Introduction to Machine Learning

Bayesian Decision Theory

The 7th lecture, 09.03.2022

Phuc Loi Luu, PhD
p.luu@garvan.org.au
luu.p.loi@googlemail.com

Roadmap for today

- Joint Distributions: Two Discrete Random Variables
- Toy example: Predict gender of a person based on height
- Bayes Decision Theory

Joint Distributions: Two Discrete Random Variables

Example 17-1

Suppose we toss a pair of fair, four-sided dice, in which one of the dice is RED and the other is **BLACK**. We'll let:

- X = the outcome on the RED die = $\{1, 2, 3, 4\}$
- Y = the outcome on the **BLACK** die = $\{1, 2, 3, 4\}$

What is the probability that X takes on a particular value x, and Y takes on a particular value y? That is, what is $P(X = x, Y = y)$?

https://online.stat.psu.edu/stat414/lesson/17/17.1

Joint Distributions: Two Discrete Random Variables

Just as we have to in the case with one discrete random variable, in order to find the "**joint probability distribution**" of X and Y, we first need to define the support of X and Y. Well, the support of X is:

$$S_1 = \{1, 2, 3, 4\}$$

And, the support of Y is:

$$S_2 = \{1, 2, 3, 4\}$$

Now, if we let (x, y) denote one of the possible outcomes of one toss of the pair of dice, then certainly (1, 1) is a possible outcome, as is (1, 2), (1, 3) and (1, 4). If we continue to enumerate all of the possible outcomes, we soon see that the joint support *S* has 16 possible outcomes:

$$S = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4), (4,1), (4,2), (4,3), (4,4)\}$$

Now, because the dice are fair, we should expect each of the 16 possible outcomes to be equally likely. Therefore, using the classical approach to assigning probability, the probability that X equals any particular x value, and Y equals any particular y value, is $\frac{1}{16}$. That is, for all (x, y) in the support S:

$$P(X = x, Y = y) = \frac{1}{16}$$

Joint Probability Mass Function (PMF)

Because we have identified the probability for each (x, y), we have found what we call the **joint probability mass function**. Perhaps, it is not too surprising that the joint probability mass function, which is typically denoted as $f(x, y)$, can be defined as a formula (as we have above), as a graph, or as a table. Here's what our joint p.m.f. would like in tabular form:

		BLACK (Y)				
f(x,y)		1	2	3	4	$f_X(x)$
RED (X)	1	1/16	1/16	1/16	1/16	4/16
	2	1/16	1/16	1/16	1/16	4/16
	3	1/16	1/16	1/16	1/16	4/16
	4	1/16	1/16	1/16	1/16	4/16
$f_Y(y)$		4/16	4/16	4/16	4/16	1

Joint probability mass function (PMF)

Let X and Y be two discrete random variables, and let S denote the two-dimensional support of X and Y. Then, the function $f(x,y) = P(X=x, Y=y)$ is a **joint probability mass function** (abbreviated p.m.f.) if it satisfies the following three conditions:

1. $0 \leq f(x,y) \leq 1$
2. $\sum\limits_{(x,y)\in S}\sum f(x,y) = 1$
3. $P[(X,Y) \in A] = \sum\limits_{(x,y)\in A}\sum f(x,y)$ where A is a subset of the support S.

The third condition tells us that in order to determine the probability of an event A, you simply sum up the probabilities of the (x,y) values in A.

Marginal Probability Mass function of X

Let X be a discrete random variable with support S_1, and let Y be a discrete random variable with support S_2. Let X and Y have the joint probability mass function $f(x, y)$ with support S. Then, the probability mass function of X alone, which is called the **marginal probability mass function of** X, is defined by:

$$f_X(x) = \sum_y f(x, y) = P(X = x), \qquad x \in S_1$$

where, for each x in the support S_1, the summation is taken over all possible values of y. Similarly, the probability mass function of Y alone, which is called the **marginal probability mass function of** Y, is defined by:

$$f_Y(y) = \sum_x f(x, y) = P(Y = y), \qquad y \in S_2$$

where, for each y in the support S_2, the summation is taken over all possible values of x.

If you again take a look back at the representation of our joint p.m.f. in tabular form, you might notice that the following holds true:

$$P(X = x, Y = y) = \frac{1}{16} = P(X = x) \cdot P(Y = y) = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$$

for all $x \in S_1, y \in S_2$. When this happens, we say that X and Y are **independent**. A formal definition of the independence of two random variables X and Y follows.

	f(x,y)	BLACK (Y) 1	2	3	4	$f_X(x)$
RED (X)	1	1/16	1/16	1/16	1/16	4/16
	2	1/16	1/16	1/16	1/16	4/16
	3	1/16	1/16	1/16	1/16	4/16
	4	1/16	1/16	1/16	1/16	4/16
	$f_Y(y)$	4/16	4/16	4/16	4/16	1

To find the probability mass function of X, we sum, for each x, the probabilities when y=1,2,3 and 4. That is, for each x, we sum f(x,1), f(x,2), f(x,3) and f(x,4).

Independent, Dependent and Expectation

The random variables X and Y are **independent** if and only if:

$$P(X=x, Y=y) = P(X=x) \times P(Y=y)$$

for all $x \in S_1, y \in S_2$. Otherwise, X and Y are said to be **dependent**.

Now, suppose we were given a joint probability mass function $f(x,y)$, and we wanted to find the mean of X. Well, one strategy would be to find the marginal p.m.f of X first, and then use the definition of the expected value that we previously learned to calculate $E(X)$. Alternatively, we could use the following definition of the mean that has been extended to accommodate joint probability mass functions.

Definition. Let X be a discrete random variable with support S_1, and let Y be a discrete random variable with support S_2. Let X and Y be discrete random variables with joint p.m.f. $f(x,y)$ on the support S. If $u(X,Y)$ is a function of these two random variables, then:

$$E[u(X,Y)] = \sum\sum_{(x,y)\in S} u(x,y)f(x,y)$$

if it exists, is called the **expected value** of $u(X,Y)$. If $u(X,Y) = X$, then:

$$\mu_X = E[X] = \sum_{x\in S_1}\sum_{y\in S_2} xf(x,y)$$

if it exists, is the **mean of** X. If $u(X,Y) = Y$, then:

$$\mu_Y = E[Y] = \sum_{x\in S_1}\sum_{y\in S_2} yf(x,y)$$

if it exists, is the **mean of** Y.

Conditioning and Independence

We have discussed conditional probability before, and you have already seen some problems regarding random variables and conditional probability. Here, we will discuss conditioning for random variables more in detail and introduce the conditional PMF, conditional CDF, and conditional expectation. We would like to emphasize that there is only one main formula regarding conditional probability which is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ when } P(B) > 0.$$

Any other formula regarding conditional probability can be derived from the above formula. Specifically, if you have two random variables X and Y, you can write

$$P(X \in C | Y \in D) = \frac{P(X \in C, Y \in D)}{P(Y \in D)}, \text{ where } C, D \subset \mathbb{R}.$$

Conditional PMF and CDF:

Remember that the PMF is by definition a probability measure, i.e., it is $P(X = x_k)$. Thus, we can talk about the **conditional PMF**. Specifically, the conditional PMF of X given event A, is defined as

$$\begin{aligned} P_{X|A}(x_i) &= P(X = x_i | A) \\ &= \frac{P(X = x_i \text{ and } A)}{P(A)}. \end{aligned}$$

Example 5.3

I roll a fair die. Let X be the observed number. Find the conditional PMF of X given that we know the observed number was less than 5.

Solution

Here, we condition on the event $A = \{X < 5\}$, where $P(A) = \frac{4}{6}$. Thus,

$$\begin{aligned} P_{X|A}(1) &= P(X = 1 | X < 5) \\ &= \frac{P(X = 1 \text{ and } X < 5)}{P(X < 5)} \\ &= \frac{P(X = 1)}{P(X < 5)} = \frac{1}{4}. \end{aligned}$$

Similarly, we have

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

Also,

$$P_{X|A}(5) = P_{X|A}(6) = 0.$$

For a discrete random variable X and event A, the **conditional PMF** of X given A is defined as

$$\begin{aligned} P_{X|A}(x_i) &= P(X = x_i | A) \\ &= \frac{P(X = x_i \text{ and } A)}{P(A)}, \quad \text{for any } x_i \in R_X. \end{aligned}$$

Similarly, we define the **conditional CDF** of X given A as

$$F_{X|A}(x) = P(X \leq x | A).$$

Conditioning and Independence

Conditional PMF of X Given Y:

In some problems, we have observed the value of a random variable Y, and we need to update the PMF of another random variable X whose value has not yet been observed. In these problems, we use the **conditional PMF** of X given Y. The conditional PMF of X given Y is defined as

$$\begin{aligned} P_{X|Y}(x_i|y_j) &= P(X=x_i|Y=y_j) \\ &= \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} \\ &= \frac{P_{XY}(x_i,y_j)}{P_Y(y_j)}. \end{aligned}$$

Similarly, we can define the conditional probability of Y given X:

$$\begin{aligned} P_{Y|X}(y_j|x_i) &= P(Y=y_j|X=x_i) \\ &= \frac{P_{XY}(x_i,y_j)}{P_X(x_i)}. \end{aligned}$$

For discrete random variables X and Y, the **conditional PMFs** of X given Y and vice versa are defined as

$$P_{X|Y}(x_i|y_j) = \frac{P_{XY}(x_i,y_j)}{P_Y(y_j)},$$
$$P_{Y|X}(y_j|x_i) = \frac{P_{XY}(x_i,y_j)}{P_X(x_i)}$$

for any $x_i \in R_X$ and $y_j \in R_Y$.

Independent Random Variables:

We have defined independent random variables previously. Now that we have seen joint PMFs and CDFs, we can restate the independence definition.

Two discrete random variables X and Y are independent if

$$P_{XY}(x,y) = P_X(x)P_Y(y), \quad \text{for all } x,y.$$

Equivalently, X and Y are independent if

$$F_{XY}(x,y) = F_X(x)F_Y(y), \quad \text{for all } x,y.$$

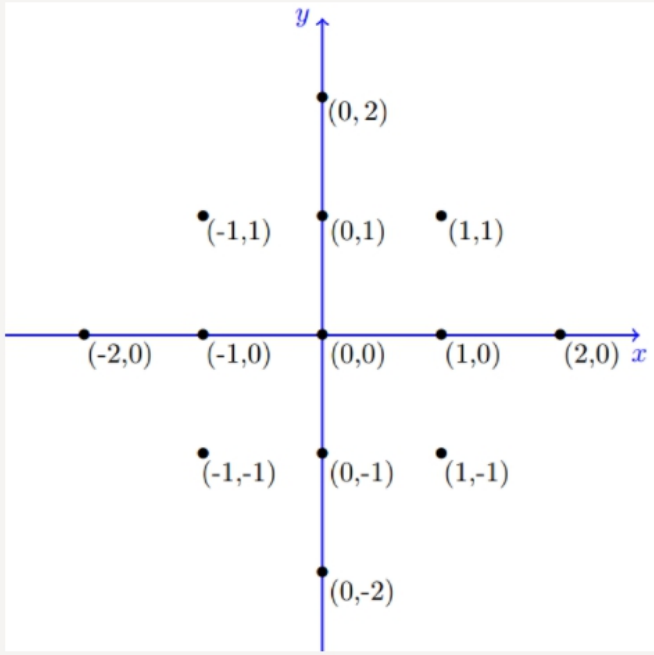


Figure 5.4: Grid for example 5.4

Conditioning and Independence

So, if X and Y are independent, we have

$$\begin{aligned} P_{X|Y}(x_i|y_j) &= P(X = x_i|Y = y_j) \\ &= \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)} \\ &= \frac{P_X(x_i)P_Y(y_j)}{P_Y(y_j)} \\ &= P_X(x_i). \end{aligned}$$

As we expect, for independent random variables, the conditional PMF is equal to the marginal PMF. In other words, knowing the value of Y does not provide any information about X.

Example 5.4

Consider the set of points in the grid shown in Figure 5.4. These are the points in set G defined as

$$G = \{(x, y)|x, y \in \mathbb{Z}, |x| + |y| \leq 2\}.$$

Suppose that we pick a point (X, Y) from this grid completely at random. Thus, each point has a probability of $\frac{1}{13}$ of being chosen.

a. Find the joint and marginal PMFs of X and Y.
b. Find the conditional PMF of X given $Y = 1$.
c. Are X and Y independent?

Conditioning and Independence

a. Here, note that

$$R_{XY} = G = \{(x, y) | x, y \in \mathbb{Z}, |x| + |y| \leq 2\}.$$

Thus, the joint PMF is given by

$$P_{XY}(x, y) = \begin{cases} \frac{1}{13} & (x, y) \in G \\ 0 & \text{otherwise} \end{cases}$$

To find the marginal PMF of X, $P_X(i)$, we use Equation 5.1. Thus,

$$\begin{aligned} P_X(-2) &= P_{XY}(-2, 0) = \frac{1}{13}, \\ P_X(-1) &= P_{XY}(-1, -1) + P_{XY}(-1, 0) + P_{XY}(-1, 1) = \frac{3}{13}, \\ P_X(0) &= P_{XY}(0, -2) + P_{XY}(0, -1) + P_{XY}(0, 0) \\ &\quad + P_{XY}(0, 1) + P_{XY}(0, 2) = \frac{5}{13}, \\ P_X(1) &= P_{XY}(1, -1) + P_{XY}(1, 0) + P_{XY}(1, -1) = \frac{3}{13}, \\ P_X(2) &= P_{XY}(2, 0) = \frac{1}{13}. \end{aligned}$$

Similarly, we can find

$$P_Y(j) = \begin{cases} \frac{1}{13} & \text{for } j = 2, -2 \\ \frac{3}{13} & \text{for } j = -1, 1 \\ \frac{5}{13} & \text{for } j = 0 \\ 0 & \text{otherwise} \end{cases}$$

We can write this in a more compact form as

$$P_X(k) = P_Y(k) = \frac{5 - 2|k|}{13}, \quad \text{for } k = -2, -1, 0, 1, 2.$$

b. For $i = -1, 0, 1$, we can write

$$\begin{aligned} P_{X|Y}(i|1) &= \frac{P_{XY}(i, 1)}{P_Y(1)} \\ &= \frac{\frac{1}{13}}{\frac{3}{13}} = \frac{1}{3}, \quad \text{for } i = -1, 0, 1. \end{aligned}$$

Thus, we conclude

$$P_{X|Y}(i|1) = \begin{cases} \frac{1}{3} & \text{for } i = -1, 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

By looking at the above conditional PMF, we conclude that, given $Y = 1$, X is uniformly distributed over the set $\{-1, 0, 1\}$.

c. X and Y are **not** independent. We can see this as the conditional PMF of X given $Y = 1$ (calculated above) is not the same as marginal PMF of X, $P_X(x)$.

Conditional Expectation

Given that we know event A has occurred, we can compute the conditional expectation of a random variable X, $E[X|A]$. Conditional expectation is similar to ordinary expectation. The only difference is that we replace the PMF by the conditional PMF. Specifically, we have

$$E[X|A] = \sum_{x_i \in R_X} x_i P_{X|A}(x_i).$$

Similarly, given that we have observed the value of random variable Y, we can compute the conditional expectation of X. Specifically, the conditional expectation of X given that $Y = y$ is

$$E[X|Y = y] = \sum_{x_i \in R_X} x_i P_{X|Y}(x_i|y).$$

Conditional Expectation of X:

$$E[X|A] = \sum_{x_i \in R_X} x_i P_{X|A}(x_i),$$
$$E[X|Y = y_j] = \sum_{x_i \in R_X} x_i P_{X|Y}(x_i|y_j)$$

Example 5.5 Let X and Y be the same as in Example 5.4.

a. Find $E[X|Y = 1|$.
b. Find $E[X| -1 < Y < 2]$.
c. Find $E[|X|| -1 < Y < 2]$.

Conditional Expectation

a. To find $E[X|Y=1]$, we have

$$E[X|Y=1]=\sum_{x_i\in R_X} x_i P_{X|Y}(x_i|1).$$

We found in Example 5.4 that given $Y=1$, X is uniformly distributed over the set $\{-1,0,1\}$. Thus, we conclude that

$$E[X|Y=1]=\frac{1}{3}(-1+0+1)=0.$$

b. To find $E[X|-1<Y<2]$, let A be the event that $-1<Y<2$, i.e., $Y\in\{0,1\}$. To find $E[X|A]$, we need to find the conditional PMF, $P_{X|A}(k)$, for $k=-2,-1,0,1,2$. First, note that

$$P(A)=P_Y(0)+P_Y(1)=\frac{5}{13}+\frac{3}{13}=\frac{8}{13}.$$

Thus, for $k=-2,1,0,1,2$, we have

$$P_{X|A}(k)=\frac{13}{8}P(X=k,A).$$

So, we can write

$$\begin{aligned}
P_{X|A}(-2)&=\frac{13}{8}P(X=-2,A)\\
&=\frac{13}{8}P_{XY}(-2,0)=\frac{1}{8},\\
P_{X|A}(-1)&=\frac{13}{8}P(X=-1,A)\\
&=\frac{13}{8}\big[P_{XY}(-1,0)+P_{XY}(-1,1)\big]=\frac{2}{8}=\frac{1}{4},\\
P_{X|A}(\ 0\ )&=\frac{13}{8}P(X=0,A)\\
&=\frac{13}{8}\big[P_{XY}(0,0)+P_{XY}(0,1)\big]=\frac{2}{8}=\frac{1}{4},\\
P_{X|A}(\ 1\ )&=\frac{13}{8}P(X=1,A)\\
&=\frac{13}{8}\big[P_{XY}(1,0)+P_{XY}(1,1)\big]=\frac{2}{8}=\frac{1}{4},\\
P_{X|A}(\ 2\ )&=\frac{13}{8}P(X=2,A)\\
&=\frac{13}{8}P_{XY}(2,0)=\frac{1}{8}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
E[X|A]&=\sum_{x_i\in R_X} x_i P_{X|A}(x_i)\\
&=(-2)\frac{1}{8}+(-1)\frac{1}{4}+(0)\frac{1}{4}+(1)\frac{1}{4}+(2)\frac{1}{8}=0.
\end{aligned}$$

c. To find $E\big[|X||-1<Y<2\big]$, we use the conditional PMF and LOTUS. We have

$$\begin{aligned}
E[|X||A]&=\sum_{x_i\in R_X} |x_i| P_{X|A}(x_i)\\
&=|-2|\cdot\frac{1}{8}+|-1|\cdot\frac{1}{4}+0\cdot\frac{1}{4}+1\cdot\frac{1}{4}+2\cdot\frac{1}{8}=1.
\end{aligned}$$

Law of Total Probability and Expectation

Law of Total Probability:

$$P(X \in A) = \sum_{y_j \in R_Y} P(X \in A | Y = y_j) P_Y(y_j), \quad \text{for any set } A.$$

Law of Total Expectation:

1. If $B_1, B_2, B_3, \ldots$ is a partition of the sample space S,

$$EX = \sum_i E[X|B_i] P(B_i) \tag{5.3}$$

2. For a random variable X and a discrete random variable Y,

$$EX = \sum_{y_j \in R_Y} E[X|Y = y_j] P_Y(y_j) \tag{5.4}$$

Example 5.6

Let $X \sim Geometric(p)$. Find EX by conditioning on the result of the first "coin toss."

Law of Total Probability and Expectation, example

Example 5.6

Let $X \sim Geometric(p)$. Find EX by conditioning on the result of the first "coin toss."

Solution

Remember that the random experiment behind $Geometric(p)$ is that we have a coin with $P(H) = p$. We toss the coin repeatedly until we observe the first heads. X is the total number of coin tosses. Now, there are two possible outcomes for the first coin toss: H or T. Thus, we can use the law of total expectation (Equation 5.3):

$$\begin{aligned} EX &= E[X|H]P(H) + E[X|T]P(T) \\ &= pE[X|H] + (1-p)E[X|T] \\ &= p \cdot 1 + (1-p)(EX+1). \end{aligned}$$

In this equation, $E[X|T] = 1 + EX$, because the tosses are independent, so if the first toss is tails, it is like starting over on the second toss. Solving for EX, we obtain

$$EX = \frac{1}{p}.$$

Law of Total Probability and Expectation, example

Example 5.7

Suppose that the number of customers visiting a fast food restaurant in a given day is $N \sim Poisson(\lambda)$. Assume that each customer purchases a drink with probability p, independently from other customers and independently from the value of N. Let X be the number of customers who purchase drinks. Find EX.

Solution

By the above information, we conclude that given $N = n$, then X is a sum of n independent $Bernoulli(p)$ random variables. Thus, given $N = n$, X has a binomial distribution with parameters n and p. We write

$$X|N = n \ \sim \ Binomial(n, p).$$

That is,

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

Thus, we conclude

$$E[X|N = n] = np.$$

Thus, using the law of total probability, we have

$$\begin{aligned} E[X] &= \sum_{n=0}^{\infty} E[X|N = n] P_N(n) \\ &= \sum_{n=0}^{\infty} np P_N(n) \\ &= p \sum_{n=0}^{\infty} n P_N(n) = pE[N] = p\lambda. \end{aligned}$$

Expectation Example

Consider again our example in which we toss a pair of fair, four-sided dice, in which one of the dice is RED and the other is **BLACK**. Again, letting:

- X = *the outcome on the RED die* = $\{1, 2, 3, 4\}$
- Y = the outcome on the **BLACK** die = $\{1, 2, 3, 4\}$

What is the mean of X ? And, what is the mean of Y?

Solution

The mean of X is calculated as:

$$\mu_X = E[X] = \sum_{x \in S_1} \sum_{y \in S_2} x f(x, y) = 1\left(\frac{1}{16}\right) + \cdots + 1\left(\frac{1}{16}\right) + \cdots + 4\left(\frac{1}{16}\right) + \cdots + 4\left(\frac{1}{16}\right)$$

which simplifies to:

$$\mu_X = E[X] = 1\left(\frac{4}{16}\right) + 2\left(\frac{4}{16}\right) + 3\left(\frac{4}{16}\right) + 4\left(\frac{4}{16}\right) = \frac{40}{16} = 2.5$$

The mean of Y is similarly calculated as:

$$\mu_Y = E[Y] = \sum_{x \in S_1} \sum_{y \in S_2} y f(x, y) = 1\left(\frac{1}{16}\right) + \cdots + 1\left(\frac{1}{16}\right) + \cdots + 4\left(\frac{1}{16}\right) + \cdots + 4\left(\frac{1}{16}\right)$$

which simplifies to:

$$\mu_Y = E[Y] = 1\left(\frac{4}{16}\right) + 2\left(\frac{4}{16}\right) + 3\left(\frac{4}{16}\right) + 4\left(\frac{4}{16}\right) = \frac{40}{16} = 2.5$$

Expectation Example

By the way, you probably shouldn't find it surprising that the formula for the mean of X reduces to:

$$\mu_X = \sum_{x \in S_1} x f(x)$$

because:

$$\mu_X = E(X) = \sum_{x \in S_1} \sum_{y \in S_2} x f(x,y) = \sum_{x \in S_1} x \sum_{y \in S_2} f(x,y) = \sum_{x \in S_1} x f(x)$$

That is, the third equality holds because the *x* values don't depend on y and therefore can be pulled through the summation over y. And, the last equality holds because of the definition of the marginal probability mass function of X. Similarly, the mean of Y reduces to:

$$\mu_Y = \sum_{y \in S_2} y f(y)$$

because:

$$\mu_Y = E(Y) = \sum_{y \in S_2} \sum_{x \in S_1} y f(x,y) = \sum_{y \in S_2} y \sum_{x \in S_1} f(x,y) = \sum_{y \in S_2} y f(y)$$

That is, again, the third equality holds because the *y* values don't depend on x and therefore can be pulled through the summation over x. And, the last equality holds because of the definition of the marginal probability mass function of Y.

Variance of Joint Probability

Now, suppose we were given a joint probability mass function $f(x,y)$, and we wanted to find the variance of X. Again, one strategy would be to find the marginal p.m.f of X first, and then use the definition of the expected value that we previously learned to calculate $\text{Var}(X)$. Alternatively, we could use the following definition of the variance that has been extended to accommodate joint probability mass functions.

Definition. Let X be a discrete random variable with support S_1, and let Y be a discrete random variable with support S_2. Let X and Y be discrete random variables with joint p.m.f. $f(x,y)$ on the support S. If $u(X,Y)$ is a function of these two random variables, then:

$$E[u(X,Y)] = \sum_{(x,y)\in S}\sum u(x,y)f(x,y)$$

if it exists, is called the **expected value** of $u(X,Y)$. If $u(X,Y)=(X-\mu_X)^2$, then:

$$\sigma_X^2 = Var[X] = \sum_{x\in S_1}\sum_{y\in S_2}(x-\mu_X)^2 f(x,y)$$

if it exists, is the **variance of** X. The variance of X can also be calculated using the shortcut formula:

$$\sigma_X^2 = E(X^2)-\mu_X^2 = \left(\sum_{x\in S_1}\sum_{y\in S_2} x^2 f(x,y)\right) - \mu_X^2$$

If $u(X,Y)=(Y-\mu_Y)^2$, then:

$$\sigma_Y^2 = Var[Y] = \sum_{x\in S_1}\sum_{y\in S_2}(y-\mu_Y)^2 f(x,y)$$

if it exists, is the **variance of** Y. The variance of Y can also be calculated using the shortcut formula:

$$\sigma_Y^2 = E(Y^2)-\mu_Y^2 = \left(\sum_{x\in S_1}\sum_{y\in S_2} y^2 f(x,y)\right) - \mu_Y^2$$

Variance Example

Consider yet again our example in which we toss a pair of fair, four-sided dice, in which one of the dice is RED and the other is **BLACK**. Again, letting:

- X = *the outcome on the RED die* = $\{1, 2, 3, 4\}$
- Y = the outcome on the **BLACK** die = $\{1, 2, 3, 4\}$

What is the variance of X ? And, what is the variance of Y?

Solution

Using the definition, the variance of X is calculated as:

$$\sigma_X^2 = \sum_{x \in S_1} \sum_{y \in S_2} (x - \mu_X)^2 f(x, y) = (1 - 2.5)^2 \left(\frac{1}{16}\right) + \cdots + (4 - 2.5)^2 \left(\frac{1}{16}\right) = 1.25$$

Thankfully, we get the same answer using the shortcut formula for the variance of X:

$$\sigma_X^2 = E(X^2) - \mu_X^2 = \left(\sum_{x \in S_1} \sum_{y \in S_2} x^2 f(x, y)\right) - \mu_X^2 = \left[1^2 \left(\frac{1}{16}\right) + \cdots + 4^2 \left(\frac{1}{16}\right)\right] - 2.5^2 = \frac{120}{16} - 6.25 = 1.25$$

Calculating the variance of Y is left for you as an exercise. You should, because of the symmetry, also get $\text{Var}(Y) = 1.25$.

Example of joint probability of 2 variables

Consider two random variables X and Y with joint PMF given in Table 5.1.

Table 5.1 Joint PMF of X and Y in Example 5.1

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$
$X=1$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$

Figure 5.1 shows $P_{XY}(x,y)$.

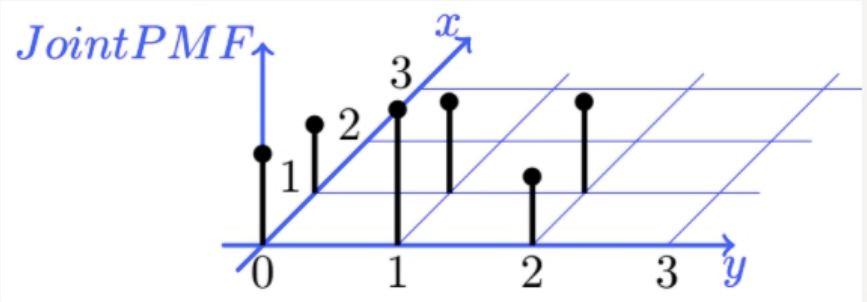


Figure 5.1: Joint PMF of X and Y (Example 5.1).

a. Find $P(X=0, Y \leq 1)$.
b. Find the marginal PMFs of X and Y.
c. Find $P(Y=1|X=0)$.
d. Are X and Y independent?

a. To find $P(X=0, Y \leq 1)$, we can write

$$P(X=0, Y \leq 1) = P_{XY}(0,0) + P_{XY}(0,1) = \frac{1}{6} + \frac{1}{4} = \frac{5}{12}.$$

b. Note that from the table,

$$R_X = \{0,1\} \qquad \text{and} \qquad R_Y = \{0,1,2\}.$$

Now we can use Equation 5.1 to find the marginal PMFs. For example, to find $P_X(0)$, we can write

$$\begin{aligned} P_X(0) &= P_{XY}(0,0) + P_{XY}(0,1) + P_{XY}(0,2) \\ &= \frac{1}{6} + \frac{1}{4} + \frac{1}{8} \\ &= \frac{13}{24}. \end{aligned}$$

We obtain

$$P_X(x) = \begin{cases} \frac{13}{24} & x=0 \\ \frac{11}{24} & x=1 \\ 0 & \text{otherwise} \end{cases}$$

$$P_Y(y) = \begin{cases} \frac{7}{24} & y=0 \\ \frac{5}{12} & y=1 \\ \frac{7}{24} & y=2 \\ 0 & \text{otherwise} \end{cases}$$

Example of joint probability of 2 variables

Consider two random variables X and Y with joint PMF given in Table 5.1.

Table 5.1 Joint PMF of X and Y in Example 5.1

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$
$X=1$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$

Figure 5.1 shows $P_{XY}(x,y)$.

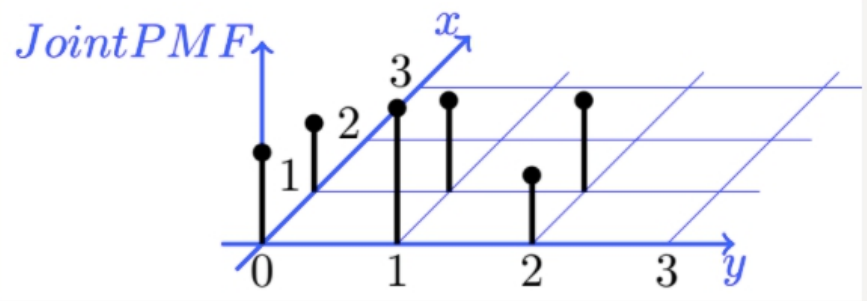


Figure 5.1: Joint PMF of X and Y (Example 5.1).

a. Find $P(X=0, Y\leq 1)$.
b. Find the marginal PMFs of X and Y.
c. Find $P(Y=1|X=0)$.
d. Are X and Y independent?

c. Find $P(Y=1|X=0)$: Using the formula for conditional probability, we have

$$\begin{aligned} P(Y=1|X=0) &= \frac{P(X=0,Y=1)}{P(X=0)} \\ &= \frac{P_{XY}(0,1)}{P_X(0)} \\ &= \frac{\frac{1}{4}}{\frac{13}{24}} = \frac{6}{13}. \end{aligned}$$

d. Are X and Y independent? X and Y are not independent, because as we just found out

$$P(Y=1|X=0) = \frac{6}{13} \neq P(Y=1) = \frac{5}{12}.$$

Caution: If we want to show that X and Y are independent, we need to check that $P(X=x_i, Y=y_j) = P(X=x_i)P(Y=y_j)$, for all $x_i \in R_X$ and all $y_j \in R_Y$. Thus, even if in the above calculation we had found $P(Y=1|X=0) = P(Y=1)$, we would not yet have been able to conclude that X and Y are independent. For that, we would need to check the independence condition for all $x_i \in R_X$ and all $y_j \in R_Y$.

https://www.probabilitycourse.com/chapter5/5_1_1_joint_pmf.php

Joint Cumulative Distributive Function (CDF)

The **joint cumulative distribution function** of two random variables X and Y is defined as

$$F_{XY}(x,y) = P(X \le x, Y \le y).$$

As usual, comma means "and," so we can write

$$\begin{aligned} F_{XY}(x,y) &= P(X \le x, Y \le y) \\ &= P\big((X \le x) \text{ and } (Y \le y)\big) = P\big((X \le x) \cap (Y \le y)\big). \end{aligned}$$

Figure 5.2 shows the region associated with $F_{XY}(x,y)$ in the two-dimensional plane. Note that the above definition of joint CDF is a general definition and is applicable to discrete, continuous, and mixed random variables. Since the joint CDF refers to the probability of an event, we must have $0 \le F_{XY}(x,y) \le 1$.

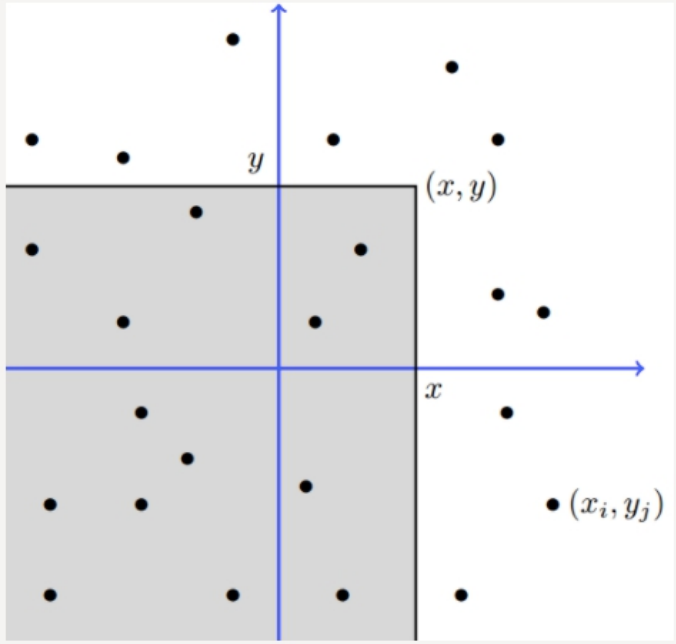


Figure 5.2: $F_{XY}(x,y)$ is the probability that (X,Y) belongs to the shaded region. The dots are the pairs (x_i, y_j) in R_{XY}.

If we know the joint CDF of X and Y, we can find the *marginal* CDFs, $F_X(x)$ and $F_Y(y)$. Specifically, for any $x \in \mathbb{R}$, we have

$$\begin{aligned} F_{XY}(x,\infty) &= P(X \le x, Y \le \infty) \\ &= P(X \le x) = F_X(x). \end{aligned}$$

Here, by $F_{XY}(x,\infty)$, we mean $\lim_{y\to\infty} F_{XY}(x,y)$. Similarly, for any $y \in \mathbb{R}$, we have

$$F_Y(y) = F_{XY}(\infty, y).$$

Marginal CDFs of X and Y:

$$\begin{aligned} F_X(x) &= F_{XY}(x,\infty) = \lim_{y\to\infty} F_{XY}(x,y), \qquad \text{for any } x, \\ F_Y(y) &= F_{XY}(\infty,y) = \lim_{x\to\infty} F_{XY}(x,y), \qquad \text{for any } y \end{aligned} \tag{5.2}$$

Also, note that we must have

$$\begin{aligned} &F_{XY}(\infty,\infty) = 1, \\ &F_{XY}(-\infty,y) = 0, \qquad \text{for any } y, \\ &F_{XY}(x,-\infty) = 0, \qquad \text{for any } x. \end{aligned}$$

Example 5.2

Let $X \sim Bernoulli(p)$ and $Y \sim Bernoulli(q)$ be independent, where $0 < p, q < 1$. Find the joint PMF and joint CDF for X and Y.

Joint Cumulative Distributive Function (CDF)

Example 5.2
Let $X \sim Bernoulli(p)$ and $Y \sim Bernoulli(q)$ be independent, where $0 < p, q < 1$. Find the joint PMF and joint CDF for X and Y.

First note that the joint range of X and Y is given by

$$R_{XY} = \{(0,0), (0,1), (1,0), (1,1)\}.$$

Since X and Y are independent, we have

$$P_{XY}(i,j) = P_X(i)P_Y(j), \qquad \text{for } i,j = 0,1.$$

Thus, we conclude

$$\begin{aligned}
P_{XY}(0,0) &= P_X(0)P_Y(0) = (1-p)(1-q),\\
P_{XY}(0,1) &= P_X(0)P_Y(1) = (1-p)q,\\
P_{XY}(1,0) &= P_X(1)P_Y(0) = p(1-q),\\
P_{XY}(1,1) &= P_X(1)P_Y(1) = pq.
\end{aligned}$$

Now that we have the joint PMF, we can find the joint CDF

$$F_{XY}(x,y) = P(X \le x, Y \le y).$$

Specifically, since $0 \le X, Y \le 1$, we conclude

$$\begin{aligned}
F_{XY}(x,y) &= 0, \qquad &&\text{if } x < 0,\\
F_{XY}(x,y) &= 0, \qquad &&\text{if } y < 0,\\
F_{XY}(x,y) &= 1, \qquad &&\text{if } x \ge 1 \text{ and } y \ge 1.
\end{aligned}$$

Now, for $0 \le x < 1$ and $y \ge 1$, we have

$$\begin{aligned}
F_{XY}(x,y) &= P(X \le x, Y \le y)\\
&= P(X = 0, y \le 1)\\
&= P(X = 0) = 1 - p.
\end{aligned}$$

Similarly, for $0 \le y < 1$ and $x \ge 1$, we have

$$\begin{aligned}
F_{XY}(x,y) &= P(X \le x, Y \le y)\\
&= P(X \le 1, y = 0)\\
&= P(Y = 0) = 1 - q.
\end{aligned}$$

Finally, for $0 \le x < 1$ and $0 \le y < 1$, we have

$$\begin{aligned}
F_{XY}(x,y) &= P(X \le x, Y \le y)\\
&= P(X = 0, y = 0)\\
&= P(X = 0)P(Y = 0) = (1-p)(1-q).
\end{aligned}$$

Figure 5.3 shows the values of $F_{XY}(x,y)$ in different regions of the two-dimensional plane. Note that, in general, we actually need a three-dimensional graph to show a joint CDF of two random variables, i.e., we need three axes: x, y, and $z = F_{XY}(x,y)$. However, because the random variables of this example are simple, and can take only two values, a two-dimensional figure suffices.

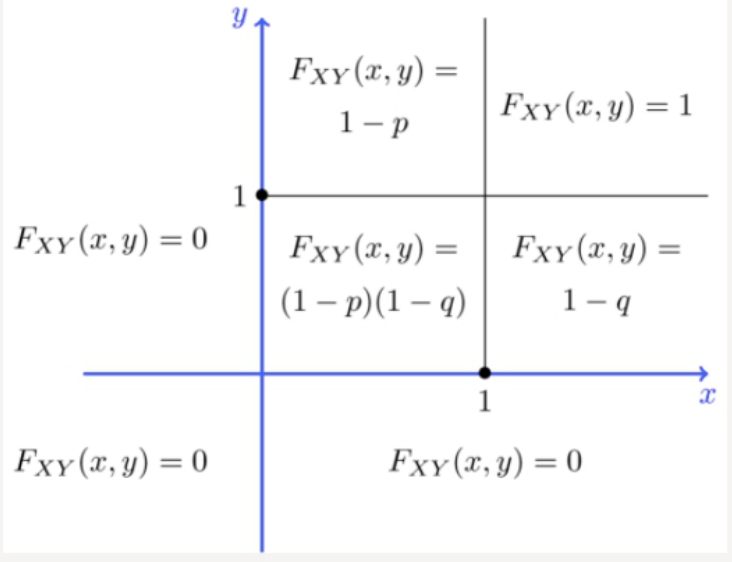


Figure 5.3 Joint CDF for X and Y in Example 5.2

Conditional Expectation as a Function of a Random Variable

Remember that the conditional expectation of X given that $Y = y$ is given by

$$E[X|Y = y] = \sum_{x_i \in R_X} x_i P_{X|Y}(x_i|y).$$

Note that $E[X|Y = y]$ depends on the value of y. In other words, by changing y, $E[X|Y = y]$ can also change. Thus, we can say $E[X|Y = y]$ is a function of y, so let's write

$$g(y) = E[X|Y = y].$$

Thus, we can think of $g(y) = E[X|Y = y]$ as a function of the value of random variable Y. We then write

$$g(Y) = E[X|Y].$$

We use this notation to indicate that $E[X|Y]$ is a random variable whose value equals $g(y) = E[X|Y = y]$ when $Y = y$. Thus, if Y is a random variable with range $R_Y = \{y_1, y_2, \cdots\}$, then $E[X|Y]$ is also a random variable with

$$E[X|Y] = \begin{cases} E[X|Y = y_1] & \text{with probability } P(Y = y_1) \\ E[X|Y = y_2] & \text{with probability } P(Y = y_2) \\ \quad . & \quad . \\ \quad . & \quad . \\ \quad . & \quad . \end{cases}$$

Conditional Expectation as a Function of a Random Variable

Example 5.10 Let $X = aY + b$. Then $E[X|Y = y] = E[aY + b|Y = y] = ay + b$. Here, we have $g(y) = ay + b$, and therefore,

$$E[X|Y] = aY + b,$$

which is a function of the random variable Y.

Since $E[X|Y]$ is a random variable, we can find its PMF, CDF, variance, etc. Let's look at an example to better understand $E[X|Y]$.

Example 5.11 Consider two random variables X and Y with joint PMF given in Table 5.2. Let $Z = E[X|Y]$.

a. Find the Marginal PMFs of X and Y.
b. Find the conditional PMF of X given $Y = 0$ and $Y = 1$, i.e., find $P_{X|Y}(x|0)$ and $P_{X|Y}(x|1)$.
c. Find the PMF of Z.
d. Find EZ, and check that $EZ = EX$.
e. Find Var(Z).

Table 5.2: Joint PMF of X and Y in example 5.11

	$Y = 0$	$Y = 1$
$X = 0$	$\frac{1}{5}$	$\frac{2}{5}$
$X = 1$	$\frac{2}{5}$	0

Conditional Expectation as a Function of a Random Variable

a. Using the table we find out

$$P_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5},$$
$$P_X(1) = \frac{2}{5} + 0 = \frac{2}{5},$$
$$P_Y(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5},$$
$$P_Y(1) = \frac{2}{5} + 0 = \frac{2}{5}.$$

Thus, the marginal distributions of X and Y are both $Bernoulli(\frac{2}{5})$. However, note that X and Y are not independent.

b. We have

$$P_{X|Y}(0|0) = \frac{P_{XY}(0,0)}{P_Y(0)}$$
$$= \frac{\frac{1}{5}}{\frac{3}{5}} = \frac{1}{3}.$$

Thus,

$$P_{X|Y}(1|0) = 1 - \frac{1}{3} = \frac{2}{3}.$$

We conclude

$$X|Y=0 \sim Bernoulli\left(\frac{2}{3}\right).$$

Similarly, we find

$$P_{X|Y}(0|1) = 1,$$
$$P_{X|Y}(1|1) = 0.$$

Thus, given $Y = 1$, we have always $X = 0$.

c. We note that the random variable Y can take two values: 0 and 1. Thus, the random variable $Z = E[X|Y]$ can take two values as it is a function of Y. Specifically,

$$Z = E[X|Y] = \begin{cases} E[X|Y=0] & \text{if } Y = 0 \\ E[X|Y=1] & \text{if } Y = 1 \end{cases}$$

Now, using the previous part, we have

$$E[X|Y=0] = \frac{2}{3}, \quad E[X|Y=1] = 0,$$

and since $P(y=0) = \frac{3}{5}$, and $P(y=1) = \frac{2}{5}$, we conclude that

$$Z = E[X|Y] = \begin{cases} \frac{2}{3} & \text{with probability } \frac{3}{5} \\ 0 & \text{with probability } \frac{2}{5} \end{cases}$$

So we can write

$$P_Z(z) = \begin{cases} \frac{3}{5} & \text{if } z = \frac{2}{3} \\ \frac{2}{5} & \text{if } z = 0 \\ 0 & \text{otherwise} \end{cases}$$

d. Now that we have found the PMF of Z, we can find its mean and variance. Specifically,

$$E[Z] = \frac{2}{3} \cdot \frac{3}{5} + 0 \cdot \frac{2}{5} = \frac{2}{5}.$$

We also note that $EX = \frac{2}{5}$. Thus, here we have

$$E[X] = E[Z] = E[E[X|Y]].$$

Conditional Expectation as a Function of a Random Variable

e. To find $\text{Var}(Z)$, we write

$$\begin{aligned}\text{Var}(Z) &= E[Z^2] - (EZ)^2 \\ &= E[Z^2] - \frac{4}{25},\end{aligned}$$

where

$$E[Z^2] = \frac{4}{9} \cdot \frac{3}{5} + 0 \cdot \frac{2}{5} = \frac{4}{15}.$$

Thus,

$$\begin{aligned}\text{Var}(Z) &= \frac{4}{15} - \frac{4}{25} \\ &= \frac{8}{75}.\end{aligned}$$

Example 5.12

Let X and Y be two random variables and g and h be two functions. Show that

$$E[g(X)h(Y)|X] = g(X)E[h(Y)|X].$$

Conditional Expectation as a Function of a Random Variable

Example 5.12

Let X and Y be two random variables and g and h be two functions. Show that

$$E[g(X)h(Y)|X] = g(X)E[h(Y)|X].$$

Solution

Note that $E[g(X)h(Y)|X]$ is a random variable that is a function of X. In particular, if $X = x$, then $E[g(X)h(Y)|X] = E[g(X)h(Y)|X = x]$. Now, we can write

$$\begin{aligned} E[g(X)h(Y)|X = x] &= E[g(x)h(Y)|X = x] \\ &= g(x)E[h(Y)|X = x] \qquad (\text{since } g(x) \text{ is a constant}). \end{aligned}$$

Thinking of this as a function of the random variable X, it can be rewritten as $E[g(X)h(Y)|X] = g(X)E[h(Y)|X]$. This rule is sometimes called "taking out what is known." The idea is that, given X, $g(X)$ is a known quantity, so it can be taken out of the conditional expectation.

$$E[g(X)h(Y)|X] = g(X)E[h(Y)|X] \qquad (5.6)$$

Iterated Expectations

Let us look again at the law of total probability for expectation. Assuming $g(Y) = E[X|Y]$, we have

$$\begin{aligned} E[X] &= \sum_{y_j \in R_Y} E[X|Y=y_j]P_Y(y_j) \\ &= \sum_{y_j \in R_Y} g(y_j)P_Y(y_j) \\ &= E[g(Y)] \qquad \text{by LOTUS (Equation 5.2)} \\ &= E[E[X|Y]]. \end{aligned}$$

Thus, we conclude

$$E[X] = E[E[X|Y]]. \qquad (5.7)$$

This equation might look a little confusing at first, but it is just another way of writing the law of total expectation (Equation 5.4). To better understand it, let's solve Example 5.7 using this terminology. In that example, we want to find EX. We can write

$$\begin{aligned} E[X] &= E[E[X|N]] \\ &= E[Np] \qquad (\text{since } X|N \sim Binomial(N,p)) \\ &= pE[N] = p\lambda. \end{aligned}$$

Equation 5.7 is called the *law of iterated expectations*. Since it is basically the same as Equation 5.4, it is also called the law of total expectation [3].

> Law of Iterated Expectations: $E[X] = E[E[X|Y]]$

Conditional Variance

Similar to the conditional expectation, we can define the conditional variance of X, $\text{Var}(X|Y=y)$, which is the variance of X in the conditional space where we know $Y=y$. If we let $\mu_{X|Y}(y) = E[X|Y=y]$, then

$$\begin{aligned}\text{Var}(X|Y=y) &= E\big[(X-\mu_{X|Y}(y))^2|Y=y\big] \\ &= \sum_{x_i \in R_X} \big(x_i-\mu_{X|Y}(y)\big)^2 P_{X|Y}(x_i) \\ &= E\big[X^2|Y=y\big]-\mu_{X|Y}(y)^2.\end{aligned}$$

Note that $\text{Var}(X|Y=y)$ is a function of y. Similar to our discussion on $E[X|Y=y]$ and $E[X|Y]$, we define $\text{Var}(X|Y)$ as a function of the random variable Y. That is, $\text{Var}(X|Y)$ is a random variable whose value equals $\text{Var}(X|Y=y)$ whenever $Y=y$. Let us look at an example.

Example 5.13

Let X, Y, and $Z=E[X|Y]$ be as in Example 5.11. Let also $V=\text{Var}(X|Y)$.

a. Find the PMF of V.
b. Find EV.
c. Check that $\text{Var}(X)=E(V)+\text{Var}(Z)$.

Conditional Variance, example

In Example 5.11, we found out that $X, Y \sim Bernoulli(\frac{2}{5})$. We also obtained

$$\begin{aligned} X|Y=0 &\sim Bernoulli\left(\frac{2}{3}\right), \\ P(X=0|Y=1) &= 1, \\ \mathrm{Var}(Z) &= \frac{8}{75}. \end{aligned}$$

a. To find the PMF of V, we note that V is a function of Y. Specifically,

$$V = \mathrm{Var}(X|Y) = \begin{cases} \mathrm{Var}(X|Y=0) & \text{if } Y=0 \\ \mathrm{Var}(X|Y=1) & \text{if } Y=1 \end{cases}$$

Therefore,

$$V = \mathrm{Var}(X|Y) = \begin{cases} \mathrm{Var}(X|Y=0) & \text{with probability } \frac{3}{5} \\ \mathrm{Var}(X|Y=1) & \text{with probability } \frac{2}{5} \end{cases}$$

Now, since $X|Y=0 \sim Bernoulli\left(\frac{2}{3}\right)$, we have

$$\mathrm{Var}(X|Y=0) = \frac{2}{3}\cdot\frac{1}{3} = \frac{2}{9},$$

and since given $Y=1$, $X=0$, we have

$$\mathrm{Var}(X|Y=1) = 0.$$

Thus,

$$V = \mathrm{Var}(X|Y) = \begin{cases} \frac{2}{9} & \text{with probability } \frac{3}{5} \\ 0 & \text{with probability } \frac{2}{5} \end{cases}$$

So we can write

$$P_V(v) = \begin{cases} \frac{3}{5} & \text{if } v = \frac{2}{9} \\ \frac{2}{5} & \text{if } v = 0 \\ 0 & \text{otherwise} \end{cases}$$

b. To find EV, we write

$$EV = \frac{2}{9}\cdot\frac{3}{5} + 0\cdot\frac{2}{5} = \frac{2}{15}.$$

c. To check that $\mathrm{Var}(X) = E(V) + \mathrm{Var}(Z)$, we just note that

$$\begin{aligned} \mathrm{Var}(X) &= \frac{2}{5}\cdot\frac{3}{5} = \frac{6}{25}, \\ EV &= \frac{2}{15}, \\ \mathrm{Var}(Z) &= \frac{8}{75}. \end{aligned}$$

Law of Total Variance

In the above example, we checked that $\text{Var}(X) = E(V) + \text{Var}(Z)$, which says

$$\text{Var}(X) = E(\text{Var}(X|Y)) + \text{Var}(E[X|Y]).$$

It turns out this is true in general and it is called *the law of total variance*, or *variance decomposition formula* [3]. Let us first prove the law of total variance, and then we explain it intuitively. Note that if $V = \text{Var}(X|Y)$, and $Z = E[X|Y]$, then

$$\begin{aligned} V &= E[X^2|Y] - (E[X|Y])^2 \\ &= E[X^2|Y] - Z^2. \end{aligned}$$

Thus,

$$\begin{aligned} EV &= E[E[X^2|Y]] - E[Z^2] \\ &= E[X^2] - E[Z^2] \qquad \big(\text{law of iterated expectations(Equation 5.7)}\big) \end{aligned} \tag{5.8}$$

Next, we have

$$\begin{aligned} \text{Var}(Z) &= E[Z^2] - (EZ)^2 \\ &= E[Z^2] - (EX)^2 \qquad (\text{law of iterated expectations}) \end{aligned} \tag{5.9}$$

Combining Equations 5.8 and 5.9, we obtain the law of total variance.

Law of Total Variance:

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y]) \tag{5.10}$$

Law of Total Variance, example

Example 5.14

Let N be the number of customers that visit a certain store in a given day. Suppose that we know $E[N]$ and Var (N). Let X_i be the amount that the ith customer spends on average. We assume X_i's are independent of each other and also independent of N. We further assume they have the same mean and variance

$$EX_i = EX,$$
$$\mathrm{Var}(X_i) = \mathrm{Var}(X).$$

Let Y be the store's total sales, i.e.,

$$Y = \sum_{i=1}^{N} X_i.$$

Find EY and Var(Y).

Law of Total Variance, example

To find EY, we cannot directly use the linearity of expectation because N is random. But, conditioned on $N = n$, we can use linearity and find $E[Y|N = n]$; so, we use the law of iterated expectations:

$$\begin{aligned}
EY &= E[E[Y|N]] && \text{(law of iterated expectations)}\\
&= E\left[E\left[\sum_{i=1}^{N} X_i|N\right]\right]\\
&= E\left[\sum_{i=1}^{N} E[X_i|N]\right] && \text{(linearity of expectation)}\\
&= E\left[\sum_{i=1}^{N} E[X_i]\right] && (X_i\text{'s and } N \text{ are indpendent})\\
&= E[NE[X]] && (\text{since } EX_i = EX\text{s})\\
&= E[X]E[N] && (\text{since } EX \text{ is not random}).
\end{aligned}$$

To find $\text{Var}(Y)$, we use the law of total variance:

$$\begin{aligned}
\text{Var}(Y) &= E(\text{Var}(Y|N)) + \text{Var}(E[Y|N])\\
&= E(\text{Var}(Y|N)) + \text{Var}(NEX) && \text{(as above)}\\
&= E(\text{Var}(Y|N)) + (EX)^2\text{Var}(N) && (5.12)
\end{aligned}$$

To find $E(\text{Var}(Y|N))$, note that, given $N = n$, Y is a sum of n independent random variables. As we discussed before, for n independent random variables, the variance of the sum is equal to sum of the variances. This fact is officially proved in Section 5.3 and also in Chapter 6, but we have occasionally used it as it simplifies the analysis. Thus, we can write

$$\begin{aligned}
\text{Var}(Y|N) &= \sum_{i=1}^{N} \text{Var}(X_i|N)\\
&= \sum_{i=1}^{N} \text{Var}(X_i) && (\text{since } X_i\text{'s are independent of } N)\\
&= NVar(X).
\end{aligned}$$

Thus, we have

$$E(\text{Var}(Y|N)) = ENVar(X) \qquad (5.13)$$

Combining Equations 5.12 and 5.13, we obtain

$$\text{Var}(Y) = ENVar(X) + (EX)^2Var(N).$$

For exercises please have a look:
https://www.probabilitycourse.com/chapter5/5_1_6_solved_prob.php

How to make decision in the prsence of uncertainty?

Decision theory is the science of making decisions. Contributions have come from a number of areas: philosophy, psychology, statistics, the social sciences, political science, economics, ... It has been mostly developed during the 20th century:

> Bernoulli (1738), de Finetti (1937), Ramsay (1931), Wald (1950), Lindley (1953), Savage (1954), Ferguson (1967), DeGroot (1970), Berger (1980,1985), Bernardo and Smith (1994), Robert (1994), ...

We make decisions all the time.

- The umbrella question—do I take umbrella today or not?
- Pregnancy and the risk of Down's syndrome.
- Should I fix or float my mortgage?
- Civil defence alerts.

How to make decision in the prsence of uncertainty?

Hypothetically, we could apply decision theory to any of these, but in many situations the difficulties may out-weigh the benefits.

In statistics, decision theory is concerned with quantifying the decision making process. We must choose a particular course of action from a set of possible decisions. The consequences of each possible decision depends on the "true" **state of nature** which we don't know. We collect data which provides information about the true state of nature and based on this we make a decision.

A **decision matrix** is often used to organise information. Consequences of alternative courses of action are tabulated against the possible "states of nature." For example,

	State of Nature: Size of Tsunami				
Action	0–0.5m	0.5–1m	1–2m	...	>5m
Issue Advisory					
Issue Warning					
Close Beaches					
Evacuation					

How to make decision in the prsence of uncertainty?

Example 7.1 (Game Show)**.** As part of a TV game show, suppose you need to choose between one of the following two options:

- Option A
  - win \$100,000 — probability $= 1$
- Option B
  - win \$500,000 — probability $= 0.10$
  - win \$100,000 — probability $= 0.89$
  - win \$0 — probability $= 0.01$

What would you do?

Now suppose you need to choose between one of the following two options:

- Option C
  - win \$100,000 — probability $= 0.11$
  - win \$0 — probability $= 0.89$
- Option D
  - win \$500,000 — probability $= 0.10$
  - win \$0 — probability $= 0.90$

What would you do? ◊

How to make decision in the prsence of uncertainty?

Example 7.1 (Game Show)**.** As part of a TV game show, suppose you need to choose between one of the following two options:

- Option A
  - win \$100,000 — probability $= 1$
- Option B
  - win \$500,000 — probability $= 0.10$
  - win \$100,000 — probability $= 0.89$
  - win \$0 — probability $= 0.01$

What would you do?

Now suppose you need to choose between one of the following two options:

- Option C
  - win \$100,000 — probability $= 0.11$
  - win \$0 — probability $= 0.89$
- Option D
  - win \$500,000 — probability $= 0.10$
  - win \$0 — probability $= 0.90$

What would you do? ◊

Elements of Decision Theory

Rice describes decision theory as "a mathematical approach for making decisions in face of uncertainty." Key elements are:

(a) An **action** a is chosen from a set of possible actions $\mathcal{A}$.

(b) This choice is based on observations from a **random variable** X which has a probability distribution that depends on the state of nature Θ.

(c) A **loss function** evaluates the consequences of the possible actions given the state of nature.

(d) A **decision rule** that maps the sample space of X into $\mathcal{A}$. Given a set of observations the decision rule identifies which action will be taken.

(e) The idea is to identify the **optimal decision rule**.

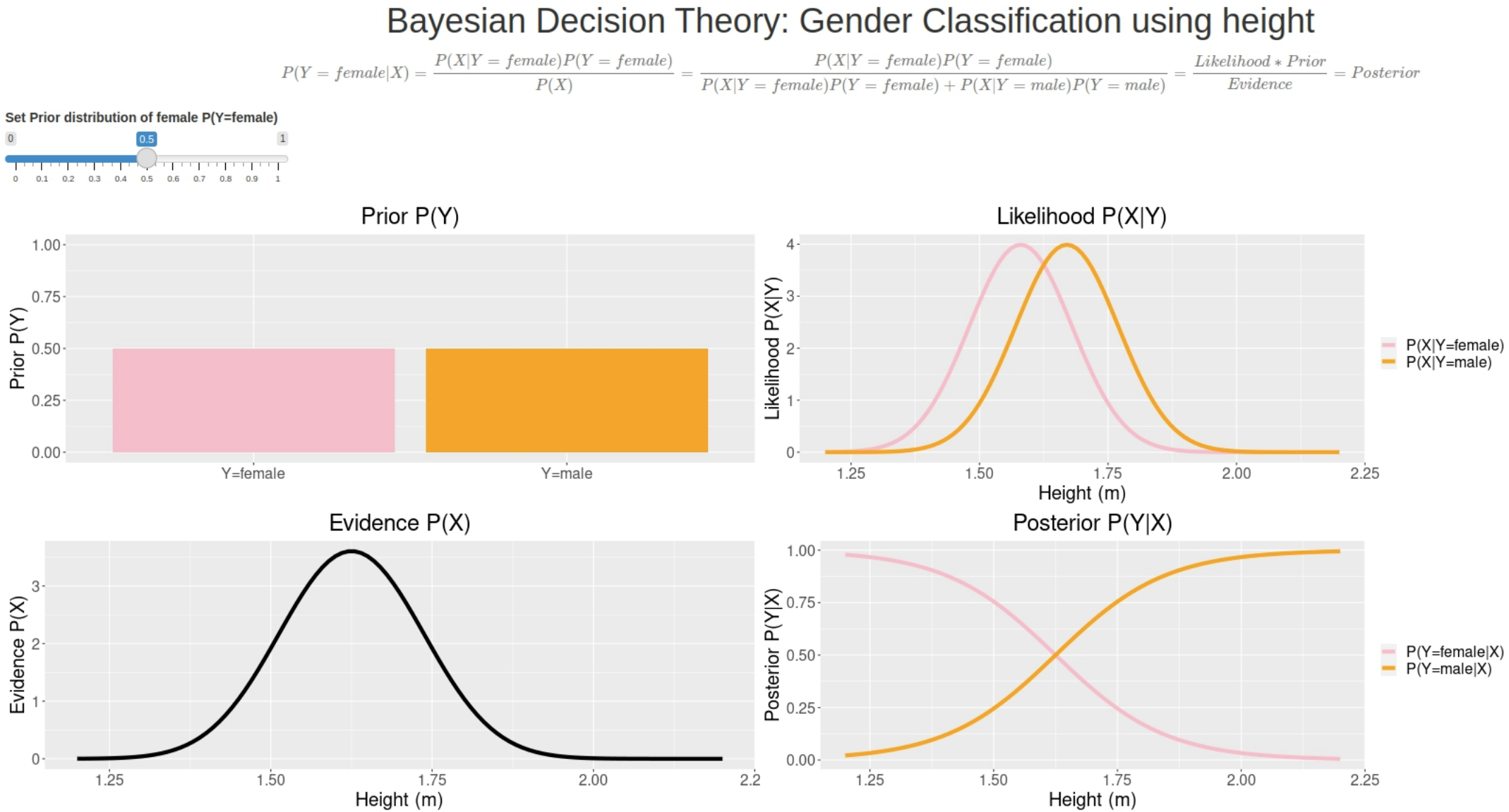
Toy example: Predict gender of a person based on height

- Setting: binary classification, that is Y = {male, female} = {M, F} = {-1, 1} and X is height of a person
- The joint density p(x, y) of the probability measure P on X × Y can be decomposed as follows
  - The **class-conditional density** or **likelihood** p(x | y). It models the occurrence of the features x of class y.
  - The **conditional probability** p(y | x) or **Posterior**. The probability that we observe y given that the input is x. The most probable class y for the features x is then used for prediction.
  - The **marginal distribution or evidence** p(x). It models the cumulated occurrence of features x over all classes.
  - The class probabilities p(y). The total probability of a class y or **Prior**.

$$P(Y = female|X) = \frac{P(X|Y = female)P(Y = female)}{P(X)} = \frac{Likelihood * Prior}{Evidence} = Posterior$$

$$P(X) = P(X|Y = female)P(Y = female) + P(X|Y = male)P(Y = male)$$

The effect of Prior on Posterior



How to decide male of female?

1. Using Prior: P(Y=female) > P(Y=male) -> female
2. Using Likelihood or Class conditional Probability with Maximum Likelihood (ML)

$$\hat{y} = \max_y P(x|y)$$

$P(x|\hat{y}) >= P(x|y)$

P(x|y=male) > P(x|y=female) -> y = male

P(x|y=male) < P(x|y=female) -> y = female

$\log \frac{P(x|y=male)}{P(x|y=female)} > 0$ (log-likelihood test)

But what if female are more likely than male in Nu Vuong country -> take into account the Prior propability P(Y=male) and P(Y=female)

$$P(Y = female|X) = \frac{P(X|Y = female)P(Y = female)}{P(X)} = \frac{Likelihood * Prior}{Evidence} = Posterior$$

How to decide male of female?

3. Using Posteria with Maximum A Posterior (MAP) of Bayes Rule

$$\hat{y} = \max_{y} P(y|x)$$

$$P(Y = female|X) = \frac{P(X|Y = female)P(Y = female)}{P(X)} = \frac{Likelihood * Prior}{Evidence} = Posterior$$

P(x|$\hat{y}$) >= P(x|y)

P(y=male|x) > P(y=female|x) -> y = male

P(y=male|x) < P(y=female|x) -> y = female

4. What does it cost if you make a mistake?

i.e. suppose you decide y = male, but really y = female

i.e. you may pay a big penalty if you decide it is a female when it is a male

Summary: How to decide male of female?

- Likelihood function: $P(x \mid y)$ $\quad x \in X, y \in Y$
- Prior $P(Y)$
- Posterior $P(y \mid x)$
- Decision rule $\alpha(x)$ $\quad \alpha(x) \in Y$
- Loss function $L(\alpha(x), y)$: cost of making decision $\alpha(x)$ if true state is Y
  - All errors penalisied the same
    - $L(\alpha(x), y) = 0$, if $\alpha(x) = y$
    - $L(\alpha(x), y) = 1$, if $\alpha(x) \neq y$
  - All errors penalisied NOT the same
    - $L(\alpha(x), y) = 0$, if $\alpha(x) = y$
    - $L(\alpha(x)=1, y=-1) = 10$
    - $L(\alpha(x)=-1, y=1) = 1000000$

Summary: How to decide male of female?

❖Risk: the risk of the decision rule α(x) is the expected loss

❖$R(\alpha) = E[L(\alpha(x), y)] = \sum_{x,y} L(\alpha(x), y)$P(x, y)

❖Bayes Decision Theory says "Pick the decision rule $\hat{\alpha}$(x) which minimize the risk"

$$\hat{\alpha} = Arg\min_{\alpha \in A} R(\alpha)$$

-> $R(\hat{\alpha}) \geq R(\alpha)\ \forall \alpha \in A$ (Exercise)

- A = set of all decision rules
- $\hat{\alpha}$ is Bayes Decision
- $R(\hat{\alpha})$ is Bayes Risk

Summary: How to decide male of female?

Bayes risk is the best you can do if

- (a) You know P(x|y)P(y) and Loss function
- (b) You can compute $\hat{\alpha} = Arg\min_{\alpha \in A} R(\alpha)$
- (c) You can afford the losses (e.g. gambling, poker)
- (d) you can make decision for a sequence of data $x_1, x_2 \ldots x_n$ with states $y_1, y_2 \ldots y_n$where each point (x_i, y_i) are independently identically distributed from P(x,y)

Bayesian Decision Theory with rules, example

Example 7.6 (Milleritis)**.** We will apply Bayesian decision theory to the problem of diagnosing and treating a fictional disease which we will call milleritis. Suppose the level of uric acid (X) in the blood is affected by this disease as follows:

- Patients without milleritis: $X \sim \mathrm{N}(\mu = 6, \sigma = 1)$ mg/100ml blood
- Patients with milleritis: $X \sim \mathrm{N}(\mu = 10, \sigma = 1.5)$ mg/100ml blood

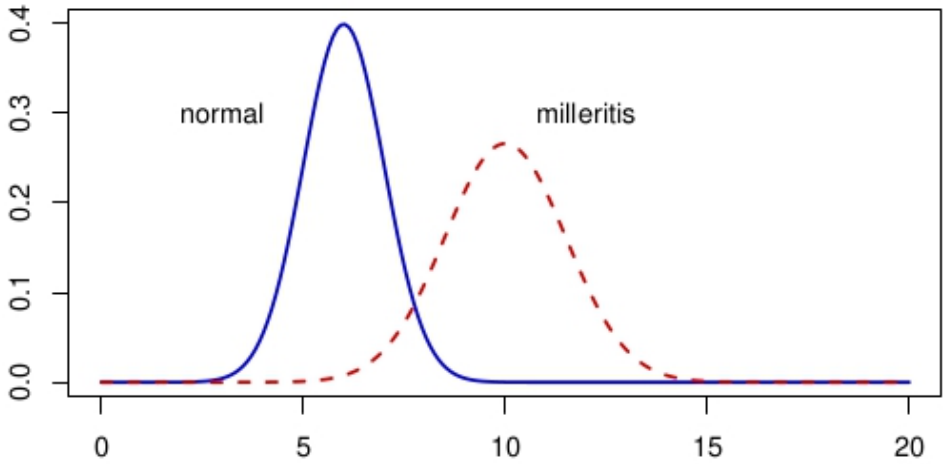


Figure 7.2: Densities of "normal" *vs.* milleritis.

Bayesian Decision Theory with rules, example

For a measured value of X, we can use the following version of Bayes theorem to find the **conditional probability** of a patient being milleritis positive (M^+):

$$\mathrm{P}(M^+|X=x) = \frac{f(x|M^+) \times \mathrm{P}(M^+)}{f(x|M^+) \times \mathrm{P}(M^+) + f(x|M^-) \times \mathrm{P}(M^-)}$$

where

$$f(x|M^+) = \frac{1}{\sqrt{4.5\pi}} e^{-\frac{(x-10)^2}{4.5}} \qquad \text{and} \qquad f(x|M^-) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-6)^2}{2}}.$$

Suppose the prevalence of milleritis is 20%. Therefore we have $\mathrm{P}(M^+) = 0.2$ and $\mathrm{P}(M^-) = 0.8$. The conditional probability of milleritis $\mathrm{P}(M^+|X=x)$ is shown in Figure 7.3.

Suppose when a patient comes to a doctor, their uric acid level is measured (this costs \$5) and (depending on the result) one of the following three courses of action is taken:

- a_1: Do not treat. This incurs no additional cost for a M^- patient. But if the patient is M^+, the condition will worsen and the cost of treating an acute case is \$200.

Bayesian Decision Theory with rules, example

For a measured value of X, we can use the following version of Bayes theorem to find the **conditional probability** of a patient being milleritis positive (M^+):

$$\mathrm{P}(M^+|X=x) = \frac{f(x|M^+) \times \mathrm{P}(M^+)}{f(x|M^+) \times \mathrm{P}(M^+) + f(x|M^-) \times \mathrm{P}(M^-)}$$

where

$$f(x|M^+) = \frac{1}{\sqrt{4.5\pi}} e^{-\frac{(x-10)^2}{4.5}} \qquad \text{and} \qquad f(x|M^-) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-6)^2}{2}}.$$

Suppose the prevalence of milleritis is 20%. Therefore we have $\mathrm{P}(M^+) = 0.2$ and $\mathrm{P}(M^-) = 0.8$. The conditional probability of milleritis $\mathrm{P}(M^+|X=x)$ is shown in Figure 7.3.

Suppose when a patient comes to a doctor, their uric acid level is measured (this costs \$5) and (depending on the result) one of the following three courses of action is taken:

- a_1: Do not treat. This incurs no additional cost for a M^- patient. But if the patient is M^+, the condition will worsen and the cost of treating an acute case is \$200.

Bayesian Decision Theory with rules, example

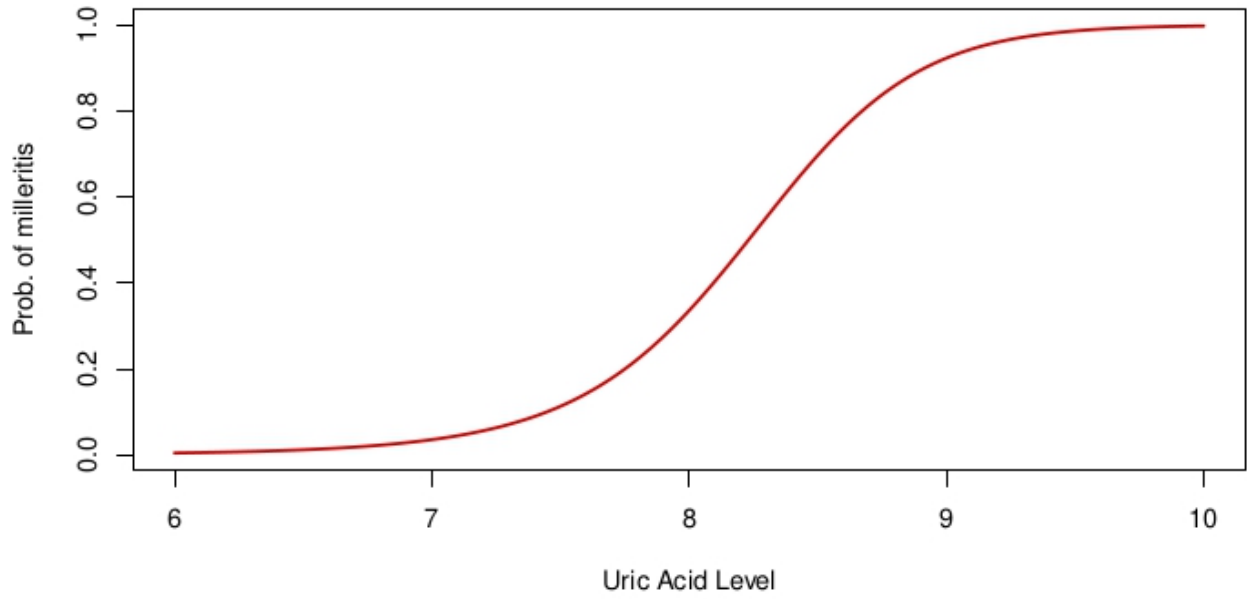


Figure 7.3: Conditional probability of milleritis $\mathrm{P}(M^+|X = x)$.

a_2: Treat for milleritis. The cost of treatment is \$55.

a_3: Do a second test which costs an additional \$25 but gives a definitive result. Only treat if the patient is M^+.

Bayesian Decision Theory

If we define the **loss** as the cost of treatment, then we have the following table of losses:

	M^-	M^+
a_1	\$5	\$205
a_2	\$60	\$60
a_3	\$30	\$85

To get the **expected loss** (as a function of x) for each action, we simply apply $\mathrm{P}(M^+|X = x)$ to these losses. As an example for a_1:

$$\mathrm{E}(\text{loss}|X = x) = 5 \times \mathrm{P}(M^-|X = x) + 205 \times \mathrm{P}(M^+|X = x),$$

where $\mathrm{P}(M^-|X = x) = 1 - \mathrm{P}(M^+|X = x)$.

Bayesian Decision Theory

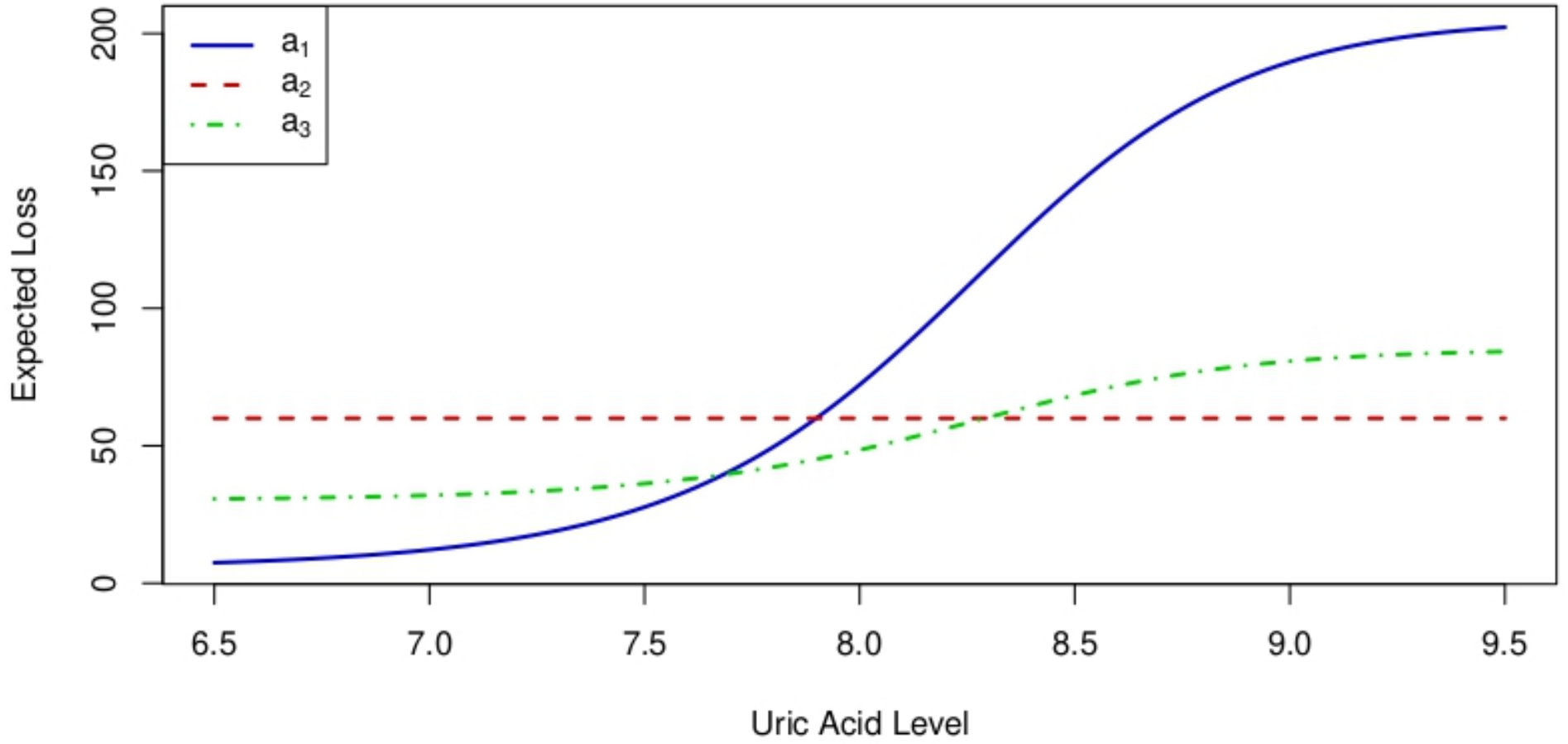


Figure 7.4: Expected losses for all three actions

Bayesian Decision Theory

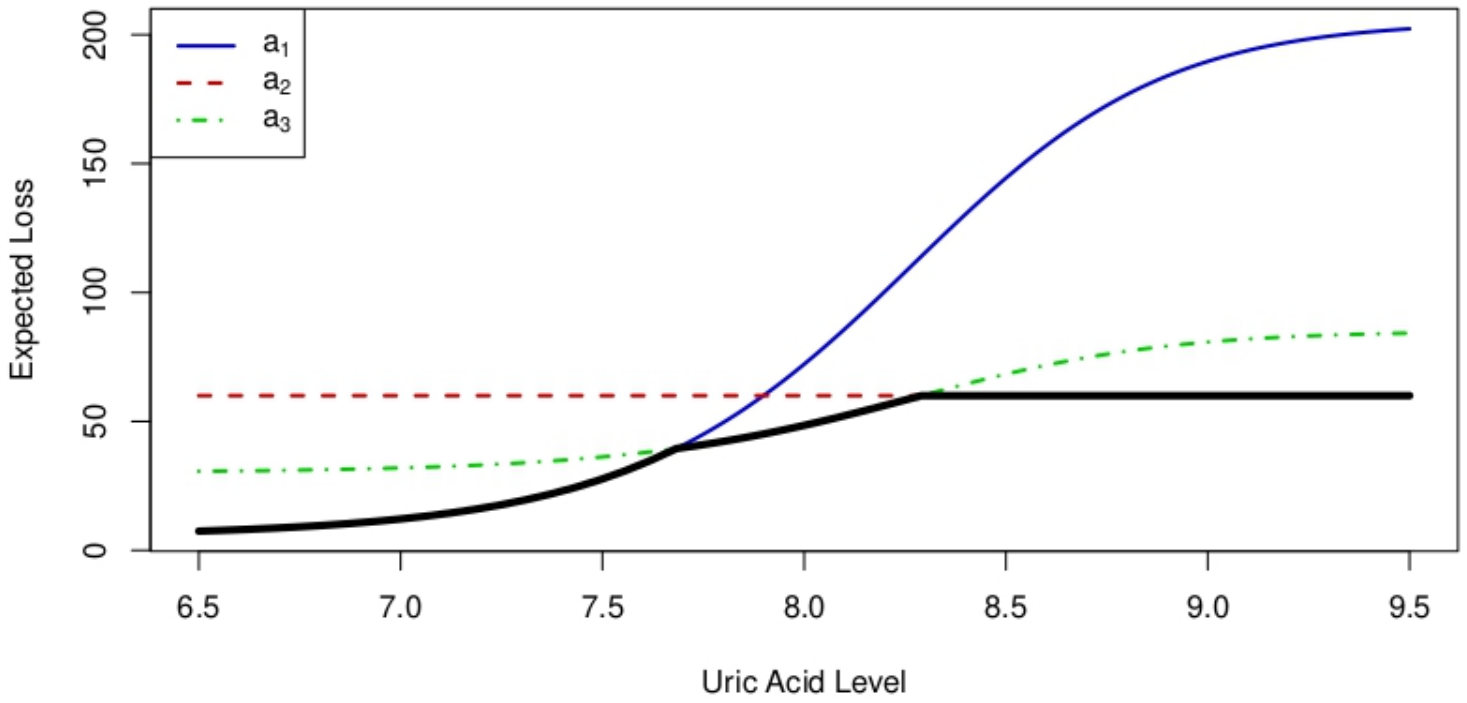


Figure 7.5: Expected losses for all three actions

Bayes rule is given by

$$a_1 \text{ for } x < 7.68, \qquad a_3 \text{ for } 7.68 < x < 8.29, \qquad \text{and } a_2 \text{ for } x > 8.29. \qquad \Diamond$$

Bayesian Decision Theory with rules, exercise

- Create a shinyApp for gender prediction example with adding plot of expected loss 1/0

Bayesian Decision Theory with rules, homework

- Create a shinyApp for this example with the slider allowing to change a_1, a_2 and a_3