, b) = \Pr[\mathbf{X} = a \text{ and } \mathbf{Y} = b].$$

Random variables with infinite range

The same thing happens when $\mathbf{X}$ and $\mathbf{Y}$ have infinite range, but we can't draw a table. The joint PMF still exists, but feels more abstract.

Here are three examples of random experiments that produce variously related $\mathbf{X}, \mathbf{Y} \sim \textit{Geometric}(\frac{1}{6})$.

1. $\mathbf{X}$ and $\mathbf{Y}$ could be independent. Xavier and Yvonne each roll a die until rolling a 6. $\mathbf{X}$ counts Xavier's rolls, and $\mathbf{Y}$ counts Yvonne's rolls.
2. $\mathbf{X}$ could entirely determine $\mathbf{Y}$ (and vice versa). Zsuzsa rolls a die until rolling a 6. $\mathbf{X}$ counts Zsuzsa's rolls, and $\mathbf{Y} = \mathbf{X}$.
3. Something more complicated. Wilhelm rolls a die until rolling both a 1 and a 6. $\mathbf{X}$ counts the number of rolls until the 1, and $\mathbf{Y}$ counts the number of rolls until the 6.

Formulas for joint distributions

If $\mathbf{X}, \mathbf{Y} \sim \textit{Geometric}(\frac{1}{6})$, then $P_{\mathbf{X}}(k) = P_{\mathbf{Y}}(k) = (\frac{1}{6})(\frac{5}{6})^{k-1}$.

1. If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then

$$P_{\mathbf{XY}}(a,b) = P_{\mathbf{X}}(a)P_{\mathbf{Y}}(b) = (\tfrac{1}{6})(\tfrac{5}{6})^{a-1}(\tfrac{1}{6})(\tfrac{5}{6})^{b-1}.$$

2. If $\mathbf{X} = \mathbf{Y}$, then

$$P_{\mathbf{XY}}(a,b) = \begin{cases} (\frac{1}{6})(\frac{5}{6})^{a-1} & a = b \\ 0 & a \neq b. \end{cases}$$

3. In the third case, we get

$$P_{\mathbf{XY}}(a,b) = \begin{cases} (\frac{4}{6})^{a-1}(\frac{1}{6})(\frac{5}{6})^{b-a-1}(\frac{1}{6}) & a < b \\ (\frac{4}{6})^{b-1}(\frac{1}{6})(\frac{5}{6})^{a-b-1}(\frac{1}{6}) & a > b \\ 0 & a = b. \end{cases}$$

Functions of two random variables

Given the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$, we often want to define a third random variable $\mathbf{Z} = g(\mathbf{X}, \mathbf{Y})$ for some function $g : R_{\mathbf{X}} \times R_{\mathbf{Y}} \to \mathbb{R}$.

Examples: the sum $\mathbf{X} + \mathbf{Y}$, the maximum $\max\{\mathbf{X}, \mathbf{Y}\}$, the product $\mathbf{XY}$, and more.

What is $P_{\mathbf{Z}}(k)$? Two-step process:

1. Let $A_k = \{(a, b) \in R_{\mathbf{X}} \times R_{\mathbf{Y}} : g(a, b) = k\}$.

   These are all the ways we can get $g(\mathbf{X}, \mathbf{Y}) = k$.

2. Compute the PMF of $\mathbf{Z}$ by the formula

$$P_{\mathbf{Z}}(k) = \sum_{(a,b)\in A_k} P_{\mathbf{XY}}(a, b).$$

Example 1: a finite example

Suppose the joint PMF of $\mathbf{X}$ and $\mathbf{Y}$ is given by the following table:

	$\mathbf{Y}=0$	$\mathbf{Y}=1$	$\mathbf{Y}=2$
$\mathbf{X}=0$	0.04	0.16	0.20
$\mathbf{X}=1$	0.02	0.06	0.12
$\mathbf{X}=2$	0.14	0.08	0.18

What is the distribution of the product $\mathbf{Z}=\mathbf{X}\cdot\mathbf{Y}$?

1. We split $R_{\mathbf{X}} \times R_{\mathbf{Y}}$ into $A_0 = \{(0,0),(0,1),(0,2),(1,0),(2,0)\}$, $A_1 = \{(1,1)\}$, $A_2 = \{(1,2),(2,1)\}$, and $A_4 = \{(2,2)\}$.
2. We sum over these sets to get $P_{\mathbf{Z}}(k)$. For example, $P_{\mathbf{Z}}(2) = P_{\mathbf{XY}}(1,2) + P_{\mathbf{XY}}(2,1) = 0.12 + 0.08 = 0.2$.

Example 2: Sum of two geometrics

Suppose $\mathbf{X}, \mathbf{Y} \sim$ *Geometric*$(\frac{1}{6})$, and $\mathbf{X}$ and $\mathbf{Y}$ are independent. What is the distribution of $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$?

1. What is A_k, the set of all ways $\mathbf{X} + \mathbf{Y}$ can equal k? It is

$$A_k = \{(a, k-a) : 1 \leq a \leq k-1\}.$$

   For example, $A_5 = \{(1,4),(2,3),(3,2),(4,1)\}$.

2. What is $P_{\mathbf{Z}}(k)$? We take the sum

$$P_{\mathbf{Z}}(k) = \sum_{a=1}^{k-1} P_{\mathbf{XY}}(a, k-a) = \sum_{a=1}^{k-1} (\tfrac{1}{6})(\tfrac{5}{6})^{a-1}(\tfrac{1}{6})(\tfrac{5}{6})^{k-a-1}.$$

Simplifying, $P_{\mathbf{Z}}(k) = (k-1)(\frac{1}{6})^2(\frac{5}{6})^{k-2}$, the PMF of a *Pascal*$(2, \frac{1}{6})$.

Example 3: Min of two geometrics

Suppose $\mathbf{X}, \mathbf{Y} \sim$ *Geometric*$(\frac{1}{6})$, and $\mathbf{X}$ and $\mathbf{Y}$ are independent. What is the distribution of $\mathbf{Z} = \min\{\mathbf{X}, \mathbf{Y}\}$?

1. What is A_k, the set of all ways $\min\{\mathbf{X}, \mathbf{Y}\}$ can equal k? It is

$$A_k = \{(k,k)\} \cup \{(a,k) : a \geq k+1\} \cup \{(k,b) : b \geq k+1\}.$$

2. What is $P_{\mathbf{Z}}(k)$? We take the sum

$$P_{\mathbf{Z}}(k) = P_{\mathbf{XY}}(k,k) + \sum_{a=k+1}^{\infty} P_{\mathbf{XY}}(a,k) + \sum_{b=k+1}^{\infty} P_{\mathbf{XY}}(k,b)$$

$$= (\tfrac{1}{6})(\tfrac{5}{6})^{k-1}(\tfrac{1}{6})(\tfrac{5}{6})^{k-1} + 2\sum_{i=k+1}^{\infty} (\tfrac{1}{6})(\tfrac{5}{6})^{k-1}(\tfrac{1}{6})(\tfrac{5}{6})^{i-1}.$$

Exercise: simplify to $(\frac{11}{36})(\frac{25}{36})^{k-1}$; conclude $\mathbf{Z} \sim$ *Geometric*$(\frac{11}{36})$.

Example 4: $\mathbf{X}$ and $\mathbf{Y}$ aren't independent

Suppose $\mathbf{X}, \mathbf{Y} \sim$ *Geometric*$(\frac{1}{6})$, but the joint distribution is that $\mathbf{X} = \mathbf{Y}$ always. What is the distribution of $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$?

1. The set $A_k = \{(a, k-a) : 1 \leq a \leq k-1\}$ is still the same.
2. However, the joint distribution is different!

   When k is even, $(\frac{k}{2}, \frac{k}{2}) \in A_k$ is the only one with a nonzero probability: $P_{\mathbf{XY}}(\frac{k}{2}, \frac{k}{2}) = (\frac{1}{6})(\frac{5}{6})^{k/2-1}$.

   When k is odd, no ordered pair in A_k has nonzero probability.

We get

$$P_{\mathbf{Z}}(k) = \begin{cases} (\frac{1}{6})(\frac{5}{6})^{k/2-1} & k \text{ is even,} \\ 0 & k \text{ is odd.} \end{cases}$$

Conditional probability and random variables

We've already seen conditional probability:

$$\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

where $\Pr[A \mid B]$ has the interpretation "probability that A happens, given that we know B happens".

There's no reason not to apply this to random variables. For example, we can say "What is $\Pr[\mathbf{X} = 0 \mid \mathbf{X} \leq 5]$?" and the same definition works just fine.

But if we want to ask many questions about $\mathbf{X}$ conditioning on the same event, we should define a **conditional distribution**.

Conditional distributions

If $\mathbf{X}$ is a random variable and A is an event, $\mathbf{X} \mid A$ is another random variable. Its distribution is the "conditional distribution of $\mathbf{X}$ given A".

- **The intuition:** $\mathbf{X} \mid A$ is the same random quantity as $\mathbf{X}$, but we have more information (we know that A occurred).
- **Example:** Say $\mathbf{X} \sim \mathit{Geometric}(\frac{1}{2})$ (for example, number of coinflips until the coin lands heads) and A is "$\mathbf{X} \leq 10$".

  Then $\mathbf{X} \mid A$ is the number of coinflips until the coin lands heads, given that it took at most 10 coinflips.
- **Formal definition:** When we condition on A, we just restrict the sample space S to the set A. $\mathbf{X}$ is a function $S \to R_{\mathbf{X}}$, and $\mathbf{X} \mid A$ is just the function $A \to R_{\mathbf{X}}$ with the same values for every outcome in A.

An example using the definition

We flip 10 fair coins. $\mathbf{X} \sim$ *Binomial*$(5, \frac{1}{2})$ is the number of heads in the first **five** flips. Meanwhile, A is the event "Out of all ten flips, six were heads" (with $\Pr[A] = \binom{10}{6}(\frac{1}{2})^{10}$).

For each k, $\Pr[\mathbf{X} = k \mid A]$ requires finding $\Pr[\mathbf{X} = k \text{ and } A]$. If "$\mathbf{X} = k$ and A" happens, then k of the first 5 flips and $6 - k$ of the last five flips are heads.

There are $\binom{5}{k}\binom{5}{6-k}$ ways to choose the heads, and $(\frac{1}{2})^{10}$ chance to see the resulting sequence, so

$$\Pr[\mathbf{X} = k \mid A] = \frac{\Pr[\mathbf{X} = k \text{ and } A]}{\Pr[A]} = \frac{\binom{5}{k}\binom{5}{6-k}(\frac{1}{2})^{10}}{\binom{10}{6}(\frac{1}{2})^{10}} = \frac{\binom{5}{k}\binom{5}{6-k}}{\binom{10}{6}}.$$

We realize that in this case, $\mathbf{X} \mid A \sim$ *Hypergeometric*$(6, 4, 5)$.

The one-variable case

The PMF of $\mathbf{X} \mid A$ is defined by the rule:

$$P_{\mathbf{X}|A}(k) = \Pr[\mathbf{X} = k \mid A] = \frac{\Pr[\mathbf{X} = k \text{ and } A]}{\Pr[A]}.$$

In the simplest case, A is an event of the form "$\mathbf{X} \in B$" for some subset $B \subseteq R_{\mathbf{X}}$. For example, A could be "$\mathbf{X} > 1$" or "$\mathbf{X}$ is even".

$\Pr[\mathbf{X} = k \text{ and } \mathbf{X} \in B]$ is $\Pr[\mathbf{X} = k]$ if $k \in B$, and 0 if $k \notin B$.

So to describe $\mathbf{X} \mid \mathbf{X} \in B$, we:

1. Set the probabilities of every value outside B to 0.
2. Divide all probabilities by $\Pr[\mathbf{X} \in B]$. (Rescale them so that they add to 1 again.)

Example: Binomial conditioning problem

Suppose that $\mathbf{X} \sim \textit{Binomial}(5, \frac{1}{2})$: its PMF is given by

$$\begin{array}{lll} P_{\mathbf{X}}(0) = \dfrac{1}{32} & P_{\mathbf{X}}(1) = \dfrac{5}{32} & P_{\mathbf{X}}(2) = \dfrac{10}{32} \\ P_{\mathbf{X}}(3) = \dfrac{10}{32} & P_{\mathbf{X}}(4) = \dfrac{5}{32} & P_{\mathbf{X}}(5) = \dfrac{1}{32}. \end{array}$$

Now, condition on the event “$\mathbf{X} \geq 2$”.

The range of $\mathbf{X}$ was $\{0, 1, 2, 3, 4, 5\}$; the range of $\mathbf{X} \mid \mathbf{X} \geq 2$ is just $\{2, 3, 4, 5\}$. The PMF is given by

$$\begin{array}{lll} & & P_{\mathbf{X}|\mathbf{X}\geq 2}(2) = \dfrac{10}{26} \\ P_{\mathbf{X}|\mathbf{X}\geq 2}(2) = \dfrac{10}{26} & P_{\mathbf{X}|\mathbf{X}\geq 2}(2) = \dfrac{5}{26} & P_{\mathbf{X}|\mathbf{X}\geq 2}(2) = \dfrac{1}{26} \end{array}$$

Two-variable conditioning

Suppose that we have two random variables $\mathbf{X}, \mathbf{Y}$. Then we can ask:

- What is the distribution of $\mathbf{X} \mid \mathbf{Y} = k$?
- If A is some more complicated event, like $\mathbf{X} > \mathbf{Y}$, what is the distribution of $\mathbf{X} \mid A$?
- What is the joint distribution of $\mathbf{X} \mid A$ and $\mathbf{Y} \mid A$ in such a case?

We can always fall back on the definition of $\mathbf{X} \mid A$, which comes down to the definition of conditional probability.

We are only looking at this case differently because there are easier ways to approach the problem.

A finite joint distribution

Suppose $\mathbf{X}, \mathbf{Y} \sim \textit{Binomial}(2, \frac{1}{2})$ and are independent. Their joint PMF is given below:

	$\mathbf{Y} = 0$	$\mathbf{Y} = 1$	$\mathbf{Y} = 2$
$\mathbf{X} = 0$	$1/16$	$1/8$	$1/16$
$\mathbf{X} = 1$	$1/8$	$1/4$	$1/8$
$\mathbf{X} = 2$	$1/16$	$1/8$	$1/16$

	$\mathbf{Y} = 0$	$\mathbf{Y} = 1$	$\mathbf{Y} = 2$
$\mathbf{X} = 0$	0	0	0
$\mathbf{X} = 1$	$2/5$	0	0
$\mathbf{X} = 2$	$1/5$	$2/5$	0

With the joint distribution in place, conditioning on an event like $\mathbf{X} > \mathbf{Y}$ is the same as the 1-variable case.

1. Set the PMF to 0 when the event does not hold.
2. Scale the remaining values to sum to 1.

Example: rolling two dice

We roll two dice; let $\mathbf{X}$ be the value of the first roll. What is the distribution of $\mathbf{X}$, given that the total is at least 10?

1. Let $\mathbf{X}, \mathbf{Y}$ to be the values of the **two** dice: uniformly distributed on $\{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$.

   We are conditioning on A which is "$\mathbf{X} + \mathbf{Y} \geq 10$".

2. The outcomes $\{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}$ are the only ones compatible with A.

   In the conditional distribution, all of them have probability $\frac{1}{6}$.

3. The marginal distribution $\mathbf{X} \mid A$ is given by

$$P_{\mathbf{X}|A}(4) = \frac{1}{6} \quad P_{\mathbf{X}|A}(5) = \frac{2}{6} \quad P_{\mathbf{X}|A}(6) = \frac{3}{6}.$$

Conditional PMF

A special case: the conditional distribution $\mathbf{X} \mid \mathbf{Y} = k$. Here, suppose we have the joint PMF:

	$\mathbf{Y} = 0$	$\mathbf{Y} = 1$	$\mathbf{Y} = 2$	$\mathbf{Y} = 3$	$\mathbf{Y} = 4$
$\mathbf{X} = 0$	0.1	0.1	0.1	0	0
$\mathbf{X} = 1$	0	0.1	0.2	0.1	0
$\mathbf{X} = 2$	0	0	0.1	0.1	0.1

Conditioning on $\mathbf{Y} = 2$ for instance means just taking the $\mathbf{Y} = 2$ column of the table.

We have probabilities 0.1, 0.2, 0.1 which we scale up to $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$.

Special notation: we write $P_{\mathbf{X}|\mathbf{Y}}(a \mid b)$ for $P_{\mathbf{X}|\mathbf{Y}=b}(a)$, which is another way to say $\Pr[\mathbf{X} = a \mid \mathbf{Y} = b]$.

The fundamentals of conditional expectation

Expected value is one of the main reasons to introduce random variables. So the first thing we do when dealing with conditional distributions like $\mathbf{X} \mid A$ is to compute their expected values $\mathbb{E}[\mathbf{X} \mid A]$.

There are some extra tricks, but the straightforward way to do it is to find the distribution of $\mathbf{X} \mid A$ first, then find its expectation.

Example. If $\mathbf{X} \sim \textit{Geometric}(\frac{1}{3})$, find $\mathbb{E}[\mathbf{X} \mid \mathbf{X} \leq 3]$.

1. Find $P_{\mathbf{X}}(1) = \frac{1}{3} = \frac{9}{27}$, $P_{\mathbf{X}}(2) = \frac{1}{3}(\frac{2}{3}) = \frac{6}{27}$, $P_{\mathbf{X}}(3) = \frac{1}{3}(\frac{2}{3})^2 = \frac{4}{27}$.
2. Rescale: $P_{\mathbf{X}|\mathbf{X}\leq 3}(1) = \frac{9}{19}$, $P_{\mathbf{X}|\mathbf{X}\leq 3}(2) = \frac{6}{19}$, and $P_{\mathbf{X}|\mathbf{X}\leq 3}(3) = \frac{4}{19}$.
3. Compute $\mathbb{E}[\mathbf{X} \mid \mathbf{X} \leq 3] = 1 \cdot \frac{9}{19} + 2 \cdot \frac{6}{19} + 3 \cdot \frac{4}{19} = \frac{33}{19}$.

Two-variable example

Suppose $\mathbf{X}, \mathbf{Y} \sim$ *Geometric*$(\frac{1}{2})$ and are independent. What is $\mathbb{E}[\mathbf{X} \mid \mathbf{X} + \mathbf{Y} = 6]$?

1. The possible values of $(\mathbf{X}, \mathbf{Y})$ consistent with $\mathbf{X} + \mathbf{Y} = 6$ are $A_6 = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$.
2. The joint PMF of $\mathbf{X}$ and $\mathbf{Y}$ is $P_{\mathbf{XY}}(a,b) = (\frac{1}{2})^a(\frac{1}{2})^b = (\frac{1}{2})^{a+b}$. For each of the pairs in A_6, this is the same value: $(\frac{1}{2})^6$.
3. So the conditional distribution of $(\mathbf{X}, \mathbf{Y})$ given $\mathbf{X} + \mathbf{Y} = 6$ assigns each of $\{(1,5), (2,4), (3,3), (4,2), (5,1)\}$ the same value: $\frac{1}{5}$.
4. Therefore $\mathbb{E}[\mathbf{X} \mid \mathbf{X} + \mathbf{Y} = 6]$ is $\frac{1+2+3+4+5}{5} = 3$.

Law of total expectation

Theorem. If $B_1, B_2, \ldots, B_n$ are events forming a partition of the sample space, and $\mathbf{X}$ is a random variable, then

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} \mathbb{E}[\mathbf{X} \mid B_i] \Pr[B_i].$$

Proof. By the law of total probability, for any $k \in R_{\mathbf{X}}$,

$$P_{\mathbf{X}}(k) = \sum_{i=1}^{n} P_{\mathbf{X}|B_i}(k) \Pr[B_i].$$

Then we can multiply by k to get

$$k \cdot P_{\mathbf{X}}(k) = \sum_{i=1}^{n} (k \cdot P_{\mathbf{X}|B_i}(k)) \Pr[B_i].$$

Summing over all $k \in R_{\mathbf{X}}$, we get the law of total expectation. □

Application: a shortcut for conditional expectation

Let $\mathbf{X} \sim \mathit{Binomial}(10, \frac{1}{2})$. What is $\mathbb{E}[\mathbf{X} \mid \mathbf{X} > 1]$?

If we reason directly, we'd first have to figure out the distribution of $\mathbf{X} \mid \mathbf{X} > 1$, which is difficult. Instead, let's find $\mathbb{E}[\mathbf{X} \mid \mathbf{X} \leq 1]$.

Since $P_{\mathbf{X}}(0) = (\frac{1}{2})^{10}$ and $P_{\mathbf{X}}(1) = 10(\frac{1}{2})^{10}$, the conditioned random variable $\mathbf{X} \mid \mathbf{X} \leq 1$ is 0 with probability $\frac{1}{11}$ and 1 with probability $\frac{10}{11}$. So $\mathbb{E}[\mathbf{X} \mid \mathbf{X} \leq 1] = \frac{10}{11}$.

By the law of total expectation,

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{X} \mid \mathbf{X} \leq 1] \Pr[\mathbf{X} \leq 1] + \mathbb{E}[\mathbf{X} \mid \mathbf{X} > 1] \Pr[\mathbf{X} > 1].$$

We know most of these; in particular, $\Pr[\mathbf{X} \leq 1]$ is $\frac{11}{2^{10}}$. We get

$$5 = \frac{10}{11}\left(\frac{11}{2^{10}}\right) + x\left(1 - \frac{11}{2^{10}}\right) \implies x = \frac{5 - 10/2^{10}}{1 - 11/2^{10}} \approx 5.044.$$

Application: fair gambling

You start with $\$10$, and you go to the casino. The casino, unusually, lets you place fair bets that have zero expected profit (with no cut for the house). You play until either you have $\$100$ or you're broke.

What is your probability of making a profit?

It seems like we need to know what kind of bets you place and then do lots of work. But we don't!

Let A be the event "you reach $\$100$ before going broke". Then

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{X} \mid A] \Pr[A] + \mathbb{E}[\mathbf{X} \mid A^c] \Pr[A^c].$$

By fairness, $\mathbb{E}[\mathbf{X}] = 10$. Meanwhile, $\mathbb{E}[\mathbf{X} \mid A] = 100$ (that's what A means) and $\mathbb{E}[\mathbf{X} \mid A^c] = 0$. So we get

$$10 = 100 \cdot \Pr[A] + 0 \cdot (1 - \Pr[A]) \implies \Pr[A] = \tfrac{1}{10}.$$

The meaning of $\mathbb{E}[\mathbf{X} \mid \mathbf{Y}]$

A common kind of conditional expectation is $\mathbb{E}[\mathbf{X} \mid \mathbf{Y} = k]$: the expected value of $\mathbf{X}$ under the assumption that $\mathbf{Y} = k$.

Building off of that, $\mathbb{E}[\mathbf{X} \mid \mathbf{Y}]$ is a more complicated object.

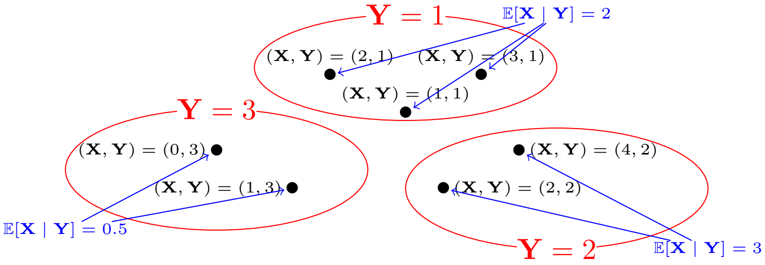
It is another random variable—in fact, it is a function of $\mathbf{Y}$. It is defined as follows:

> In the outcomes where $\mathbf{Y} = k$, $\mathbb{E}[\mathbf{X} \mid \mathbf{Y}]$ is equal to $\mathbb{E}[\mathbf{X} \mid \mathbf{Y} = k]$.

As a special case, if $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $\mathbf{X} \mid \mathbf{Y} = k$ is just $\mathbf{X}$ for any possible k. Therefore $\mathbb{E}[\mathbf{X} \mid \mathbf{Y}] = \mathbb{E}[\mathbf{X}]$. (It is still technically a random variable, but it is a random variable that always has the same value.)

An intuition-building picture

To help intuition, suppose that our sample space is the dots below (a dot is chosen uniformly at random). $\mathbf{X}$ and $\mathbf{Y}$ have the values specified at each dot.


$\mathbf{Y} = 1$
$\mathbb{E}[\mathbf{X} \mid \mathbf{Y}] = 2$
$(\mathbf{X}, \mathbf{Y}) = (2, 1)$
$(\mathbf{X}, \mathbf{Y}) = (3, 1)$
$(\mathbf{X}, \mathbf{Y}) = (1, 1)$
$\mathbf{Y} = 3$
$(\mathbf{X}, \mathbf{Y}) = (0, 3)$
$(\mathbf{X}, \mathbf{Y}) = (1, 3)$
$\mathbb{E}[\mathbf{X} \mid \mathbf{Y}] = 0.5$
$(\mathbf{X}, \mathbf{Y}) = (4, 2)$
$(\mathbf{X}, \mathbf{Y}) = (2, 2)$
$\mathbf{Y} = 2$
$\mathbb{E}[\mathbf{X} \mid \mathbf{Y}] = 3$

First, partition the sample space according to the value of $\mathbf{Y}$.

Then, set $\mathbb{E}[\mathbf{X} \mid \mathbf{Y}]$ to be the expected value of $\mathbf{X}$ in each part.

Application: two-stage experiments

This weird $\mathbb{E}[\mathbf{X} \mid \mathbf{Y}]$ may be easier to think about in two-stage experiments, where we:

1. Sample $\mathbf{Y}$ from a known distribution.
2. Sample $\mathbf{X}$ from a distribution that depends on $\mathbf{Y}$.

(We'll see more of these in the next lecture.)

Example. We choose $\mathbf{P}$ uniformly at random from $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}\}$. Then, choose $\mathbf{X} \sim \textit{Binomial}(10, \mathbf{P})$.

We say $\mathbb{E}[\mathbf{X} \mid \mathbf{P}] = 10\mathbf{P}$ as a shorthand for "$\mathbb{E}[\mathbf{X} \mid \mathbf{P} = p] = 10p$" or "If $\mathbf{P} = p$ in the first step, the expected value of $\mathbf{X}$ from the second step is $10p$."

Law of iterated expectation

If $R_{\mathbf{Y}} = \{y_1, y_2, \dots, y_n\}$, then the events $\mathbf{Y} = y_1, \dots, \mathbf{Y} = y_n$ are a partition of the sample space. So by the law of total expectation,

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} \mathbb{E}[\mathbf{X} \mid \mathbf{Y} = y_i] \Pr[\mathbf{Y} = y_i].$$

But remember: $\mathbb{E}[\mathbf{X} \mid \mathbf{Y}]$ is a random variable that's equal to $\mathbb{E}[\mathbf{X} \mid \mathbf{Y} = y_i]$ when $\mathbf{Y} = y_i$, which happens with probability $\Pr[\mathbf{Y} = y_i]$.

Therefore $\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbb{E}[\mathbf{X} \mid \mathbf{Y}]]$.

This is called the "law of iterated expectation".

Bayes' rule, again

Everything we do today comes down to the definition of conditional probability and to Bayes' rule:

$$\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[B \mid A] \Pr[A]}{\Pr[B]}.$$

In particular, say that $\mathbf{X}$ and $\mathbf{Y}$ are discrete random variables, A is the event $\mathbf{X} = a$, and B is the event $\mathbf{Y} = b$. We've already seen the notation

$$P_{\mathbf{X}|\mathbf{Y}}(a, b) = \Pr[\mathbf{X} = a \mid \mathbf{Y} = b].$$

With it, we can write the formula for $\Pr[A \mid B]$ very concisely:

$$P_{\mathbf{X}|\mathbf{Y}}(a, b) = \frac{P_{\mathbf{X}\mathbf{Y}}(a, b)}{P_{\mathbf{Y}}(b)} = \frac{P_{\mathbf{Y}|\mathbf{X}}(b, a) P_{\mathbf{X}}(a)}{P_{\mathbf{Y}}(b)}.$$

When to use it

Bayes' rule is the right tool for the job when it's easier to find $P_{\mathbf{Y}|\mathbf{X}}$, but it's more useful to know $P_{\mathbf{X}|\mathbf{Y}}$. For example:

1. We know or assume a distribution for $\mathbf{X}$, but we don't get to observe $\mathbf{X}$ directly.
2. We know the conditional distribution $P_{\mathbf{Y}|\mathbf{X}}$: each value of $\mathbf{X}$ is a "hypothesis" under which $\mathbf{Y}$ is easier to predict.
3. Our experiment lets observe the value of $\mathbf{Y}$, and we'd like to use it to predict $\mathbf{X}$.

In a setup like this, we can use Bayes' rule to flip things around and try to find $P_{\mathbf{X}|\mathbf{Y}}$: this will tell us the probability distribution of $\mathbf{X}$, given the information we observed.

A finite example: higher roll

We roll two fair dice—just for the sake of making tables easier, let's say they're 3-sided dice. Let $\mathbf{X}_1, \mathbf{X}_2$ be the two values rolled, and let $\mathbf{Y} = \max\{\mathbf{X}_1, \mathbf{X}_2\}$.

Our goal: given the value of $\mathbf{Y}$, make a guess about $\mathbf{X}_1$.

But we'll start with the reverse table, because going from $\mathbf{X}_1$ to $\mathbf{Y}$ is easier:

$P_{\mathbf{Y}\mid\mathbf{X}_1}$	$\mathbf{Y} = 1$	$\mathbf{Y} = 2$	$\mathbf{Y} = 3$
$\mathbf{X}_1 = 1$	$1/3$	$1/3$	$1/3$
$\mathbf{X}_1 = 2$	0	$2/3$	$1/3$
$\mathbf{X}_1 = 3$	0	0	1

Each row in this table represents the distribution of $\mathbf{Y}$, given some particular value of $\mathbf{X}_1$.

Solving the problem

In this problem, $P_{\mathbf{X}_1}(1) = P_{\mathbf{X}_1}(2) = P_{\mathbf{X}_1}(3) = \frac{1}{3}$. So we can multiply the $P_{\mathbf{Y}|\mathbf{X}_1}$ table by $\frac{1}{3}$ to get the $P_{\mathbf{X}_1\mathbf{Y}}$ joint PMF table:

$P_{\mathbf{Y}\|\mathbf{X}_1}$	$\mathbf{Y}{=}1$	$\mathbf{Y}{=}2$	$\mathbf{Y}{=}3$
$\mathbf{X}_1{=}1$	$1/3$	$1/3$	$1/3$
$\mathbf{X}_1{=}2$	0	$2/3$	$1/3$
$\mathbf{X}_1{=}3$	0	0	1

$P_{\mathbf{X}_1\mathbf{Y}}$	$\mathbf{Y}{=}1$	$\mathbf{Y}{=}2$	$\mathbf{Y}{=}3$
$\mathbf{X}_1{=}1$	$1/9$	$1/9$	$1/9$
$\mathbf{X}_1{=}2$	0	$2/9$	$1/9$
$\mathbf{X}_1{=}3$	0	0	$1/3$

This is the numerator of Bayes' rule. Finally, divide by the denominator $P_{\mathbf{Y}}$: the marginal probability, found by summing the columns.

$P_{\mathbf{X}_1\|\mathbf{Y}}$	$\mathbf{Y}{=}1$	$\mathbf{Y}{=}2$	$\mathbf{Y}{=}3$
$\mathbf{X}_1{=}1$	1	$1/3$	$1/5$
$\mathbf{X}_1{=}2$	0	$2/3$	$1/5$
$\mathbf{X}_1{=}3$	0	0	$3/5$

An extended example

We will look at the following random experiment:

1. Flip n fair coins, and let $\mathbf{X}$ be the number of heads.
2. Re-flip all $\mathbf{X}$ coins that landed heads, and let $\mathbf{Y}$ be the number of heads among the newly flipped coins.

We have $\mathbf{X} \sim \mathit{Binomial}(n, \frac{1}{2})$ and $\mathbf{Y} \mid \mathbf{X} \sim \mathit{Binomial}(\mathbf{X}, \frac{1}{2})$.

(In other words, $\mathbf{Y} \mid \mathbf{X} = a \sim \mathit{Binomial}(a, \frac{1}{2})$ for each a.)

In formulas:

$$P_{\mathbf{X}}(a) = \binom{n}{a}\left(\frac{1}{2}\right)^n \quad P_{\mathbf{Y}|\mathbf{X}}(b, a) = \binom{a}{b}\left(\frac{1}{2}\right)^a$$

When $\mathbf{Y} = 0$

What is the distribution of the intermediate headcount $\mathbf{X}$ when at the end, we see $\mathbf{Y} = 0$?

In applying Bayes' rule, we first find the joint PMF $P_{\mathbf{XY}}(a, b)$ by multiplying together $P_{\mathbf{X}}$ and $P_{\mathbf{Y}|\mathbf{X}}$:

$$P_{\mathbf{XY}}(a, b) = \binom{n}{a}\left(\frac{1}{2}\right)^{n}\binom{a}{b}\left(\frac{1}{2}\right)^{a}.$$

In particular, setting $b = 0$, we get

$$P_{\mathbf{XY}}(a, 0) = \binom{n}{a}\left(\frac{1}{2}\right)^{n+a}.$$

Our goal is to find $P_{\mathbf{X}|\mathbf{Y}}(a, 0)$. To get it, we must divide by $P_{\mathbf{Y}}(0)$.

How to find the denominator

$P_{\mathbf{Y}}(0)$ is the marginal probability: the sum $\displaystyle\sum_{k=0}^{n} P_{\mathbf{XY}}(k,0)$.

In other words, we need to "scale down" $\binom{n}{a}\left(\frac{1}{2}\right)^{n+a}$ by a constant independent of a, so that the sum adds up to 1.

In this problem, we have a shortcut. I claim that $\mathbf{Y} \sim$ *Binomial*$(n, \frac{1}{4})$ because in order for a coin to be counted by $\mathbf{Y}$, it has to land heads twice in a row. Therefore we must divide by $P_{\mathbf{Y}}(0) = (\frac{3}{4})^n$.

We get

$$P_{\mathbf{X}|\mathbf{Y}=0}(a) = \binom{n}{a}\left(\frac{1}{2}\right)^{n+a} \Big/ \left(\frac{3}{4}\right)^{n} = \binom{n}{a}\left(\frac{1}{3}\right)^{a}\left(\frac{2}{3}\right)^{n-a}$$

concluding that $\mathbf{X} \mid \mathbf{Y} = 0 \sim$ *Binomial*$(n, \frac{1}{3})$.

(Optional) Finding the scaling factor the hard way

Without noticing the shortcut, what would we have done?

We know that we want $P_{\mathbf{X}|\mathbf{Y}=0}(a)$ to be proportional to $\binom{n}{a}\left(\frac{1}{2}\right)^{n+a}$.

$$P_{\mathbf{X}|\mathbf{Y}=0}(a) \propto \binom{n}{a}\left(\frac{1}{2}\right)^{n+a} \propto \binom{n}{a}\left(\frac{1}{2}\right)^{a}.$$

This looks like a binomial. If $\mathbf{Z} \sim \mathit{Binomial}(n,p)$, then

$$P_{\mathbf{Z}}(k) = \binom{n}{k}p^k(1-p)^{n-k} = \binom{n}{k}\left(\frac{p}{1-p}\right)^{k}(1-p)^n \propto \binom{n}{k}\left(\frac{p}{1-p}\right)^{k}.$$

So we want $\frac{p}{1-p} = \frac{1}{2}$, which we can solve to get $p = \frac{1}{3}$, as before.

The general answer

More generally, what is $\mathbf{X} \mid \mathbf{Y} = k$? We can do everything we just did, again, but there is a shortcut.

If k of the coins remained heads at the end of the experiment, they were certainly heads in the intermediate step. It's only the remaining $n - k$ heads we need to worry about.

For these $n - k$ heads, we are finding the number of them counted by $\mathbf{X}$, given that 0 of them were counted by $\mathbf{Y}$.

This is exactly the problem we solved, except with n replaced by $n - k$. So the general answer is that

$$\mathbf{X} - k \mid \mathbf{Y} = k \sim \textit{Binomial}\left(n - k, \frac{1}{3}\right).$$