CHAPTER 5

DISCRETE RANDOM VARIABLES

In the previous chapter, we looked at random experiments in terms of events. We also introduced probability defined on events as a tool for understanding random experiments. We showed how conditional probability is the logical way to change our belief about an unobserved event given that we observed another related event. In this chapter we introduce discrete random variables and probability distributions.

A random variable describes the outcome of the experiment in terms of a number. If the only possible outcomes of the experiment are distinct numbers separated from each other (e.g., counts), then we say that the random variable is discrete. There are good reasons why we introduce random variables and their notation:

- It is quicker to describe an outcome as a random variable having a particular value than to describe that outcome in words. Any event can be formed from outcomes described by the random variable using union, intersection, and complements.
- The probability distribution of the discrete random variable is a numerical function. It is easier to deal with a numerical function than with

Introduction to Bayesian Statistics, 3^{rd} *ed.*
By Bolstad, W. M. and Curran, J. M. Copyright © 2016 John Wiley & Sons, Inc.

probabilities being a function defined on sets (events). The probability of any possible event can be found from the probability distribution of the random variable using the rules of probability. So instead of having to know the probability of every possible event, we only have to know the probability distribution of the random variable.

- It becomes much easier to deal with compound events made up from repetitions of the experiment.

5.1 Discrete Random Variables

A number that is determined by the outcome of a random experiment is called a random variable. Random variables are denoted by uppercase letters, e.g., Y. The value the random variable takes is denoted by lowercase letters, e.g., y. A discrete random variable, Y, can only take on the distinct values y_k. There can be a finite possible number of values; for example, the random variable defined as "number of heads in n tosses of a coin" has possible values $0, 1, \ldots, n$. Or there can be a countably infinite number of possible values; for example, the random variable defined as "number of tosses until the first head" has possible values $1, 2, \ldots, \infty$. The key thing for discrete random variables is that the possible values are separated by gaps.

Thought Experiment 1: ***Roll of a die***
Suppose we have a fair six sided die. Our random experiment is to roll it, and we let the random variable Y be the number on the top face. There are six possible values $1, 2, \ldots, 6$. *Since the die is fair, those six values are equally likely. Now, suppose we take independent repetitions of the random variable and record each occurrence of Y. Table 5.1 shows the proportion of times each face has occurred in a typical sequence of rolls of the die, after 10, 100, 1,000, and 10,000 rolls. The last column shows the true probabilities for a fair die.*

Table 5.1 Typical results of rolling a fair die

	Proportion After					
Value	10 Rolls	100 Rolls	1,000 Rolls	10,000 Rolls	...	Probability
1	0.1	0.17	0.182	0.1668	...	0.1666
2	0.2	0.13	0.182	0.1739	...	0.1666
3	0.3	0.20	0.176	0.1716	...	0.1666
4	0.1	0.21	0.159	0.1685	...	0.1666
5	0.1	0.09	0.150	0.1592	...	0.1666
6	0.2	0.20	0.151	0.1600	...	0.1666

We note that the proportions taking any value are getting closer and closer to the true probability of that value as n increases to ∞. We could draw graphs of the proportions having each value. These are shown in Figure 5.1. The graphs are at zero for any other y value, and they have a spike at each possible value where the spike height equals the proportion of times that value occurred. The sum of spike heights equals one.

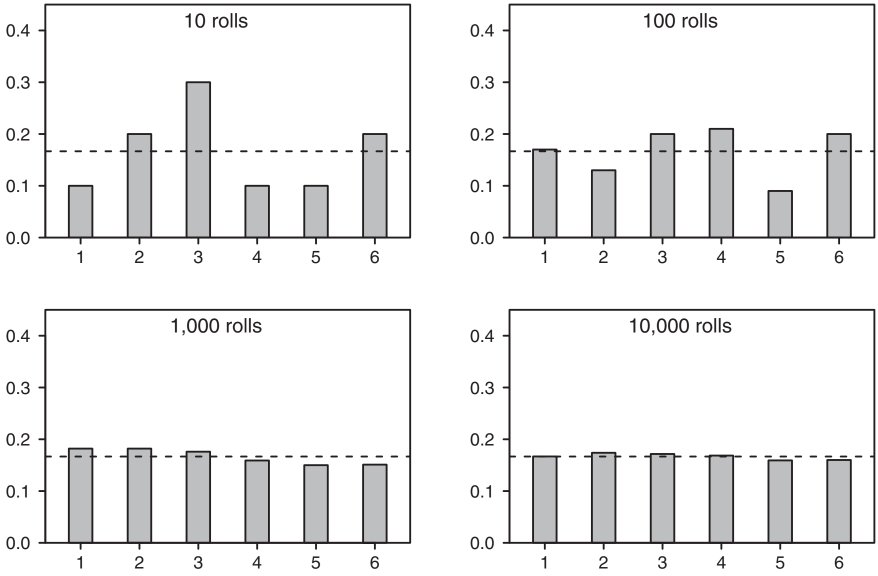

Figure 5.1 Proportions resulting from 10, 100, 1,000, and 10,000 rolls of a fair die.

Thought Experiment 2: *Random sampling from a finite population* *Suppose we have a finite population of size N. There can be at most a finite number of possible values, and they must be discrete, since there must be a gap between every pair of two real numbers. Some members of the population have the same value, so there are only K possible values $y_1, \ldots, y_K$. The probability of observing the value y_k is the proportion of population having that value.*

We start by randomly drawing from the population with replacement. Each draw is done under identical conditions. If we continue doing the sampling, then eventually we will have seen all possible values. After each draw we update the proportions in the accumulated sample that have each value. We sketch a graph with a spike at each value in the sample equal to the proportion in the sample having that value. The updating of the graph at step n is made by scaling all the existing spikes down by the ratio $\frac{n-1}{n}$ and adding $\frac{1}{n}$ to the spike at the value observed. The scaling changes the proportions after the first $n-1$ observations to the proportions after the first n observations. As the sample size increases, the sample proportions get less variable. In the limit as *the sample size n approaches infinity, the spike at each value approaches its probability.*

Thought Experiment 3: ***Number of tails before first head from independent coin tosses***
Each toss of a coin results in either a head or a tail. The probability of getting a head remains the same on each toss. The outcomes of each toss are independent of each other. This is an example of what we call Bernoulli trials. The outcome of a trial is either a success (head) or failure (tail), the probability of success remains constant over all trials, and we are taking independent trials. We are counting the number of failures before the first success. Every nonnegative integer is a possible value, and there are an infinite number of them. They must be discrete, since there is a gap between every pair of nonnegative integers.

We start by tossing the coin and counting the number of tails until the first head occurs. Then we repeat the whole process. Eventually we reach a state where most of the time we get a value we have gotten before. After each sequence of trials until the first head, we update the proportions that have each value. We sketch a graph with a spike at each value equal to the proportion having that value. As in the previous example, the updating of the graph at step n is made by scaling all the existing spikes down by the ratio $(n-1)/n$ and adding $1/n$ to the spike at the value observed. The sample proportions get less variable as the sample size increases, and in the limit as n approaches infinity, the spike at each value approaches its probability.

5.2 Probability Distribution of a Discrete Random Variable

The proportion functions that we have seen in the three thought experiments are spike functions. They have a spike at each possible value, zero at all other values, and the sum of the spike heights equals one. In the limit as the sample size approaches infinity, the proportion of times a value occurs approaches the probability of that value, and the proportion graphs approach the probability function

$$f(y_k) = P(Y = y_k)$$

for all possible values $y_1, \ldots, y_k$ of the discrete random variable. For any other value y, it equals zero.

Expected Value of a Discrete Random Variable

The expected value of a discrete random variable Y is defined to be the sum over all possible values of each possible value times its probability:

$$\mathrm{E}[Y] = \sum_{k=1} y_k \times f(y_k)\,. \tag{5.1}$$

The expected value of a random variable is often called the mean of the random variable and is denoted μ. It is like the sample mean of an infinite sample of independent repetitions of the random variable. The sample mean of a random sample of size n repetitions of the random variable is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \,.$$

Here y_i is the value that occurs on the i^{th} repetition. We are summing over all repetitions. Grouping together all repetitions that have the same possible value, we get

$$\bar{y} = \sum_{k} \frac{n_k}{n} \times y_k \,,$$

where n_k is the number of observations that have value y_k, and we are now summing over all possible values. Note that each of the y_i (observed values) equals one of the y_k (possible values). But in the limit as n approaches ∞, the relative frequency $\frac{n_k}{n}$ approaches the probability $f(y_k)$, so the sample mean, $\bar{y}$, approaches the expected value, $\mathrm{E}[Y]$. This shows that the expected value of a random variable is like the sample mean of an infinite size random sample of that variable.

The Variance of a Discrete Random Variable

The variance of a random variable is the expected value of square of the variable minus its mean.

$$\begin{aligned}\mathrm{Var}[Y] &= E[Y - \mathrm{E}[Y]]^2 \\ &= \sum_{k} (y_k - \mu)^2 \times f(y_k)\,. \end{aligned} \tag{5.2}$$

This is like the sample variance of an infinite size random sample of that variable. We note that if we square the term in brackets, break the sum into three sums, and factor the constant terms out of each sum, we get

$$\begin{aligned}\mathrm{Var}[Y] &= \sum_{k} y_k^2 \times f(y_k) - 2\mu \times \sum_{k} y_k f(y_k) + \mu^2 \times \sum_{k} f(y_k) \\ &= \mathrm{E}[Y^2] - \mu^2\,.\end{aligned}$$

Since $\mu = \mathrm{E}[Y]$, this gives another useful formula for computing the variance.

$$\mathrm{Var}[Y] = \mathrm{E}[Y^2] - [\mathrm{E}[Y]]^2\,. \tag{5.3}$$

EXAMPLE 5.1

Let Y be a discrete random variable with probability function given in the following table.

y_i	$f(y_i)$
0	.20
1	.15
2	.25
3	.35
4	.05

To find E[Y] we use Equation 5.1, which gives

$$\begin{aligned} \mathrm{E}[Y] &= 0 \times .20 + 1 \times .15 + 2 \times .25 + 3 \times .35 + 4 \times .05 \\ &= 1.90\,. \end{aligned}$$

Note that the expected value does not have to be a possible value of the random variable Y. It represents an average. We will find Var[Y] in two ways and see that they give equivalent results. First, we use the definition of variance given in Equation 5.2.

$$\begin{aligned} \mathrm{Var}[Y] &= (0-1.90)^2 \times .20 + (1-1.90)^2 \times .15 + (2-1.90)^2 \times .25 \\ &\quad + (3-1.90)^2 \times .35 + (4-1.90)^2 \times .05 \\ &= 1.49\,. \end{aligned}$$

Second, we will use Equation 5.3. We calculate

$$\begin{aligned} \mathrm{E}[Y^2] &= 0^2 \times .20 + 1^2 \times .15 + 2^2 \times .25 + 3^2 \times .35 + 4^2 \times .05 \\ &= 5.10\,. \end{aligned}$$

Putting that result in Equation 5.3, we get

$$\begin{aligned} \mathrm{Var}[Y] &= 5.10 - 1.90^2 \\ &= 1.49\,. \end{aligned}$$

■

The Mean and Variance of a Linear Function of a Random Variable

Suppose $W = a \times Y + b$, where Y is a discrete random variable. Clearly, W is another number that is the outcome of the same random experiment that

Y came from. Thus W, a linear function of a random variable Y, is another random variable. We wish to find its mean.

$$\begin{aligned}
\mathrm{E}[aY + b] &= \sum_k (ay_k + b) \times f(y_k) \\
&= \sum_k ay_k \times f(y_k) + \sum b \times f(y_k) \\
&= a \sum y_k f(y_k) + b \sum f(y_k)\,.
\end{aligned}$$

Since $\sum y_k f(y_k) = \mu$ and $\sum f(y_k) = 1$, the mean of the linear function is the linear function of the mean:

$$\mathrm{E}[aY + b] = a\,\mathrm{E}[Y] + b\,. \tag{5.4}$$

Similarly, we may wish to know its variance.

$$\begin{aligned}
\mathrm{Var}[aY + b] &= \sum_k (ay_k + b - \mathrm{E}[aY + b])^2 f(y_k) \\
&= \sum_k [a(y_k - \mathrm{E}[Y]) + b - b)]^2 f(y_k) \\
&= a^2 \sum_k (y_k - \mathrm{E}[Y])^2 f(y_k)\,.
\end{aligned}$$

Thus the variance of a linear function is the square of the multiplicative constant a times the variance :

$$\mathrm{Var}[aY + b] = a^2\,\mathrm{Var}[Y]\,. \tag{5.5}$$

The additive constant b does not enter into it.

EXAMPLE 5.1 (continued)

Suppose $W = -2Y + 3$. Then from Equation 5.4 we have

$$\begin{aligned}
\mathrm{E}[W] &= -2\,\mathrm{E}[Y] + 3 \\
&= -2 \times 1.90 + 3 \\
&= -.80
\end{aligned}$$

and from Equation 5.5 we have

$$\begin{aligned}
\mathrm{Var}[W] &= (-2)^2 \times \mathrm{Var}[Y] \\
&= 4 \times 1.49 \\
&= 5.96\,.
\end{aligned}$$

■

5.3 Binomial Distribution

Let us look at three situations and see what characteristics they have in common.

Coin tossing. Suppose we toss the same coin n times, and count the number of heads that occur. We consider that any one toss is not influenced by the outcomes of previous tosses; in other words, the outcome of one toss is independent of the outcomes of previous tosses. Since we are always tossing the same coin, the probability of getting a head on any particular toss remains constant for all tosses. The possible values of the total number of heads observed in the n tosses are $0, \ldots, n$.

Drawing from an urn with replacement. An urn contains balls of two colors, red and green. The proportion of red balls is π. We draw a ball at random from the urn, record its color, then return it to the urn, and remix the balls before the next random draw. We make a total of n draws and count the number of times we drew a red ball. Since we replace and remix the balls between draws, each draw takes place under identical conditions. The outcome of any particular draw is not influenced by the previous draw outcomes. The probability of getting a red ball on any particular draw remains equal to π, the proportion of red balls in the urn. The possible values of the total number of red balls drawn are $0, \ldots, n$.

Random sampling from a very large population. Suppose we draw a random sample of size n from a very large population. The proportion of items in the population having some attribute is π. We count the number of items in the sample that have the attribute. Since the population is very large compared to the sample size, removing a few items from the population does not perceptibly change the proportion of remaining items having the attribute. For all intents and purposes it remains π. The random draws are taken under almost identical conditions. The outcome of any draw is not influenced by the previous outcomes. The possible values of the number of items drawn that have the attribute is $0, \ldots, n$.

Characteristics of the Binomial Distribution

These three cases all have the following things in common.

- There are n *independent* trials. Each trial can result either in a "success" or a "failure."
- The probability of "success" is constant over all the trials. Let π be the probability of "success."
- Y is the number of "successes" that occurred in the n trials. Y can take on integer values $0, 1, \ldots, n$.

These are the characteristics of the *binomial*(n, π) distribution. The binomial probability function can be found from these characteristics using the laws of probability. Any sequence having exactly y successes out of the n independent trials has probability equal to $\pi^y(1-\pi)^{n-y}$, no matter in which order they occur. The event $\{Y = y\}$ is the union of all sequences such sequences. The sequences are disjoint, so the probability function of the binomial random variable Y given the parameter value π is written as

$$f(y|\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \tag{5.6}$$

for $y = 0, 1, \ldots, n$ where the binomial coefficient

$$\binom{n}{y} = \frac{n!}{y! \times (n-y)!}$$

represents the number of sequences having exactly y successes out of n trials and $\pi^y(1-\pi)^{n-y}$ is the probability of any particular sequence having exactly y successes out of n trials.

Mean of binomial. The mean of the *binomial*(n, π) distribution is the sample size times the probability of success since

$$\begin{aligned} \mathrm{E}[Y|\pi] &= \sum_{y=0}^{n} y \times f(y|\pi) \\ &= \sum_{y=0}^{n} y \times \binom{n}{y} \pi^y (1-\pi)^{n-y} \,. \end{aligned}$$

We write this as a conditional mean because it is the mean of Y given the value of the parameter π. The first term in the sum is 0, so we can start the sum at $y = 1$. We cancel y in the remaining terms, and factor out $n\pi$. This gives

$$\mathrm{E}[Y|\pi] = \sum_{y=1}^{n} n\pi \binom{n-1}{y-1} \pi^{y-1} (1-\pi)^{n-y} \,.$$

Factoring $n\pi$ out of the sum and substituting $n' = n - 1$ and $y' = y - 1$, we get

$$\mathrm{E}[Y|\pi] = n\pi \sum_{y'=0}^{n'} \binom{n'}{y'} \pi^{y'} (1-\pi)^{n'-y'} \,.$$

We see the sum is a binomial probability function summed over all possible values. Hence it equals one, and the mean of the binomial is

$$\mathrm{E}[Y|\pi] = n\pi \,. \tag{5.7}$$

Variance of binomial. The variance is the sample size times the probability of success times the probability of failure. We write this as a conditional variance since it is the variance of Y given the value of the parameter π. Note that

$$\begin{aligned} \mathrm{E}[Y(Y-1)|\pi] &= \sum_{y=0}^{n} y(y-1) \times f(y|\pi) \\ &= \sum_{y=0}^{n} y(y-1) \times \binom{n}{y} \pi^{y}(1-\pi)^{n-y} . \end{aligned}$$

The first two terms in the sum equal 0, so we can start summing at $y = 2$. We cancel $y(y-1)$ out of the remaining terms and factor out $n(n-1)\pi^2$ to get

$$\mathrm{E}[Y(Y-1)|\pi] = \sum_{y=2}^{n} n(n-1)\pi^2 \binom{n-2}{y-2} \pi^{y-2}(1-\pi)^{n-y} .$$

Substituting $y' = y - 2$ and $n' = n - 2$, we get

$$\begin{aligned} \mathrm{E}[Y(Y-1)|\pi] &= n(n-1)\pi^2 \sum_{y'=0}^{n-2} \binom{n'}{y'} \pi^{y'}(1-\pi)^{n'} \\ &= n(n-1)\pi^2 \end{aligned}$$

since we are summing a binomial distribution over all possible values. The variance can be found by

$$\begin{aligned} \mathrm{Var}[Y|\pi] &= \mathrm{E}[Y^2|\pi] - [\mathrm{E}[Y|\pi]]^2 \\ &= \mathrm{E}[Y(Y-1)|\pi] + \mathrm{E}[Y|\pi] - [\mathrm{E}[Y|\pi]]^2 \\ &= n(n-1)\pi^2 + n\pi - [n\pi]^2 . \end{aligned}$$

Hence the variance of the binomial is the sample size times the probability of success times the probability of failure.

$$\mathrm{Var}[Y|\pi] = n\pi(1-\pi) . \tag{5.8}$$

5.4 Hypergeometric Distribution

The hypergeometric distribution models sampling from an urn without replacement. There is an urn containing N balls, R of which are red. A sequence of n balls is drawn randomly from the urn *without replacement*. Drawing a red ball is called a "success." The probability of success π does not stay constant over all the draws. At each draw the probability of "success" is the proportion of red balls remaining in the urn, which does depend on the outcomes of previous draws. Y is the number of "successes" in the n trials. Y can take on integer values $0, 1, \ldots, n$.

Probability Function of Hypergeometric

The probability function of the hypergeometric random variable Y given the parameters N, n, R is written as

$$f(y|N, R, n) = \frac{\binom{R}{y} \times \binom{N-R}{n-y}}{\binom{N}{n}}$$

for possible values $y = 0, 1, \ldots, n$.

Mean and variance of hypergeometric. The conditional mean of the hypergeometric distribution is given by

$$\text{E}[Y|N, R, n] = n \times \frac{R}{N}\,.$$

The conditional variance of the hypergeometric distribution is given by

$$\text{Var}[Y|N, R, n] = n \times \frac{R}{N} \times \left(1 - \frac{R}{N}\right) \times \left(\frac{N-n}{N-1}\right)$$

We note that $\frac{R}{N}$ is the proportion of red balls in the urn. The mean and variance of the hypergeometric are similar to that of the binomial, except that the variance is smaller due to the finite population correction factor $\frac{N-n}{N-1}$.

5.5 Poisson Distribution

The Poisson distribution is another distribution for counts.[1] Specifically, the Poisson is a distribution which counts the number of occurrences of rare events over a period of time or space. Unlike the binomial which counts the number of events (successes) in a known number of independent trials, the number of trials in the Poisson is so large that it is not known. Nevertheless, looking at the binomial gives us way to start our investigation of the Poisson. Let Y be a binomial random variable where n is very large, and π is very small. The binomial probability function is

$$\begin{aligned} P(Y = y|\pi) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ &= \frac{n!}{(n-y)!y!} \pi^y (1-\pi)^{n-y} \end{aligned}$$

[1] First studied by Simeon Poisson (1781–1840).

for $y = 0, \ldots, n$. Since π is small, the only terms that have appreciable probability are those where y is much smaller than n. We will look at the probabilities for those small values of y. Let $\mu = n\pi$. The probability function is

$$P(Y = y|\mu) = \frac{n!}{(n-y)!y!}\left(\frac{\mu}{n}\right)^y\left(1-\frac{\mu}{n}\right)^{n-y}.$$

Rearranging the terms, we get

$$P(Y = y|\mu) = \frac{n}{n} \times \frac{n-1}{n} \times \ldots \times \frac{n-y+1}{n} \times \frac{\mu^y}{y!}\left(1-\frac{\mu}{n}\right)^n\left(1-\frac{\mu}{n}\right)^{-y}.$$

But all the values $\frac{n}{n}, \frac{n-1}{n}, \ldots, \frac{n-y+1}{n}$ are approximately equal to 1 since y is much smaller than n. We let n approach infinity, and π approach 0 in such a way that $\mu = n\pi$ is constant. We know that

$$\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^n = e^{-\mu} \quad \text{and} \quad \lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^{-y} = 1,$$

so the Poisson probability function is given by

$$f(y|\mu) = \frac{\mu^y e^{-\mu}}{y!} \tag{5.9}$$

for $y = 0, 1, \ldots$. Thus the *Poisson*(μ) distribution can be used to approximate a *binomial*(n, π) when n is large, π is very small, and $\mu = n\pi$.

Characteristics of the Poisson Distribution

Think of the period of time (or space) divided into n equal parts. The total number of occurrences is the sum of the number of occurrences in all n parts. We see from the Poisson approximation to the binomial that the Poisson distribution is a limiting case of the binomial distribution as $n \to \infty$ and $\pi \to 0$ at such a rate that $n\pi = \mu$ is constant.

- In the binomial, the probability of success remains constant over all the trials. It follows that the instantaneous rate of occurrences per unit time (or space) for the Poisson is constant.

- In the binomial, the trials are independent. Thus the Poisson occurrences in any two non-overlapping intervals will be independent of each other. It follows that the Poisson occurrences are randomly occurring through time at the constant instantaneous rate.

- In the binomial each trial contributes either one success or one failure. It follows that Poisson counts occur one at a time.

- The possible values are $y = 0, 1, \ldots$.

Mean and variance of Poisson. The mean of the $Poisson(\mu)$ can be found by

$$\begin{aligned}\mathrm{E}[y|\mu] &= \sum_{y=0}^{\infty} y\frac{\mu^y e^{-\mu}}{y!}\\ &= \sum_{y=1}^{\infty} \frac{\mu^y e^{-\mu}}{(y-1)!}\end{aligned}$$

We let $y' = y - 1$ and factor out μ:

$$\mathrm{E}[y|\mu] = \mu \sum_{y'=0}^{\infty} \frac{\mu^{y'} e^{-\mu}}{y'!}\,.$$

The sum equals one since it is the the sum is over all possible values of a Poisson distribution, so the mean of the $Poisson(\mu)$ is

$$\mathrm{E}[y|\mu] = \mu\,.$$

Similarly, we can evaluate

$$\begin{aligned}\mathrm{E}[y \times (y-1)|\mu] &= \sum_{y=0}^{\infty} y \times (y-1] \times \frac{\mu^y e^{-\mu}}{y!}\\ &= \sum_{y=2}^{\infty} \frac{\mu^y e^{-\mu}}{(y-2)!}\end{aligned}$$

We let $y' = y - 2$, and factor out μ^2

$$\mathrm{E}[y \times (y-1)|\mu] = \mu^2 \sum_{y'=0}^{\infty} \frac{\mu^{y'} e^{-\mu}}{y'!}\,.$$

The sum equals one since it is the the sum is over all possible values of a Poisson distribution, so $\mathrm{E}[y \times (y-1)|\mu]$ for a $Poisson(\mu)$ is given by

$$\mathrm{E}[y \times (y-1)|\mu] = \mu^2\,.$$

The Poisson variance is given by

$$\begin{aligned}\mathrm{Var}[y|\mu] &= \mathrm{E}[y^2|\mu) - [\mathrm{E}[y|\mu]]^2\\ &= E[y \times (y-1)|\mu] + \mathrm{E}[y|\mu] - [E(y|\mu]]^2\\ &= \mu^2 + \mu - \mu^2\\ &= \mu\,.\end{aligned}$$

Thus we see the mean and variance of a $Poisson(\mu)$ are both equal to μ.

Table 5.2 Universe of joint experiment

(x_1, y_1)	.	.	.	(x_1, y_j)	.	.	.	(x_1, y_J)
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
(x_i, y_1)	.	.	.	(x_i, y_j)	.	.	.	(x_i, y_J)
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
(x_I, y_1)	.	.	.	(x_I, y_j)	.	.	.	(x_I, y_J)

5.6 Joint Random Variables

When two (or more) numbers are determined from the outcome of a random experiment, we call it a joint experiment. The two numbers are called joint random variables and denoted X, Y. If both the random variables are discrete, they each have separated possible values x_i for $i = 1, \ldots, I$ and y_j for $j = 1, \ldots, J$. The *universe* for the experiment is the set of all possible outcomes of the experiment which are all possible ordered pairs of possible values. The *universe* of the joint experiment is shown in Table 5.2.

The joint probability function of two discrete joint random variables is defined at each point in the universe:

$$f(x_i, y_j) = P(X = x_i, Y = y_j)$$

for $i = 1, \ldots, I$, and $j = 1, \ldots, J$. This is the probability that $X = x_i$ and $Y = y_j$ simultaneously, in other words, the probability of the intersection of the events $X = x_i$ and $Y = y_j$. These joint probabilities can be put in a table.

We might want to consider the probability distribution of just one of the joint random variables, for instance, Y. The event $Y = y_j$ for some fixed value y_j is the union of all events $X = x_i, Y = y_j$, where $i = 1, \ldots, I$, and they are all disjoint. Thus

$$P(Y = y_j) = P(\cup_i (X = x_i, Y = y_j)) = \sum_i P(X = x_i, Y = y_j)$$

for $j = 1, \ldots, J$, since probability is additive over a disjoint union. This probability distribution of Y by itself is called the *marginal* distribution of Y. Putting this relationship in terms of the probability function, we get

$$f(y_j) = \sum_i f(x_i, y_j) \tag{5.10}$$

Table 5.3 Joint and marginal probability distributions

	y_1	.	.	.	y_j	.	.	.	y_J	
x_1	$f(x_1, y_1)$	.	.	.	$f(x_1, y_j)$	.	.	.	$f(x_1, y_J)$	$f(x_1)$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
x_i	$f(x_i, y_1)$	.	.	.	$f(x_i, y_j)$	.	.	.	$f(x_i, y_J)$	$f(x_i)$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
x_I	$f(x_I, y_1)$	.	.	.	$f(x_I, y_j)$	.	.	.	$f(x_I, y_J)$	$f(x_I)$
	$f(y_1)$	.	.	.	$f(y_j)$	.	.	.	$f(y_J)$	

for $j = 1, \ldots J$. So we see that the individual probabilities of Y is found by summing the joint probabilities down the columns. Similarly the individual probabilities of X can be found by summing the joint probabilities across the rows. We can write them on the margins of the table, hence the names *marginal* probability distribution of Y and X respectively. The joint probability distribution and the marginal probability distributions are shown in Table 5.3. The joint probabilities are in the main body of the table, and the marginal probabilities for X and Y are in the right column and bottom row, respectively.

The expected value of a function of the joint random variables is given by

$$\mathrm{E}[h(X,Y)] = \sum_i \sum_j h(x_i, y_j) \times f(x_i, y_j)\,.$$

Often we wish to find the expected value of a sum of random variables. In that case

$$\begin{aligned}
\mathrm{E}[X+Y] &= \sum_i \sum_j (x_i + y_j) \times f(x_i, y_j] \\
&= \sum_i \sum_j x_i \times f(x_i, y_j) + \sum_i \sum_j y_j \times f(x_i, y_j) \\
&= \sum_i x_i \sum_j f(x_i, y_j) + \sum_j y_j \sum_i f(x_i, y_j) \\
&= \sum_i x_i \times f(x_i) + \sum_j y_j \times f(y_j)\,.
\end{aligned}$$

We see the mean of the sum of two random variables is the sum of the means.

$$\mathrm{E}[X+Y]=\mathrm{E}[X]+\mathrm{E}[Y]\,. \tag{5.11}$$

This equation always holds.

Independent Random Variables

Two (discrete) random variables X and Y are independent of each other if and only if every element in the joint distribution table equals the product of the corresponding marginal distributions. In other words,

$$f(x_i,y_j)=f(x_i)\times f(y_j)$$

for all possible x_i and y_j.
The variance of a sum of random variables is given by

$$\begin{aligned}\mathrm{Var}[X+Y]&=E(X+Y-\mathrm{E}[X+Y])^2\\&=\sum_i\sum_j(x_i+y_j-(\mathrm{E}[X]+\mathrm{E}[Y])^2\times f(x_i,y_j)\\&=\sum_i\sum_j[(x_i-\mathrm{E}[X])+(y_j-\mathrm{E}[Y])]^2\times f(x_i,y_j)\,.\end{aligned}$$

Multiplying this out and breaking it into three separate sums gives

$$\begin{aligned}\mathrm{Var}[X+Y]=&\sum_i\sum_j(x_i-\mathrm{E}[X])^2\times f(x_i,y_j)\\&+\sum_i\sum_j 2(x_i-\mathrm{E}[X])(y_j-\mathrm{E}[Y])f(x_i,y_j)\\&+\sum_i\sum_j(y_j-\mathrm{E}[Y])^2\times f(x_i,y_j)\,.\end{aligned}$$

The middle term is 2 × the covariance of the random variables. For independent random variables the covariance is given by

$$\begin{aligned}\mathrm{Cov}[X,Y]&=\sum_i\sum_j(x_i-\mathrm{E}[X])\times(y_j-\mathrm{E}[Y])f(x_i,y_j)\\&=\sum_i(x_i-\mathrm{E}[X])f(x_i)\times\sum_j(y_j-\mathrm{E}[Y])f(y_j)\,.\end{aligned}$$

This is clearly equal to 0. Hence for independent random variables we have

$$\mathrm{Var}[X+Y]=\sum_i(x_i-\mathrm{E}[X]))^2\times f(x_i)+\sum_j(y_j-\mathrm{E}[Y])^2\times f(y_j)\,.$$

We see the variance of the sum of two independent random variables is the sum of the variances.

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]\,. \tag{5.12}$$

This equation only holds for independent[2] random variables!

EXAMPLE 5.2

Let X and Y be jointly distributed discrete random variables. Their joint probability distribution is given in the following table:

		Y				
		1	2	3	4	$f(x)$
X	1	.02	.04	.06	.08	
	2	.03	.01	.09	.17	
	3	.05	.15	.15	.15	
	$f(y)$					

We find the marginal distributions of X and Y by summing across the rows and summing down the columns, respectively. That gives the table

		Y				
		1	2	3	4	$f(x)$
X	1	.02	.04	.06	.08	.2
	2	.03	.01	.09	.17	.3
	3	.05	.15	.15	.15	.5
	$f(y)$	.1	.2	.3	.4	

We see that the joint probability $f(x_i, y_j)$ is not always equal to the product of the marginal probabilities $f(x_i) \times f(y_j)$. Therefore the two random variables X and Y are not independent. ■

Mean and variance of a difference between two independent random variables. When we combine the results of Equations 5.10 and 5.11 with the results of Equations 5.4 and 5.5, we find the that mean of a difference between random variables is

$$\mathrm{E}[X-Y] = \mathrm{E}[X] - \mathrm{E}[Y]\,. \tag{5.13}$$

If the two random variables are independent, we find that the variance of their difference is

$$\mathrm{Var}[X-Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]\,. \tag{5.14}$$

[2] In general, the variance of a sum of two random variables is given by $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + 2 \times \mathrm{Cov}[X,Y] + \mathrm{Var}[Y]$.

Variability always adds for independent random variables, regardless of whether we are taking the sum or taking the difference.

5.7 Conditional Probability for Joint Random Variables

If we are given $Y = y_j$, the reduced universe is the set of ordered pairs where the second element is y_j. This is shown in Table 5.4. It is the only part of the universe that remains, given $Y = y_j$. The only part of the event $X = x_i$ that remains is the part in the reduced universe. This is the intersection of the events $X = x_i$ and $Y = y_j$. Table 5.5 shows the original joint probability function in the reduced universe, along with the marginal probability. We see that this is not a probability distribution. The sum of the probabilities in the reduced universe sums to the marginal probability, not to one!

The conditional probability that random variable $X = x_i$, given $Y = y_j$ is the probability of the intersection of the events $X = x_i$ and $Y = y_j$ divided by the probability that $Y = y_j$ from Equation 4.1. Dividing the joint probability by the marginal probability scales it up so the probability of the reduced universe equals 1. The conditional probability is given by

$$f(x_i|y_j) = P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}\,. \tag{5.15}$$

When we put this in terms of the joint and marginal probability functions, we get

$$f(x_i|y_j) = \frac{f(x_i, y_j)}{f(y_j)}\,. \tag{5.16}$$

Table 5.4 Reduced universe given $Y = y_j$

·	·	·	·	(x_1, y_j)	·	·	·	·
·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·
·	·	·	·	(x_i, y_j)	·	·	·	·
·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·
·	·	·	·	(x_I, y_j)	·	·	·	·

Table 5.5 Joint probability function values in the reduced universe $Y = y_j$. The marginal probability is found by summing down the column.

.	.	.	.	$f(x_1, y_j)$	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	$f(x_i, y_j)$	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	$f(x_I, y_j)$	.	.	.	.
.	.	.	.	$f(y_j)$	.	.	.	.

The conditional probability distribution. Letting x_i vary across all possible values of X gives us the conditional probability distribution of $X|Y = y_j$. The conditional probability distribution is defined on the reduced universe given $Y = y_j$. The conditional probability distribution is shown in Table 5.6. Each entry was found by dividing the *i,j* entry in the joint probability table by j^{th} element in the marginal probability. The marginal probability is $f(y_j) = \sum_i f(x_i, y_j)$ and is found by summing down the j^{th} column of the joint probability table. So the conditional probability of x_i given y_j is the j^{th} column in the joint probability table, divided by the sum of the joint probabilities in the j^{th} column.

EXAMPLE 5.2 (continued)

If we want to determine the conditional probability $P(X = 2|Y = 2)$, we plug in the joint and marginal probabilities into Equation 5.15. This gives

$$\begin{aligned} P(X = 2|Y = 2) &= \frac{P(X = 2, Y = 2)}{P(Y = 2)} \\ &= \frac{.01}{.2} \\ &= .05\,. \end{aligned}$$

■

Table 5.6 The conditional probability function defined on the reduced universe $Y = y_j$

.	.	.	.	$f(x_1 \mid y_j)$	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	$f(x_i \mid y_j)$	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	$f(x_I \mid y_j)$	.	.	.	.

Conditional probability as multiplication rule. Using similar arguments, we could find that the conditional probability function of Y given $X = x_i$ is given by

$$f(y_j|x_i) = \frac{f(x_i, y_j)}{f(x_i)} \,.$$

However, we will not use the relationship in this form, since we do not consider the random variables interchangeably. In Bayesian statistics, the random variable X is the unobservable parameter. The random variable Y is an observable random variable that has a probability distribution depending on the parameter. In the next chapter we will use the conditional probability relationship as the multiplication rule

$$f(x_i, y_j) = f(x_i) \times f(y_j|x_i) \tag{5.17}$$

when we develop Bayes' theorem for discrete random variables.

Main Points

- A random variable Y is a number associated with the outcome of a random experiment.
- If the only possible values of the random variable are a finite set of separated values, $y_1, \ldots, y_K$ the random variable is said to be discrete.
- The probability distribution of the discrete random variable gives the probability associated with each possible value.
- The probability of any event associated with the random experiment can be calculated from the probability function of the random variable using the laws of probability.

- The expected value of a discrete random variable is

$$\mathrm{E}[Y] \quad = \quad \sum_k y_k f(y_k)\,,$$

  where the sum is over all possible values of the random variable. It is the mean of the distribution of the random variable.

- The variance of a discrete random variable is the expected value of the squared deviation of the random variable from its mean.

$$\mathrm{Var}[Y] \;=\; E(Y - \mathrm{E}[Y])^2 \;=\; \sum_k (y_k - \mathrm{E}[Y])^2 f(y_k)\,.$$

  Another formula for the variance is

$$\mathrm{Var}[Y] \;=\; \mathrm{E}[Y^2] - [\mathrm{E}[Y]]^2\,.$$

- The mean and variance of a linear function of a random variable $aY + b$ are

$$\mathrm{E}[aY + b] \;=\; a\,\mathrm{E}[Y] + b$$

  and

$$\mathrm{Var}[aY + b] \;=\; a^2 \times \mathrm{Var}[Y]\,.$$

- The *binomial*(n, π) distribution models the number of successes in n independent trials where each trial has the same success probability, π.

- The *binomial* distribution is used for sampling from a finite population with replacement.

- The *hypergeometric* distribution is used for sampling from a finite population without replacement.

- The *Poisson*(μ) distribution counts the number of occurrences of a rare event. Occurrences are occurring randomly through time (or space) at a constant rate and occur one at a time. It is also used to approximate the *binomial*(n, π) where n is large and π is small and we let $\mu = n\pi$.

- The joint probability distribution of two discrete random variables X and Y is written as joint probability function

$$f(x_i, y_j) \quad = \quad P(X = x_i, Y = y_j)\,.$$

  Note: $(X = x_i, Y = y_j)$ is another way of writing the intersection $(X = x_i \cap Y = y_j)$. This joint probability function can be put in a table.

- The marginal probability distribution of one of the random variables can be found by summing the joint probability distribution across rows (for X) or by summing down columns (for Y).

- The mean and variance of a sum of independent random variables are

$$\mathrm{E}[X+Y] \;=\; \mathrm{E}[X]+\mathrm{E}[Y]$$

  and

$$\mathrm{Var}[X+Y] \;=\; \mathrm{Var}[X]+\mathrm{Var}[Y]\,.$$

- The mean and variance of a difference between independent random variables are

$$\mathrm{E}[X-Y] \;=\; \mathrm{E}[X]-\mathrm{E}[Y]$$

  and

$$\mathrm{Var}[X-Y] \;=\; \mathrm{Var}[X]+\mathrm{Var}[Y]\,.$$

- Conditional probability function of X given $Y=y_j$ is found by

$$f(x_i|y_j) \;=\; \frac{f(x_i,y_j)}{f(y_j)}\,.$$

  This is the joint probability divided by the marginal probability that $Y=y_j$.

- The joint probabilities on the reduced universe $Y=y_j$ are not a probability distribution. They sum to the marginal probability $f(y_j)$, not to one.

- Dividing the joint probabilities by the marginal probability scales up the probabilities, so the sum of probabilities in the reduced universe is one.

Exercises

5.1. A discrete random variable Y has discrete distribution given in the following table:

y_i	$f(y_i)$
0	.2
1	.3
2	.3
3	.1
4	.1

(a) Calculate $P(1<Y\leq 3)$.

(b) Calculate $\mathrm{E}[Y]$.

(c) Calculate Var$[Y]$.

(d) Let $W = 2Y + 3$. Calculate E$[W]$.

(e) Calculate Var$[W]$.

5.2. A discrete random variable Y has discrete distribution given in the following table:

y_i	$f(y_i)$
0	.1
1	.2
2	.3
5	.4

(a) Calculate $P(0 < Y < 2)$.

(b) Calculate E$[Y]$.

(c) Calculate Var$[Y]$.

(d) Let $W = 3Y - 1$. Calculate E$[W]$.

(e) Calculate Var$[W]$.

5.3. Let Y be $binomial(n = 5, \pi = .6)$.

(a) Calculate the mean and variance by filling in the following table:

y_i	$f(y_i)$	$y_i \times f(y_i)$	$y_i^2 \times f(y_i)$
0			
1			
2			
3			
4			
5			
Sum			

i. E$[Y] =$

ii. Var$[Y] =$

(b) Calculate the mean and variance of Y using Equations 5.7 and 5.8, respectively. Do you get the same results as in part (a)?

5.4. Let Y be $binomial(n = 4, \pi = .3)$.

(a) Calculate the mean and variance by filling in the following table:

y_i	$f(y_i)$	$y_i \times f(y_i)$	$y_i^2 \times f(y_i)$
0			
1			
2			
3			
4			
Sum			

i. $\mathrm{E}[Y] =$

ii. $\mathrm{Var}[Y] =$

(b) Calculate the mean and variance of Y using Equations 5.7 and 5.8, respectively. Do you get the same as you got in part (a)?

5.5. Suppose there is an urn containing 20 green balls and 30 red balls. A single trial consists of drawing a ball randomly from the urn, recording its color, and then putting it back in the urn. The experiment consists of 4 independent trials.

(a) List each outcome (sequence of 4 trials) in the sample space together with its probability. What do you notice about the probabilities of outcomes that have the same number of green balls?

(b) Let Y be the number of green balls drawn. List the outcomes that make up each of the following events:
$Y = 0 \quad Y = 1 \quad Y = 2 \quad Y = 3 \quad Y = 4$

(c) What can you say about $P(Y = y)$ in terms of "number of outcomes where $Y = y$, and the probability of any particular sequence of outcomes where $Y = y$.

(d) Explain how this relates to the binomial probability function.

5.6. Suppose there is an urn containing 20 green balls and 30 red balls. A single trial consists of drawing a ball randomly from the urn, recording its color. This time the ball is not returned to the urn. The experiment consists of 4 independent trials.

(a) List each outcome (sequence of 4 trials) in the sample space together with its probability. What do you notice about the probabilities of outcomes that have the same number of green balls.

(b) Let Y be the number of green balls drawn. List the outcomes that make up each of the following events:
$Y = 0 \quad Y = 1 \quad Y = 2 \quad Y = 3 \quad Y = 4$

(c) What can you say about $P(Y = y)$ in terms of "number of outcomes where $Y = y$, and the probability of any particular sequence of outcomes where $Y = y$.

(d) Explain what this means in terms of the hypergeometric distribution. Hint: write this in terms of factorials, then rearrange the terms.

5.7. Let Y have the $Poisson(\mu = 2)$ distribution.

(a) Calculate $P(Y = 2)$.

(b) Calculate $P(Y \leq 2)$.

(c) Calculate $P(1 \leq Y < 4)$.

5.8. Let Y have the $Poisson(\mu = 3)$ distribution.

(a) Calculate $P(Y = 3)$.

(b) Calculate $P(Y \leq 3)$.

(c) Calculate $P(1 \leq Y < 5)$.

5.9. Let X and Y be jointly distributed discrete random variables. Their joint probability distribution is given in the following table:

X			Y			
	1	2	3	4	5	$f(x)$
1	.02	.04	.06	.08	.05	
2	.08	.02	.10	.02	.03	
3	.05	.05	.03	.02	.10	
4	.10	.04	.05	.03	.03	
$f(y)$						

(a) Calculate the marginal probability distribution of X.

(b) Calculate the marginal probability distribution of Y.

(c) Are X and Y independent random variables? Explain why or why not.

(d) Calculate the conditional probability $P(X = 3|Y = 1)$.

5.10. Let X and Y be jointly distributed discrete random variables. Their joint probability distribution is given in the following table:

X	Y					
	1	2	3	4	5	$f(x)$
1	.015	.030	.010	.020	.025	
2	.030	.060	.020	.040	.050	
3	.045	.090	.030	.060	.075	
4	.060	.120	.040	.080	.100	
$f(y)$						

(a) Calculate the marginal probability distribution of X.

(b) Calculate the marginal probability distribution of Y.

(c) Are X and Y independent random variables? Explain why or why not.

(d) Calculate the conditional probability $P(X = 2|Y = 3)$.