Lecture 4: Conditioning, Expectation, and Variance for Discrete Random Variables

1 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 (color value 1), Red-heads (color value 2), Brunettes (color value 3), and Black-haired people (color value 4). 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 1 *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:*

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

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.

We define the conditional expectation of X given A as follows:

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

Question: For our example, viewing Blonde as having value 1 and Red-haired as having value 2, what is $\mathbf{E}\{X \mid A\}$?

Answer:

$$\mathbf{E}\{X \mid A\} = 1 \cdot \frac{5}{7} + 2 \cdot \frac{2}{7} = \frac{9}{7}$$

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:

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

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:

$$\begin{aligned}
\mathbf{E}\{X \mid Y \neq 1\} &= \sum_x x \cdot p_{X|Y\neq 1}(x) \\
&= \sum_x x \cdot \mathbf{P}\{X = x \mid Y \neq 1\} \\
&= 0 \cdot \frac{\mathbf{P}\{X = 0 \,\&\, Y \neq 1\}}{\mathbf{P}\{Y \neq 1\}} + 1 \cdot \frac{\mathbf{P}\{X = 1 \,\&\, Y \neq 1\}}{\mathbf{P}\{Y \neq 1\}} + 2 \cdot \frac{\mathbf{P}\{X = 2 \,\&\, Y \neq 1\}}{\mathbf{P}\{Y \neq 1\}} \\
&= 1 \cdot \frac{\frac{1}{6} + \frac{1}{8}}{\frac{14}{24}} + 2 \cdot \frac{\frac{1}{8}}{\frac{14}{24}} \\
&= \frac{7}{14} + \frac{6}{14} \\
&= \frac{13}{14}
\end{aligned}$$

2 Probabilities and expectations via conditioning

Recall from the Law of Total Probability, if $F_1, \ldots, F_n$ partition the sample space S then

$$\mathbf{P}\{E\} = \sum_{i=1}^{n} \mathbf{P}\{E \mid F_i\}\,\mathbf{P}\{F_i\}\,.$$

Unsurprisingly, this extends to random variables, since "$X = k$" is an event.

The **Law of Total Probability for Discrete Random Variables** says:

$$\mathbf{P}\{X = k\} = \sum_y \mathbf{P}\{X = k \mid Y = y\}\,\mathbf{P}\{Y = y\}$$

This is a *huge* tool! It allows us to break a problem into a number of simpler problems. The trick, as usual, is knowing what to condition on.

Example: Which head will come up first?

Suppose I have a coin which comes up Heads with probability $p_1 = \frac{1}{4}$. I have a second (independent) coin which comes up Heads with probability $p_2 = \frac{1}{3}$. Every second I flip both coins.

Question: What is the probability that the first coin comes up Heads (strictly) before the second coin?

Hint: We're talking about Geometric random variables.

Answer: Let $X_1 \sim \text{Geometric}(p_1)$ and let $X_2 \sim \text{Geometric}(p_2)$. The question is asking what is $\mathbf{P}\{X_1 < X_2\}$.

$$\begin{aligned}
\mathbf{P}\{X_1 < X_2\} &= \sum_{k=1}^{\infty} \mathbf{P}\{X_1 < X_2 \mid X_1 = k\} \cdot \mathbf{P}\{X_1 = k\} \\
&= \sum_{k=1}^{\infty} \frac{\mathbf{P}\{k < X_2 \;\&\; X_1 = k\}}{\mathbf{P}\{X_1 = k\}} \cdot \mathbf{P}\{X_1 = k\} \\
&= \sum_{k=1}^{\infty} \frac{\mathbf{P}\{k < X_2\} \cdot \mathbf{P}\{X_1 = k\}}{\mathbf{P}\{X_1 = k\}} \cdot \mathbf{P}\{X_1 = k\} \\
&= \sum_{k=1}^{\infty} \mathbf{P}\{k < X_2\} \cdot \mathbf{P}\{X_1 = k\} \\
&= \sum_{k=1}^{\infty} (1-p_2)^k \cdot (1-p_1)^{k-1} p_1 \\
&= p_1(1-p_2) \sum_{k=1}^{\infty} (1-p_2)^{k-1} \cdot (1-p_1)^{k-1} \\
&= \frac{p_1(1-p_2)}{1-(1-p_2)\cdot(1-p_1)}
\end{aligned}$$

Question: Can you explain this answer? (Think about coin flips)

Answer: Observe that all the flips are irrelevant until the final flip. In the final flip, we know that at least one coin comes up Heads, and that event determines the probability. The above is the probability that, on a single flip, where we're given that at least one coin came up Heads, it is the case that the first coin comes up Heads and the second one Tails. Thus the expression above is the probability that on the final flip we see the outcome we want.

Theorem 2 *For discrete random variables:*

$$\mathbf{E}\{X\} = \sum_y \mathbf{E}\{X \mid Y = y\}\,\mathbf{P}\{Y = y\}$$

Proof:

$$\begin{aligned}
\mathbf{E}\{X\} &= \sum_x x\mathbf{P}\{X = x\} \\
&= \sum_x x \sum_y \mathbf{P}\{X = x \mid Y = y\}\,\mathbf{P}\{Y = y\} \\
&= \sum_x \sum_y x\mathbf{P}\{X = x \mid Y = y\}\,\mathbf{P}\{Y = y\} \\
&= \sum_y \sum_x x\mathbf{P}\{X = x \mid Y = y\}\,\mathbf{P}\{Y = y\} \\
&= \sum_y \mathbf{P}\{Y = y\} \sum_x x\mathbf{P}\{X = x \mid Y = y\} \\
&= \sum_y \mathbf{P}\{Y = y\}\,\mathbf{E}\{X \mid Y = y\}
\end{aligned}$$

■

This proof generalizes to:

$$\mathbf{E}\{g(X)\} = \sum_y \mathbf{E}\{g(X) \mid Y = y\}\,\mathbf{P}\{Y = y\}$$

which is very important when we need to compute the second moment of X, such as when computing variance.

Example: Mean of Geometric

Suppose we want to use conditioning to *easily* compute the mean of the Geometric distribution with parameter p. That is, if N = number of flips required to get first head, we want $\mathbf{E}\{N\}$.

Question: What do we condition on?

Answer: We condition on the value of the first flip, Y, as follows:

$$\begin{aligned}
\mathbf{E}\{N\} &= \mathbf{E}\{N \mid Y=1\}\mathbf{P}\{Y=1\} + \mathbf{E}\{N \mid Y=0\}\mathbf{P}\{Y=0\} \\
&= 1p + (1+\mathbf{E}\{N\})(1-p) \\
p\mathbf{E}\{N\} &= p + (1-p) \\
\mathbf{E}\{N\} &= \frac{1}{p}
\end{aligned}$$

Note how much simpler this derivation is than our original derivation of the mean of a Geometric!

3 Linearity of expectation

The following is one of the most powerful theorems of probability:

Theorem 3 (Linearity of Expectation) *For random variables, X and Y,*

$$\mathbf{E}\{X+Y\} = \mathbf{E}\{X\} + \mathbf{E}\{Y\}$$

Question: Do we need $X \perp Y$?

Question: No! Recall that we *do* need independence for simplifying $\mathbf{E}\{XY\}$, but not for $\mathbf{E}\{X+Y\}$.

Proof: Here's a proof in the case where X and Y are discrete.

$$\begin{aligned}
\mathbf{E}\{X+Y\} &= \sum_y \sum_x (x+y) p_{X,Y}(x,y) \\
&= \sum_y \sum_x x p_{X,Y}(x,y) + \sum_y \sum_x y p_{X,Y}(x,y) \\
&= \sum_x \sum_y x p_{X,Y}(x,y) + \sum_y \sum_x y p_{X,Y}(x,y) \\
&= \sum_x x \sum_y p_{X,Y}(x,y) + \sum_y y \sum_x p_{X,Y}(x,y) \\
&= \sum_x x p_X(x) + \sum_y y p_Y(y) \\
&= \mathbf{E}\{X\} + \mathbf{E}\{Y\}
\end{aligned}$$

■

This simple identity can simplify a lot of proofs. Consider the example below:

Example: Binomial

$X \sim \text{Binomial}(n, p)$. What is $\mathbf{E}\{X\}$?

Question: If we simply use the definition of the Binomial, what expression do we have for $\mathbf{E}\{X\}$?

Answer: $\mathbf{E}\{X\} = \sum_{i=0}^{n} i\binom{n}{i}p^i(1-p)^{n-i}$. This does not seem easy to simplify.

Question: Is there a way to think of Binomial(n, p) as a sum of random variables?

Answer: Let

$$X = \text{number of successes in n trials}$$

$$X = X_1 + X_2 + \cdots + X_n$$

where

$$\begin{aligned} X_i &= \begin{cases} 1 & \text{if trial i is successful} \\ 0 & \text{otherwise} \end{cases} \\ \mathbf{E}\{X_i\} &= p \end{aligned}$$

$$\begin{aligned} \mathbf{E}\{X\} &= \mathbf{E}\{X_1\} + \mathbf{E}\{X_2\} + \cdots + \mathbf{E}\{X_n\} \\ &= n\mathbf{E}\{X_i\} \\ &= np \end{aligned}$$

This result should make sense, since n coin-flips, each with probability p of coming up heads, should result in an average of np heads.

The X_i's above are called **indicator random variables** because they take on values 0 or 1.

In the above example, the X_i's were i.i.d. (independent and identically distributed). Even if the trials weren't independent, we would have

$$\mathbf{E}\{X\} = \mathbf{E}\{X_1\} + \cdots + \mathbf{E}\{X_n\}$$

The following example makes this clear.

Example: Hats

At a party, n people throw their hat into the middle of a circle. Each closes his or her eyes and picks a random hat. Let X denote the number of people who get back their own hat. Our goal is to determine $\mathbf{E}\{X\}$.

Question: How can we express X as a sum of indicator random variables?

Answer: $X = I_1 + I_2 + \cdots + I_n$, where

$$I_i = \begin{cases} 1 & \text{if ith person gets the right hat} \\ 0 & \text{otherwise} \end{cases}$$

Observe that while the I_i's have the same distribution (by symmetry), they are *not* independent of each other! Nevertheless, we can still use Linearity of Expectation to say:

$$\begin{aligned} \mathbf{E}\{X\} &= \mathbf{E}\{I_1\} + \mathbf{E}\{I_2\} + \cdots + \mathbf{E}\{I_n\} \\ &= n\mathbf{E}\{I_i\} \\ &= n\left(\frac{1}{n} \cdot 1 + \frac{n-1}{n} \cdot 0\right) \\ &= 1 \end{aligned}$$

Linearity of Expectations can also be used to show that the two definitions we gave for variance are identical as follows:

$$\begin{aligned} \mathbf{Var}(X) &= \mathbf{E}\left\{(X - \mathbf{E}\{X\})^2\right\} \\ &= \mathbf{E}\left\{X^2 - 2X\mathbf{E}\{X\} + \mathbf{E}\{X\}^2\right\} \\ &= \mathbf{E}\left\{X^2\right\} - 2\mathbf{E}\{X\}\mathbf{E}\{X\} + \mathbf{E}\{X\}^2 \\ &= \mathbf{E}\left\{X^2\right\} - \mathbf{E}\{X\}^2 \end{aligned}$$

Observe that Linearity of Expectation also implies that:

$$\mathbf{E}\left\{X^2 + Y^2\right\} = \mathbf{E}\left\{X^2\right\} + \mathbf{E}\left\{Y^2\right\}$$

Nonetheless this *does not* imply that linearity holds for variance. For that, we require an independence assumption, as in the theorem below:

Theorem 4 *Let X and Y be random variables where $X \perp Y$. Then*

$$\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y)$$

Proof:

$$\begin{aligned}
\mathbf{Var}(X+Y) &= \mathbf{E}\left\{(X+Y)^2\right\} - \mathbf{E}\left\{X+Y\right\}^2 \\
&= \mathbf{E}\left\{X^2\right\} + \mathbf{E}\left\{Y^2\right\} + 2\mathbf{E}\left\{XY\right\} - (\mathbf{E}\left\{X\right\} + \mathbf{E}\left\{Y\right\})^2 \\
&= \mathbf{E}\left\{X^2\right\} + \mathbf{E}\left\{Y^2\right\} + 2\mathbf{E}\left\{XY\right\} \\
&\quad -\mathbf{E}\left\{X\right\}^2 - \mathbf{E}\left\{Y\right\}^2 - 2\mathbf{E}\left\{X\right\}\mathbf{E}\left\{Y\right\} \\
&= \mathbf{Var}(X) + \mathbf{Var}(Y) \\
&\quad \underbrace{+2\mathbf{E}\left\{XY\right\} - 2\mathbf{E}\left\{X\right\}\mathbf{E}\left\{Y\right\}}_{\text{equals 0 if X} \perp \text{ Y}}
\end{aligned}$$

■

Example: Variance of Binomial

Let $X \sim$ Binomial(n, p).

Question: What is $\mathbf{Var}(X)$?

Answer: We will use the same trick we used in deriving the mean of Binomial(n, p).

Let

$$X = \text{number of successes in n trials}$$

$$X = X_1 + X_2 + \cdots + X_n$$

where

$$\begin{aligned}
X_i &= \begin{cases} 1 & \text{if trial i is successful} \\ 0 & \text{otherwise} \end{cases} \\
\mathbf{E}\left\{X_i\right\} &= p \\
\mathbf{Var}(X_i) &= \mathbf{E}\left\{X_i^2\right\} - \mathbf{E}\left\{X_i\right\}^2 \\
&= p - p^2 \\
&= p(1-p)
\end{aligned}$$

Since the X_i's are independent, we have that:

$$\begin{aligned}
\mathbf{Var}(X) &= \mathbf{Var}(X_1) + \mathbf{Var}(X_2) + \cdots + \mathbf{Var}(X_n) \\
&= n\mathbf{Var}(X_i) \\
&= np(1-p)
\end{aligned}$$

4 Sum of a random number of random variables

It is common in many applications that one needs to add up a number of i.i.d. random variables, where the number of these variables is itself a random variable. Specifically, we're talking about the quantity S below.

Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables. Let

$$S = \sum_{i=1}^{N} X_i \qquad N \perp X_i$$

where N is a non-negative, integer-valued random variable.

In this section we discuss how to derive quantities like $\mathbf{E}\{S\}$ and $\mathbf{E}\left\{S^2\right\}$.

Question: Why can't we directly apply linearity of expectations?

Answer: Linearity equations only apply when N is a constant.

Question: Does this give you any ideas?

Answer: Let's condition on the value of N, and then apply linearity of expectations.

$$\begin{aligned}
\mathbf{E}\{S\} = \mathbf{E}\left\{\sum_{i=1}^{N} X_i\right\} &= \sum_n \mathbf{E}\left\{\sum_{i=1}^{N} X_i \mid N = n\right\} \cdot \mathbf{P}\{N = n\} \\
&= \sum_n \mathbf{E}\left\{\sum_{i=1}^{n} X_i\right\} \cdot \mathbf{P}\{N = n\} \\
&= \sum_n n\mathbf{E}\{X\} \cdot \mathbf{P}\{N = n\} \\
&= \mathbf{E}\{X\} \cdot \mathbf{E}\{N\} \qquad (1)
\end{aligned}$$

********** Reading beyond this point is optional ***********

Question: Can we use the same trick to get $\mathbf{E}\left\{S^2\right\}$?

Answer: The difficulty with conditioning on N is that we end up with a big sum that we need to square, and it's not obvious how to do that. Consider the following:

$$\begin{aligned}\mathbf{E}\left\{S^2\right\} &= \sum_n \mathbf{E}\left\{S^2 \mid N=n\right\} \cdot \mathbf{P}\{N=n\} \\ &= \sum_n \mathbf{E}\left\{\left(\sum_{i=1}^{n} X_i\right)^2\right\} \cdot \mathbf{P}\{N=n\}\end{aligned}$$

A better idea is to first derive $\mathbf{Var}(S \mid N=n)$ and then use that to get $\mathbf{E}\left\{S^2 \mid N=n\right\}$.

Question: What is $\mathbf{Var}(S \mid N=n)$?

Answer: By independence,

$$\mathbf{Var}(S \mid N=n) = n\mathbf{Var}(X)$$

Observe also that

$$\begin{aligned}n\mathbf{Var}(X) = \mathbf{Var}(S \mid N=n) &= \mathbf{E}\left\{S^2 \mid N=n\right\} - (\mathbf{E}\{S \mid N=n\})^2 \\ &= \mathbf{E}\left\{S^2 \mid N=n\right\} - (n\mathbf{E}\{X\})^2\end{aligned}$$

From the above expression, we have that:

$$\mathbf{E}\left\{S^2 \mid N=n\right\} = n\mathbf{Var}(X) + n^2 (\mathbf{E}\{X\})^2$$

It follows that:

$$\begin{aligned}\mathbf{E}\left\{S^2\right\} &= \sum_n \mathbf{E}\left\{S^2 \mid N=n\right\} \cdot \mathbf{P}\{N=n\} \\ &= \sum_n \left(n\mathbf{Var}(X) + n^2 (\mathbf{E}\{X\})^2\right) \mathbf{P}\{N=n\} \\ &= \mathbf{E}\{N\}\,\mathbf{Var}(X) + \mathbf{E}\left\{N^2\right\} (\mathbf{E}\{X\})^2\end{aligned}$$

Furthermore:

$$\begin{aligned}
\mathbf{Var}(S) &= \mathbf{E}\left\{S^2\right\} - (\mathbf{E}\{S\})^2 \\
&= \mathbf{E}\{N\}\,\mathbf{Var}(X) + \mathbf{E}\left\{N^2\right\}(\mathbf{E}\{X\})^2 - (\mathbf{E}\{N\}\,\mathbf{E}\{X\})^2 \\
&= \mathbf{E}\{N\}\,\mathbf{Var}(X) + \mathbf{Var}(N)\,(\mathbf{E}\{X\})^2
\end{aligned}$$

We have proven the following theorem:

Theorem 5 *Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables. Let*

$$S = \sum_{i=1}^{N} X_i \qquad N \perp X_i$$

Then

$$\mathbf{E}\{S\} = \mathbf{E}\{N\}\,\mathbf{E}\{X\}$$

$$\mathbf{E}\left\{S^2\right\} = \mathbf{E}\{N\}\,\mathbf{Var}(X) + \mathbf{E}\left\{N^2\right\}(\mathbf{E}\{X\})^2$$

$$\mathbf{Var}(S) = \mathbf{E}\{N\}\,\mathbf{Var}(X) + \mathbf{Var}(N)\,(\mathbf{E}\{X\})^2$$

The variance trick was pretty cool. You may be wondering how we would get the third moment, $\mathbf{E}\left\{S^3\right\}$, if we ever needed it, given that the variance trick won't work there. The answer is to use transform analysis (generating functions), which will easily provide any moment of S. Depending on time, we may get to cover this extremely useful topic later in this class.