Lecture 3: Discrete Random Variables

Contents

We start with a few more topics on probability on events, and then we move on to random variables.

1 Independent events and conditionally independent events

Definition 1 *Events E and F are* **independent** *if*

$$\mathbf{P}\{E \cap F\} = \mathbf{P}\{E\}\,\mathbf{P}\{F\}$$

Question: If E and F are independent, what is $\mathbf{P}\{E \mid F\}$?

Answer: Assuming $\mathbf{P}\{F\} > 0$, we have:

$$\mathbf{P}\{E \mid F\} = \frac{\mathbf{P}\{E \cap F\}}{\mathbf{P}\{F\}} \overset{indpt}{=} \frac{\mathbf{P}\{E\} \cdot \mathbf{P}\{F\}}{\mathbf{P}\{F\}} = \mathbf{P}\{E\}$$

That is, $\mathbf{P}\{E\}$ is not affected by whether F is true or not.

A different notion of independence which comes up frequently in problems (see the healthcare problem in Homework 1) is that of conditional independence.

Definition 2 *Two events E and F are said to be* **conditionally independent** *given event G, where $\mathbf{P}\{G\} > 0$, if*

$$\mathbf{P}\{E \cap F \mid G\} = \mathbf{P}\{E \mid G\} \cdot \mathbf{P}\{F \mid G\}$$

Observe that independence does not imply conditional independence and vice-versa.

2 Bayes Law

Sometimes, one needs to know $\mathbf{P}\{F|E\}$, but all one knows is the reverse direction: $\mathbf{P}\{E|F\}$. Is it possible to get $\mathbf{P}\{F|E\}$ from $\mathbf{P}\{E|F\}$? It turns out that it is, as follows:

Theorem 3 (Bayes Law)

$$\mathbf{P}\{F|E\} = \frac{\mathbf{P}\{E|F\} \cdot \mathbf{P}\{F\}}{\mathbf{P}\{E\}}$$

Proof:

$$\mathbf{P}\{F|E\} = \frac{\mathbf{P}\{E \cap F\}}{\mathbf{P}\{E\}} = \frac{\mathbf{P}\{E|F\} \cdot \mathbf{P}\{F\}}{\mathbf{P}\{E\}}$$

■

The Law of Total Probability can be combined with Bayes Law as follows: Let $F_1, F_2, \ldots, F_n$ be a partition of E. Then we can write: $\mathbf{P}\{E\} = \sum_{j=1}^{n} \mathbf{P}\left\{E \mid F_j\right\} \cdot \mathbf{P}\left\{F_j\right\}$ This yields:

Theorem 4 (Extended Bayes Law)

$$\mathbf{P}\{F|E\} = \frac{\mathbf{P}\{E|F\} \cdot \mathbf{P}\{F_i\}}{\mathbf{P}\{E\}} = \frac{\mathbf{P}\{E|F\} \cdot \mathbf{P}\{F\}}{\sum_{j=1}^{n} \mathbf{P}\left\{E|F_j\right\} \mathbf{P}\left\{F_j\right\}}$$

Example: A test is used to diagnose a rare disease. The test is only 95% accurate, meaning that, in a person who has the disease it will report "positive" with probability 95% (and negative otherwise), and in a person who does not have the disease, it will report "negative" with probability 95% (and positive otherwise). Suppose that 1 in a 10,000 children get the disease.

Question: A mom brings in her child to be tested. Given that the test comes back positive, how worried should the mom be?

Answer:

$$
\begin{aligned}
&\mathbf{P}\{\text{Child has disease} \mid \text{Test is positive}\} \\
&= \frac{\mathbf{P}\{\text{Test is Positive} \mid \text{Disease}\} \cdot \mathbf{P}\{\text{Disease}\}}{\mathbf{P}\{\text{Test is positive} \mid \text{Disease}\} \cdot \mathbf{P}\{\text{Disease}\} + \mathbf{P}\{\text{Test is positive} \mid \text{Healthy}\} \cdot \mathbf{P}\{\text{Healthy}\}} \\
&= \frac{0.95 \cdot \frac{1}{10000}}{0.95 \cdot \frac{1}{10000} + 0.05 \cdot \frac{9999}{10000}} \\
&= 0.0019
\end{aligned}
$$

Thus the probability that the child has the disease is only about 2 out of 1000.

3 Random variables

Consider an experiment, such as rolling two dice. Suppose that we are interested in the sum of the two rolls. That sum could range anywhere from 2 to 12, with each of these events having a different probability. A random variable, e.g. X, associated with this experiment, is a way to represent the value of the experiment (in this case the sum of the rolls). Specifically, when we write X, it is understood that X has many instances, ranging from 2 to 12 and that different instances occur with different probabilities, e.g., $\mathbf{P}\{X = 3\} = \frac{2}{36}$.

Formally, we say:

Definition 5 *A* **random variable** *(r.v.) is a real-valued function of the outcome of an experiment.*

Definition 6 *A* **discrete random variable** *can take on at most a countably-infinite number of possible values, whereas a* **continuous random variable** *can take on an uncountable set of possible values.*

Question: Which of these random variables is discrete and which is continuous?

1. The sum of the rolls of two dice.
2. The number of arrivals at a web site by time t.
3. The time until the next arrival at a web site.

4. The CPU requirement of an HTTP request.

Answer: The first of these can take on only a finite number of values, those between 2 and 12, so it clearly is a discrete r.v. The number of arrivals at a web site can take on the values: $0, 1, 2, 3, \ldots$ namely a countable set, hence this is discrete as well. Time, in general is modeled as a continuous quantity, even though there is a non-zero granularity in our ability to measure time via a computer. Thus quantities three and four above are continuous r.v.'s.

We use capital letters to denote random variables. For example, we might define X to be a random variable equal to the sum of two dice. Then,

$$\mathbf{P}\{X = 7\} = \mathbf{P}\{(1,6) \text{ or } (2,5) \text{ or } (3,4), \ldots, \text{ or } (6,1)\} = \frac{1}{6}$$

.

Important: Since the "outcome of the experiment" is just an event, all the theorems that we learned about events apply to random variables as well. In particular, the Law of Total Probability holds. For example, if N denotes the number of arrivals at a web site by time t, then $N > 10$ is an event. We can then use conditioning on events to get:

$$\mathbf{P}\{N > 10\} = \mathbf{P}\{N > 10 \mid \text{ weekday }\} \cdot \frac{5}{7} + \mathbf{P}\{N > 10 \mid \text{ weekend }\} \cdot \frac{2}{7}$$

All of this will become more concrete once we study examples of common random variables, coming up next.

4 Probability mass for discrete random variables

Discrete random variables take on a countable number of values, each with some probability.

Definition 7 *Let X be a discrete r.v. Then the* **probability mass function (p.m.f.)**, *$p_X(\cdot)$ of X, is defined as follows:*

$$p_X(a) = \mathbf{P}\{X = a\}, \text{where} \sum_x p_X(x) = 1$$

The **cumulative distribution function** *of X is defined as:*

$$F_X(a) = \mathbf{P}\{X \le a\} = \sum_{x \le a} p_X(x)$$

We also write:

$$\overline{F}_X(a) = \mathbf{P}\{X > a\} = \sum_{x>a} p_X(x) = 1 - F_X(a)$$

Common discrete distributions include the Bernoulli, the Binomial, the Geometric, and the Poisson, all of which are discussed below.

Bernoulli(p) represents the result of a single coin flip, where the coin has probability p of coming up heads (we map this event to the value 1) and $1 - p$ of coming up tails (we map this event to the value 0). If X is a r.v. drawn from the Bernoulli(p) distribution, then we write: $X \sim$ Bernoulli(p) and define X as follows:

$$X = \begin{cases} 1 & \text{w/ prob } p \\ 0 & \text{otherwise} \end{cases}$$

The p.m.f. of r.v. X is defined as follows:

$$\begin{aligned} p_X(1) &= p \\ p_X(0) &= 1 - p \end{aligned}$$

The p.m.f. is depicted in Figure 4.

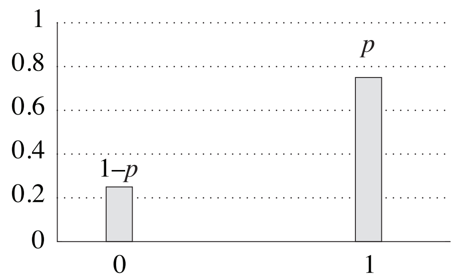

Figure 1: *Probability mass function of Bernoulli(p) distribution, with* $p = 0.78$.

Binomial(n,p) builds upon Bernoulli(p). Given a coin with probability p of coming up heads (success), we flip the coin n times (these are independent flips). If $X \sim$ Binomial(n, p), then X represents the number of heads (successes) when flipping a Bernoulli(p) coin n times. Observe that X can take on discrete values: $0, 1, 2, \ldots, n$.

The p.m.f. of r.v. X is defined as follows:

$$\begin{aligned} p_X(i) &= \mathbf{P}\{X = i\} \\ &= \binom{n}{i} p^i (1 - p)^{n-i}, \text{ where } i = 0, 1, 2, \ldots, n \end{aligned}$$

The p.m.f. is shown in Figure 2.

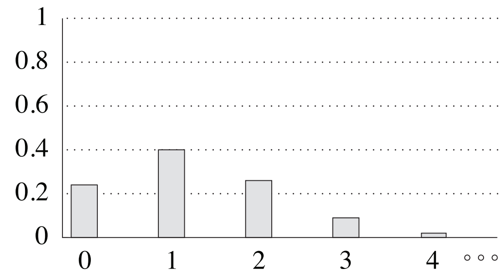

Figure 2: *Probability mass function of the Binomial*(n, p) *distribution, with* $n = 4$ *and* $p = 0.3$.

Geometric(p) also builds upon Bernoulli(p). Again we have a coin with probability p of coming up heads (success). We now flip it until we get a success; these are independent trials, each Bernoulli(p). If $X \sim$ Geometric(p), then X represents the number of flips until we get a success.

The p.m.f. of r.v. X is defined as follows:

$$\begin{aligned} p_X(i) &= \mathbf{P}\{X = i\} \\ &= (1-p)^{i-1}p, \text{ where } i = 1, 2, 3, \ldots \end{aligned}$$

The p.m.f. is shown in Figure 3.

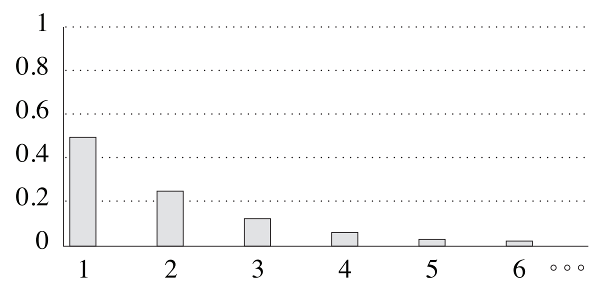

Figure 3: *Probability mass function of the Geometric*(p) *distribution, with* $p = 0.5$.

Question: Let's review. Suppose you have a room of n disks. Each disk independently dies with probability p each year. How are the following quantities distributed:

1. The number of disks that die in the first year.
2. The number of years until a particular disk dies.
3. The state of a particular disk after one year.

Answer: The distributions are: 1. Binomial(n, p), 2. Geometric(p), 3. Bernoulli(p).

Poisson(λ) is another discrete distribution that is very common in computer systems analysis. We will define Poisson(λ) via its p.m.f. While the p.m.f. doesn't appear to have any meaning at present, we will show many interesting properties of this distribution later in on the course. The Poisson distribution occurs naturally when looking at a mixture of a very large number of independent sources, each with a very small individual probability. It can therefore be a reasonable approximation for the number of arrivals to a web site or a router per unit time.

If $X \sim$ Poisson(λ), then

$$p_X(i) = \frac{e^{-\lambda}\lambda^i}{i!}, \text{ where } i = 0, 1, 2, \ldots$$

The p.m.f. for the Poisson(λ) distribution is shown in Figure 4.

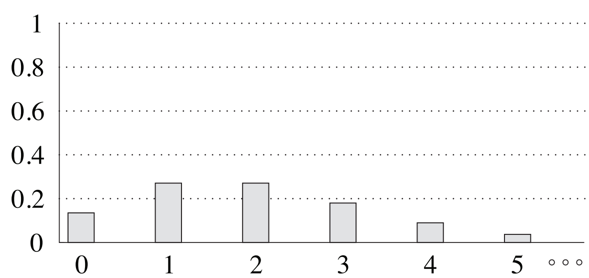

Figure 4: *Probability mass function of the Poisson(λ) distribution with* $\lambda = 2$.

You may have noticed that the Poisson distribution doesn't look all that different from the Binomial distribution. It turns out that if n is large and p is small, then Binomial(n, p) is actually very close to Poisson(np). You will prove this fact either in Homework 2 or in recitation this week!

5 Expectation & variance

The *mean* of a distribution, also known as its *expectation*, follows immediately from the probability mass function. For r.v. X, we write $\mathbf{E}\{X\}$ to denote its mean:

$$\mathbf{E}\{X\} = \sum_x x \cdot p_X(x) = \sum_x x\mathbf{P}\{X = x\}$$

The expectation of X can be viewed as a sum of the possible outcomes, each weighted by their probability.

To see this, consider the following example:

Example: Average cost of lunch

What is the average cost of my lunch?

Monday	\$7
Tuesday	\$7
Wednesday	\$5
Thursday	\$5
Friday	\$5
Saturday	\$0
Sunday	\$2

$$\boxed{\text{Avg} = \frac{7+7+5+5+5+0+2}{7}}$$

$$|||$$

$$\boxed{\mathbf{E}\{\text{Cost}\} = \frac{2}{7}(7) + \frac{3}{7}(5) + \frac{1}{7}(2) + \frac{1}{7}(0)}$$

Question: If $X \sim \text{Bernoulli}(p)$, what is $\mathbf{E}\{X\}$?

Answer: $\mathbf{E}\{X\} = 0 \cdot (1-p) + 1 \cdot (p) = p$

Question: Suppose a coin has probability $\frac{1}{3}$ of coming up heads. In expectation, how many times do I need to toss the coin to get a head?

Answer: This question is just looking for the $\mathbf{E}\{X\}$, where $X \sim \text{Geometric}(p)$, where $p = \frac{1}{3}$. Assuming $X \sim \text{Geometric}(p)$, we have:

$$\begin{aligned}
\mathbf{E}\{X\} &= \sum_{n=1}^{\infty} n(1-p)^{n-1} p \\
&= p \cdot \sum_{n=1}^{\infty} n \cdot q^{n-1} \qquad \text{where } q = (1-p) \\
&= p \cdot \sum_{n=1}^{\infty} \frac{d}{dq}(q^n) \\
&= p \cdot \frac{d}{dq} \sum_{n=1}^{\infty} q^n \\
&= p \cdot \frac{d}{dq}\left(q \sum_{n=0}^{\infty} q^n \right) \\
&= p \cdot \frac{d}{dq}\left(\frac{q}{1-q} \right) \\
&= \frac{p}{(1-q)^2} \\
&= \frac{1}{p}
\end{aligned}$$

So when $p = \frac{1}{3}$, the expected number of flips is 3.

Question: If $X \sim \text{Poisson}(\lambda)$, what is $\mathbf{E}\{X\}$?

Answer:

$$\begin{aligned}
\mathbf{E}\{X\} &= \sum_{i=0}^{\infty} i \frac{e^{-\lambda}\lambda^i}{i!} \\
&= \sum_{i=1}^{\infty} i \frac{e^{-\lambda}\lambda^i}{i!} \\
&= \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda
\end{aligned}$$

Thus the λ parameter for the Poisson distribution is also its mean.

We can also think about *higher moments* of a random variable X. The ith moment of discrete r.v. X, denoted by $\mathbf{E}\left\{X^i\right\}$ is defined as follows:

$$\mathbf{E}\left\{X^i\right\} = \sum_x x^i \cdot p_X(x)$$

More generally, we can talk about the *expectation of a function of a random variable* X. This is defined as follows for a discrete r.v. X:

$$\mathbf{E}\{g(X)\} = \sum_x g(x) \cdot p_X(x)$$

Question: Suppose X is defined as shown below,

$$X = \begin{cases} 0 & \text{w/ prob. } 0.2 \\ 1 & \text{w/ prob. } 0.5 \\ 2 & \text{w/ prob. } 0.3 \end{cases}$$

What is $\mathbf{E}\{X\}$ and what is $\mathbf{E}\left\{2X^2+3\right\}$?

Answer:

$$\begin{aligned} \mathbf{E}\{X\} &= (0)(.2) + (1)(.5) + (2)(.3) \\ \mathbf{E}\left\{2X^2+3\right\} &= \left(2 \cdot 0^2 + 3\right)(.2) + \left(2 \cdot 1^2 + 3\right)(.5) + \left(2 \cdot 2^2 + 3\right)(.3) \end{aligned}$$

You might have noticed that $\mathbf{E}\left\{2X^2+3\right\} = 2\mathbf{E}\left\{X^2\right\} + 3$. This is no coincidence and is due to "Linearity of Expectations" to be discussed a little later in the class.

The *variance* of a r.v. X, written as $\mathbf{Var}(X)$, is the expected squared difference of X from its mean, i.e., the square of how much we expect X to differ from its mean, $\mathbf{E}\{X\}$. This is defined as follows:

$$\mathbf{Var}(X) = \mathbf{E}\left\{(X - \mathbf{E}\{X\})^2\right\}$$

and can be equivalently expressed as follows:

$$\boxed{\mathbf{Var}(X) = \mathbf{E}\left\{X^2\right\} - (\mathbf{E}\{X\})^2}$$

(This proof of this equivalence will become obvious after we cover linearity of expectations).

Question: If $X \sim$ Bernoulli(p), what is $\mathbf{Var}(X)$?

Answer:

$$\begin{aligned}\mathbf{Var}(X) &= \mathbf{E}\left\{X^2\right\} - (\mathbf{E}\{X\})^2 \\ &= [0^2 \cdot (1-p) + 1^2 \cdot p] - p^2 \\ &= p(1-p)\end{aligned}$$

6 Joint probabilities and independence

We are often interested in probability statements concerning two or more r.v.'s simultaneously. For example, we might want to know the probability that two disks both crash within some time interval. The behavior of the two disks might be correlated or not. As another example, computer systems performance is often measured in terms of the Energy-Delay product, namely the product of the Energy used by the system and the delay experienced by users. Energy and delay typically are correlated with each other, and one can imagine a joint distribution between these two random variables. In this section and the next, we will present the definitions needed to formally express these ideas.

Definition 8 *The* **joint probability mass function** *between discrete r.v.'s X and Y is defined by:*

$$p_{X,Y}(x,y) = \mathbf{P}\{X = x \,\&\, Y = y\}$$

This is typically written as $\mathbf{P}\{X = x, Y = y\}$.

Question: What is the relationship between $p_X(x)$ and $p_{X,Y}(x,y)$?

Answer: Applying the Law of Total Probability, we have:

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

Similarly to the way we defined two events E and F as being independent, we can likewise define two r.v.'s as being independent.

Definition 9 *We say that discrete r.v.'s X and Y are* **independent**, *written $X \perp Y$, if*

$$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$$

Theorem 10 *If* $X \perp Y$, $\mathbf{E}\{XY\} = \mathbf{E}\{X\} \cdot \mathbf{E}\{Y\}$

Proof:

$$\begin{aligned}
\mathbf{E}\{XY\} &= \sum_x \sum_y xy \cdot \mathbf{P}\{X = x, Y = y\} \\
&= \sum_x \sum_y xy \cdot \mathbf{P}\{X = x\}\, \mathbf{P}\{Y = y\} \quad \text{(by definition of } \perp\text{)} \\
&= \sum_x x\mathbf{P}\{X = x\} \cdot \sum_y y\mathbf{P}\{Y = y\} \\
&= \mathbf{E}\{X\}\, \mathbf{E}\{Y\}
\end{aligned}$$

■

The same proof shows that if $X \perp Y$, then

$$\mathbf{E}\{g(X) f(Y)\} = \mathbf{E}\{g(X)\} \cdot \mathbf{E}\{f(Y)\}$$

Question: If $\mathbf{E}\{XY\} = \mathbf{E}\{X\}\,\mathbf{E}\{Y\}$, does that imply that $X \perp Y$?

Answer: No, figure out a counter-example.

7 Conditional probabilities and expectations

Just as we studied conditional probabilities of events, that is, the probability that one event occurs, given that another has occurred, we can also extend this to conditional probabilities in random variables.

The following example will help motivate the idea:

Example: Hair color

Suppose we divide the people in the class into Blondes, Red-heads, Brunettes, and Black-haired people. Let's say that 5 are Blondes, 2 are Red-heads, 17 are Brunettes, and 14 are Black-haired. Let X be a random variable whose value is hair color. Then the probability mass function for X looks like this:

$$\begin{aligned}
p_X(\text{Blonde}) &= 5/38 \\
p_X(\text{Red}) &= 2/38 \\
p_X(\text{Brown}) &= 17/38 \\
p_X(\text{Black}) &= 14/38
\end{aligned}$$

Now let's say that a person has *light-colored* hair if their hair color is either Blonde or Red. Let's say that a person has *dark-colored* hair if their hair color is either Brown or Black. Let A denote the event that a person's hair color is light.

$$\begin{aligned}\mathbf{P}\{A\} &= 7/38 \\ \mathbf{P}\left\{\overline{A}\right\} &= 31/38\end{aligned}$$

Definition 11 *Let X be a discrete r.v. with p.m.f. $p_X(\cdot)$ defined over a countable space. Let A be an event. Then $p_{X|A}(\cdot)$ is the* **conditional p.m.f.** *of X given event A. We define:*

$$p_{X|A}(x) \quad = \quad \mathbf{P}\{X = x \mid A\} = \frac{\mathbf{P}\{(X = x) \cap A\}}{\mathbf{P}\{A\}}$$

More formally, if Ω denotes the sample space and ω represents a sample point in the sample space, and $\{\omega : X(\omega) = x\}$ is the set of sample points that result in X having value x, then:

$$p_{X|A}(x) \quad = \quad \mathbf{P}\{X = x \mid A\} = \frac{\mathbf{P}\{\ \{\omega : X(\omega) = x\}\ \cap A\}}{\mathbf{P}\{A\}}$$

A conditional probability thus involves narrowing down the probability space.

For example

$$p_{X|A}(\text{Blonde}) = \frac{\mathbf{P}\{(X = \text{Blonde}) \cap A\}}{\mathbf{P}\{A\}} = \frac{\frac{5}{38}}{\frac{7}{38}} = \frac{5}{7}$$

Likewise $p_{X|A}(\text{Red}) = 2/7$.

As another example:

$$p_{X|A}(\text{Brown}) = \frac{\mathbf{P}\{(X = \text{Brown}) \cap A\}}{\mathbf{P}\{A\}} = \frac{0}{\frac{7}{38}} = 0$$

Likewise $p_{X|A}(\text{Black}) = 0$.

Question: If we sum $p_{X|A}(x)$ over all x what do we get?

Answer:

$$\sum_x p_{X|A}(x) = \sum_x \frac{\mathbf{P}\{(X = x) \cap A\}}{\mathbf{P}\{A\}} = \frac{\mathbf{P}\{A\}}{\mathbf{P}\{A\}} = 1$$

Thus $p_{X|A}(x)$ is a valid p.m.f.

Question: So how do we use this to define the conditional expectation of X given A?

Answer:

$$\mathbf{E}\{X|A\} \quad = \quad \sum_x x p_{X|A}(x) = \sum_x x \cdot \frac{\mathbf{P}\{(X = x) \cap A\}}{\mathbf{P}\{A\}}$$

We can also consider the case where the event, A, is an instance of a random variable, e.g., $(Y = y)$. It is then common to write the conditional p.m.f. of X given $(Y = y)$ as:

$$p_{X|Y}(x|y) = \mathbf{P}\{X = x \mid Y = y\} = \frac{\mathbf{P}\{X = x \;\&\; Y = y\}}{\mathbf{P}\{Y = y\}} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

where

$$\mathbf{E}\{X \mid Y = y\} = \sum_x x \cdot p_{X|Y}(x|y)$$

Here's an example of conditioning on random variables:

Example:

Two discrete random variables X and Y taking the values $\{0, 1, 2\}$ have a joint probability mass function given by the following table:

	2	0	$\frac{1}{6}$	$\frac{1}{8}$
Y	1	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{8}$
	0	$\frac{1}{6}$	$\frac{1}{8}$	0
		0	1	2
			X	

Question: Compute the conditional expectation $\mathbf{E}\{X \mid Y = 2\}$.

Answer:

$$
\begin{aligned}
\mathbf{E}\{X \mid Y = 2\} &= \sum_x x \cdot p_{X|Y}(x, 2) \\
&= \sum_x x \cdot \mathbf{P}\{X = x \mid Y = 2\} \\
&= 0 \cdot \frac{\mathbf{P}\{X = 0 \,\&\, Y = 2\}}{\mathbf{P}\{Y = 2\}} + 1 \cdot \frac{\mathbf{P}\{X = 1 \,\&\, Y = 2\}}{\mathbf{P}\{Y = 2\}} + 2 \cdot \frac{\mathbf{P}\{X = 2 \,\&\, Y = 2\}}{\mathbf{P}\{Y = 2\}} \\
&= 1 \cdot \frac{\frac{1}{6}}{\frac{7}{24}} + 2 \cdot \frac{\frac{1}{8}}{\frac{7}{24}} \\
&= 1 \cdot \frac{4}{7} + 2 \cdot \frac{3}{7} \\
&= \frac{10}{7}
\end{aligned}
$$

Question: Compute the conditional expectation $\mathbf{E}\{X \mid Y \neq 1\}$.

Answer: You should get $\frac{13}{14}$.