4

SUMMARIZING DISCRETE RANDOM VARIABLES

When we first looked at Bernoulli trials in Example 2.1.2 we asked the question "On average how many successes will there be after n trials?" In order to answer this question, a specific definition of "average" must be developed.

To begin, consider how to extend the basic notion of the average of a list of numbers to the situation of equally likely outcomes. For instance, if we want to know what the average roll of a die will be, it makes sense to declare it to be 3.5, the average value of $1, 2, 3, 4, 5$, and 6. A motivation for a more general definition of average comes from a rewriting of this calculation.

$$\frac{1+2+3+4+5+6}{6} = 1(\frac{1}{6}) + 2(\frac{1}{6}) + 3(\frac{1}{6}) + 4(\frac{1}{6}) + 5(\frac{1}{6}) + 6(\frac{1}{6}).$$

From the perspective of the right hand side of the equation, the results of all outcomes are added together after being weighted, each according to its probability. In the case of a die, all six outcomes have probability $\frac{1}{6}$.

4.1 EXPECTED VALUE

DEFINITION 4.1.1. *Let $X : S \to T$ be a discrete random variable (so T is countable). Then the expected value (or average) of X is written as $E[X]$ and is given by*

$$E[X] = \sum_{t \in T} t \cdot P(X = t)$$

provided that the sum converges absolutely. In this case we say that X has "finite expectation". If the sum diverges to $\pm\infty$ we say the random variable has infinite expectation. If the sum diverges, but not to infinity, we say the expected value is undefined.

EXAMPLE 4.1.2. In the previous chapter, Example 3.1.4 described a lottery for which a ticket could be worth nothing, or it could be worth either \$20 or \$200. What is the average value of such a ticket?

We calculated the distribution of ticket values as $P(X = 200) = \frac{1}{1000}$, $P(X = 20) = \frac{27}{1000}$, and $P(X = 0) = \frac{972}{1000}$. Applying the definition of expected value results in

$$E[X] = 200(\frac{1}{1000}) + 20(\frac{27}{1000}) + 0(\frac{972}{1000}) = 0.74,$$

so a ticket has an expected value of 56 cents. ■

It is possible to think of a constant as a random variable. If $c \in \mathbb{R}$ then we could define a random variable X with a distribution such that $P(X = c) = 1$. It is a slight abuse of notation, but in this case we will simply write c for both the real number as well as the constant random variable. Such random variables have the obvious expected value.

THEOREM 4.1.3. *Let c be a real number. Then $E[c] = c$.*

Proof - By definition $E[c]$ is a sum over all possible values of c, but in this case that is just a single value, so $E[c] = c \cdot P(c = c) = c \cdot 1 = c$. ■

When the range of X is finite, $E[X]$ always exists since it is a finite sum. When the range of X is infinite there is a possibility that the infinite series will not be absolutely convergent and therefore that $E[X]$ will be infinite or undefined. In fact, when proving theorems about how expected values behave, most of the complications arise from the fact that one must know that an infinite sum converges absolutely in order to rearrange terms within that sum with equality. The next examples explore ways in which expected values may misbehave.

EXAMPLE 4.1.4. Suppose X is a random variable taking values in the range $T = \{2, 4, 8, 16, \dots\}$ such that $P(X = 2^n) = \frac{1}{2^n}$ for all integers $n \geq 1$.

This is the distribution of a random variable since

$$\sum_{n=1}^{\infty} P(X = 2^n) = \sum_{n=1}^{\infty} \frac{1}{2^n} = 1.$$

But note that

$$\sum_{n=1}^{\infty} 2^n \cdot P(X = 2^n) = \sum_{n=1}^{\infty} 2^n \frac{1}{2^n} = \sum_{n=1}^{\infty} 1$$

which diverges to infinity, so this random variable has an infinite expected value. ■

EXAMPLE 4.1.5. Suppose X is a random variable taking values in the range $T = \{-2, 4, -8, 16, \dots\}$ such that $P(X = (-2)^n) = \frac{1}{2^n}$ for all integers $n \geq 1$.

$$\sum_{n=1}^{\infty} (-2)^n \cdot P(X = 2^n) = \sum_{n=1}^{\infty} (-2)^n \frac{1}{2^n} = \sum_{n=1}^{\infty} (-1)^n.$$

This infinite sum diverges (not to $\pm\infty$), so the expected value of this random variable is undefined. ■

The examples above were specifically constructed to produce series which clearly diverged, but in general it can be complicated to check whether an infinite sum is absolutely convergent or not. The next technical lemma provides a condition that is often simpler to check. The convenience of this lemma is that, since $|X|$ is always positive, the terms of the series for $E[|X|]$ may be freely rearranged without changing the value of (or the convergence of) the sum.

LEMMA 4.1.6. *$E[X]$ is a real number if and only if $E[|X|] < \infty$.*

Proof - Let T be the range of X. So $U = \{|t| : t \in T\}$ is the range of $|X|$. By definition

$$E[|X|] = \sum_{u \in U} u \cdot P(|X| = u), \quad \text{while}$$

$$E[X] = \sum_{t \in T} t \cdot P(X = t).$$

To more easilly relate these two sums, define $\hat{T} = \{t : |t| \in U\}$. Since every $u \in U$ came from some $t \in T$ the new set $\hat{T}$ contains every element of T. For every $t \in \hat{T}$ for which $t \notin T$, the element is outside of the range of X and so $P(X = t) = 0$ for such elements. Because of this $E[X]$ may be written as

$$E[X] = \sum_{t \in \hat{T}} t \cdot P(X = t)$$

since any additional terms in the series are zero.
Note that for each $u \in U$, the event $(|X| = u)$ is equal to $(X = u) \cup (X = -u)$ where each of u and $-u$ is an element of $\hat{T}$. Therefore,

$$\begin{aligned} u \cdot P(U = u) &= u \cdot (P(X = u) + P(X = -u)) \\ &= u \cdot P(X = u) + u \cdot P(X = -u) \\ &= |u| \cdot P(X = u) + |-u| \cdot P(X = -u) \end{aligned}$$

(When $u = 0$ the quantities $P(|X| = 0)$ and $P(X = 0) + P(X = -0)$ are typically not equal, but the equation is still true since both sides of the equation are zero). Summing over all $u \in U$ then yields

$$\begin{aligned} \sum_{u \in U} u \cdot P(|X| = u) &= \sum_{u \in U} |u| \cdot P(X = u) + |-u| \cdot P(X = -u) \\ &= \sum_{t \in \hat{T}} |t| \cdot P(X = t) \\ &= \sum_{t \in T} |t \cdot P(X = t)|. \end{aligned}$$

Therefore the series describing $E[X]$ is absolutely convergent exactly when $E[|X|] < \infty$. ■

4.1.1 *Properties of the Expected Value*

We will eventually wish to calculate the expected values of functions of multiple random variables. Of particular interest to statistics is an understanding of expected values of sums and averages of i.i.d. sequences. That understanding will be made easier by first learning something about how expected values behave for simple combinations of variables.

THEOREM 4.1.7. *Suppose that X and Y are discrete random variables, both with finite expected value and both defined on the same sample space S. If a and b are real numbers then*

(1) $E[aX] = aE[X]$;

(2) $E[X + Y] = E[X] + E[Y]$; *and*

(3) $E[aX + bY] = aE[X] + bE[Y]$.

(4) If $X \geq 0$ *then* $E[X] \geq 0$.

Proof of (1) - If $a = 0$ then both sides of the equation are zero, so assume $a \neq 0$. We know that X is a function from S to some range U. So aX is also a random variable and its range is $T = \{au : u \in U\}$.
By definition $E[aX] = \sum_{t \in T} t \cdot P(aX = t)$, but because of how T is defined, adding values indexed by $t \in T$ is equivalent to adding values indexed by $u \in U$ where $t = au$. In other words

$$\begin{aligned} E[aX] &= \sum_{t \in T} t \cdot P(aX = t) \\ &= \sum_{u \in U} au \cdot P(aX = au) \\ &= a \cdot \sum_{u \in U} u \cdot P(X = u) \\ &= aE[X]. \end{aligned}$$

Proof of (2) - We are assuming that X and Y have the same domain, but they typically have different ranges. Suppose $X : S \to U$ and $Y : S \to V$. Then the random variable $X + Y$ is also defined on S and takes values in $T = \{u + v : u \in U, v \in V\}$. Therefore, adding values indexed by $t \in T$ is equivalent to adding values indexed by u and v as they range over U and V respectively. So,

$$\begin{aligned} E[X+Y] &= \sum_{t\in T} t \cdot P(X+Y=t) \\ &= \sum_{u\in U, v\in V} (u+v)\cdot P(X=u, Y=v) \\ &= \sum_{u\in U}\sum_{v\in V} (u+v)\cdot P(X=u, Y=v) \\ &= \sum_{u\in U}\sum_{v\in V} u\cdot P(X=u, Y=v) + \sum_{u\in U}\sum_{v\in V} v\cdot P(X=u, Y=v) \\ &= \sum_{u\in U}\sum_{v\in V} u\cdot P(X=u, Y=v) + \sum_{v\in V}\sum_{u\in U} v\cdot P(X=u, Y=v) \end{aligned}$$

where the rearrangement of summation is legitimate since the series converges absolutely. Notice that as u ranges over all of U the sets $(X = u, Y = v)$ partition the set $(Y = v)$ into disjoint pieces based on the value of X. Likewise the event $(X = u)$ is partitioned by $(X = u, Y = v)$ as v ranges over all values of $v \in V$. Therefore, as a disjoint union,

$$(Y=v) = \bigcup_{u\in U} (X=u, Y=v) \qquad \text{and} \qquad (X=u) = \bigcup_{v\in V} (X=u, Y=v),$$

and so

$$P(Y=v) = \sum_{u\in U} P(X=u, Y=v) \text{ and } P(X=u) = \sum_{v\in V} P(X=u, Y=v).$$

From there the proof may be completed, since

$$\begin{aligned} E[X+Y] &= \sum_{u\in U} u \sum_{v\in V} P(X=u, Y=v) + \sum_{v\in V} v \sum_{u\in U} P(X=u, Y=v) \\ &= \sum_{u\in U} u\cdot P(X=u) + \sum_{v\in V} v\cdot P(Y=v) \\ &= E[X] + E[Y]. \end{aligned}$$

Proof of (3) - This is an easy consequence of (1) and (2). From (2) the expected value $E[aX + bY]$ may be rewritten as $E[aX] + E[bY]$. From there, applying (1) shows this is also equal to $aE[X] + bE[Y]$. (Using induction this theorem may be extended to any finite linear combination of random variables, a fact which we leave as an exercise below).
Proof of (4) - We know that X is a function from S to T where $t \in T$ implies that $t \geq 0$. As,

$$E[X] = \sum_{t\in T} t \cdot P(X=t),$$

it follows by definition of series (in the case T is countable) that $E[X] \geq 0$. ■

EXAMPLE 4.1.8. What is the average value of the sum of a pair of dice?

To answer this question by appealing to the definition of expected value would require summing over the eleven possible outcomes $\{2, 3, \ldots, 12\}$ and computing the probabilities of each of those outcomes. Theorem 4.1.7 makes things much simpler. We began this section by noting that a single die roll has an expected value of 3.5. The sum of two dice is $X + Y$ where each of X and Y represents the outcome of a single die. So the average value of the sum of a pair of dice is $E[X + Y] = E[X] + E[Y] = 3.5 + 3.5 = 7$. ■

EXAMPLE 4.1.9. Consider a game in which a player might either gain or lose money based on the result. A game is considered "fair" if it is described by a random variable with an expected value of zero. Such a game is fair in the sense that, on average, the player will have no net change in money after playing.

Suppose a particular game is played with one player (the roller) throwing a die. If the die comes up an even number, the roller wins that dollar amount from his opponent. If the die is odd, the roller wins nothing. Obviously the game as stated is not "fair" since the roller cannot lose money and may win something. How much should the roller pay his opponent to play this game in order to make it a fair game?

Let X be the amount of money the rolling player gains by the result on the die. The set of possible outcomes is $T = \{0, 2, 4, 6\}$ and it should be routine at this point to verify that $E[X] = 2$. Let c be the amount of money the roller should pay to play in order to make the game fair. Since X is the amount of money gained by the roll, the net change of money for the roller is $X - c$ after accounting for how much was paid to play. A fair game requires

$$0 = E[X - c] = E[X] - E[c] = 2 - c.$$

So the roller should pay his opponent \$2 to make the game fair. ■

4.1.2 *Expected Value of a Product*

Theorem 4.1.7 showed that $E[X + Y] = E[X] + E[Y]$. It is natural to ask whether a similar rule exists for the product of variables. While it is not generally the case that the expected value of a product is the product of the expected values, if X and Y happen to be independent, the result is true.

THEOREM 4.1.10. *Suppose that X and Y are discrete random variables, both with finite expected value and both defined on the same sample space S. If X and Y are independent, then $E[XY] = E[X]E[Y]$.*

Proof - Suppose $X : S \to U$ and $Y : S \to V$. Then the random variable XY takes values in $T = \{uv : u \in U, v \in V\}$. So,

$$\begin{aligned} E[XY] &= \sum_{t \in T} t \cdot P(XY = t) \\ &= \sum_{u \in U} \sum_{v \in V} (uv) \cdot P(X = u, Y = v) \\ &= \sum_{u \in U} \sum_{v \in V} (uv) \cdot P(X = u) P(Y = v) \\ &= \sum_{u \in U} u \cdot P(X = u) \sum_{v \in V} v \cdot P(Y = v) \\ &= \left(\sum_{u \in U} u \cdot P(X = u) \right) \left(\sum_{v \in V} v \cdot P(Y = v) \right) \\ &= E[X]E[Y]. \end{aligned}$$

Before showing an example of how this theorem might be used, we provide a demonstration that the result will not typically hold without the assumption of independence.

Example 4.1.11. Let $X \sim \text{Uniform}(\{1, 2, 3\})$ and let $Y = 4 - X$. It is easy to verify $Y \sim \text{Uniform}(\{1, 2, 3\})$ as well, but X and Y are certainly dependent. A routine computation shows $E[X] = E[Y] = 2$, and so $E[X]E[Y] = 4$.

However, the random variable XY can only take on two possible values. It may equal 3 (if either $X = 1$ and $Y = 3$ or vica versa) or it may equal 4 (if $X = Y = 2$). So, $P(XY = 3) = \frac{2}{3}$ and $P(XY = 4) = \frac{1}{3}$. Therefore,

$$E[XY] = 3(\frac{2}{3}) + 4(\frac{1}{3}) = \frac{10}{3} \neq 4.$$

The conclusion of Theorem 4.1.10 fails since X and Y are dependent. ■

Example 4.1.12. Suppose an insurance company assumes that, for a given month, both the number of customer claims X and the average cost per claim Y are independent random variables. Suppose further the company is able to estimate that $E[X] = 100$ and $E[Y] = \$1,250$. How should the company estimate the total cost of all claims that month?

The total cost should be the number of claims times the average cost per claim, or XY. Using Theorem 4.1.10 the expected value of XY is simply the product of the separate expected values.

$$E[XY] = E[X]E[Y] = 100 \cdot \$1,250 = \$125,000.$$

Notice, though, that the assumption of independence played a critical role in this computation. Such an assumption might not be valid for many practical problems. Consider, for example, if a weather event such as a tornado tends to cause both a larger-than-average number of claims and also a larger-than-average value per claim. This could cause the variables X and Y to be dependent and, in such a case, estimating the total cost would not be as simple as taking the product of the separate expected values. ■

4.1.3 *Expected Values of Common Distributions*

A quick glance at the definition of expected value shows that it only depends on the distribution of the random variable. Therefore one can compute the expected values for the various common distributions we defined in the previous chapter.

EXAMPLE 4.1.13. **(Expected Value of a Bernoulli(p))**
Let $X \sim \text{Bernoulli}(p)$. So $P(X=0) = 1-p$ and $P(X=1) = p$.
Therefore $E[X] = 0(1-p) + 1(p) = p$. ■

EXAMPLE 4.1.14. **(Expected Value of a Binomial(n,p))**
We will show two ways to calculate this expected value – the first is more computationally complicated, but follows from the definition of the binomial distribution directly; the second is simpler, but requires using the relationship between the binomial and Bernoulli random variables. In algebraic terms, if $Y \sim \text{Binomial}(n,p)$ then

$$\begin{aligned}
E[Y] &= \sum_{k=0}^{n} k \cdot P(Y=k) \\
&= \sum_{k=1}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= np \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} p^{k} (1-p)^{(n-1)-k}
\end{aligned}$$

where the last equality is a shift of variables. But now, by the binomial theorem, the sum $\sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{(n-1)-k}$ is equal to 1 and therefore $E[Y] = np$.

Alternatively, recall that the binomial distribution first came about as the total number of successes in n independent Bernoulli trials. Therefore a $\text{Binomial}(n,p)$ distribution results from adding together n independent $\text{Bernoulli}(p)$ random variables. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\text{Bernoulli}(p)$ and let $Y = X_1 + X_2 + \cdots + X_n$. Then $Y \sim \text{Binomial}(n,p)$ and

$$\begin{aligned}
E[Y] &= E[X_1 + X_2 + \cdots + X_n] \\
&= E[X_1] + E[X_2] + \cdots + E[X_n] \\
&= p + p + \cdots + p = np.
\end{aligned}$$

This also provides the answer to part (d) of Example 2.1.2. The expected number of successes in a series of n independent $\text{Bernoulli}(p)$ trials is np. ■

In the next example we will calculate the expected value of a geometric random variable. The computation illustrates a common technique from calculus for simplifying power series by differentiating the sum term-by-term in order to rewrite a complicated series in a simpler way.

EXAMPLE 4.1.15. **(Expected Value of a Geometric(p))**
If $X \sim Geometric(p)$ and $0 < p < 1$, then

$$E[X] = \sum_{k=1}^{\infty} k \cdot p(1-p)^{k-1}$$

To evaluate the sum of the series we will need to work the partial sums of the same. For any $n \geq 1$, let

$$\begin{aligned} T_n &= \sum_{k=1}^{n} kp(1-p)^{k-1} = \sum_{k=1}^{n} k(1-(1-p))(1-p)^{k-1} \\ &= \sum_{k=1}^{n} k(1-p)^{k-1} - \sum_{k=1}^{n} k(1-p)^{k} \\ &= \sum_{k=1}^{n} (1-p)^{k-1} - n(1-p)^n = \frac{1-(1-p)^n}{p} - n(1-p)^n. \end{aligned}$$

Using standard results from analysis we know that for $0 < p < 1$,

$$\lim_{n\to\infty} (1-p)^n = 0 \text{ and } \lim_{n\to\infty} n(1-p)^n = 0.$$

Therefore $T_n \to \frac{1}{p}$ as $n \to \infty$. Hence

$$E[X] = \frac{1}{p}.$$

For instance, suppose we wanted to know on average how many rolls of a die it would take before we observed a 5. Each roll is a Bernoulli trial with a probability $\frac{1}{6}$ of success. The time it takes to observe the first success is distributed as a $Geometric(\frac{1}{6})$ and so has expected value $\frac{1}{1/6} = 6$. On average it should take six rolls before observing this outcome. ■

EXAMPLE 4.1.16. **(Expected Value of a Poisson(λ))**

We can make a reasonable guess at the expected value of a Poisson(λ) random variable by recalling that such a distribution was created to approximate a binomial when n was large and p was small. The parameter $\lambda = np$ remained fixed as we took a limit. Since we showed above that a $Binomial(n, p)$ has an expected value of np, it seems plausible that a $Poisson(\lambda)$ should have an expected value of λ. This is indeed true and it is possible to prove the fact by using the idea that the Poisson random variable is the limit of a sequence of binomial random variables. However, this proof requires an understanding of how limits and expected values interact, a concept that has not yet been introduced in the text. Instead we leave a proof based on a direct algebraic computation as Exercise 4.1.12.

Taking the result as a given, we will illustrate how this expected value might be used for an applied problem. Suppose an insurance company wants to model catastrophic floods using a Poisson(λ) random variable. Since floods are rare in any given year, and since the company is considering what might occur over a long span of years, this may be a reasonable assumption.

As its name implies a "50-year flood" is a flood so substantial that it should occur, on average, only once every fifty years. However, this is just an average; it may be possible to have two "50-year floods" in consecutive years, though such an event would be quite rare. Suppose the insurance company wants to know how likely it is that there will be two or more "50-year floods" in the next decade, how should this be calculated?

There is an average of one such flood every fifty years, so by proportional reasoning, in the next ten years there should be an average of 0.2 floods. In other words, the number of floods in the next ten years should a random variable $X \sim Poisson(0.2)$ and we wish to calculate $P(X \geq 2)$.

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-0.2} - e^{-0.2}(0.2) \\ &\approx 0.0002. \end{aligned}$$

So assuming the Poisson random variable is an accurate model, there is only about a 0.02% chance that two or more such disastrous floods would occur in the next decade. ■

For a hypergeometric random variable, we will demonstrate another proof technique common to probability. An expected value may involve a complicated (or infinite) sum which must be computed. However, this sum includes within it the probabilities of each outcome of the random variable, and those probabilities must therefore add to 1. It is sometimes possible to simplify the sum describing the expected value using the fact that a related sum is already known.

EXAMPLE 4.1.17. **(Expected Value of a HyperGeo(N,r,m))** Let m and r be positive integers an d let N be an integer for which $N > \max\{m,r\}$. Let X be a random variable with $X \sim \text{HyperGeo}(N,r,m)$. To calculate the expected value of X, we begin with two facts. The first is an identity involving combinations. If $n \geq k > 0$ then

$$\begin{aligned}\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\ &= \frac{n}{k}\frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} \\ &= \frac{n}{k}\binom{n-1}{k-1}.\end{aligned}$$

The second comes from the consideration of the probabilities associated with a HyperGeo$(N-1,r-1,m-1)$ distribution. Specifically, as k ranges over all possible values of such a distribution, we have

$$\sum_k \frac{\binom{r-1}{k}\binom{(N-1)-(r-1)}{(m-1)-k}}{\binom{N-1}{m-1}} = 1$$

since this is the sum over all outcomes of the random variable.

To calculate $E[X]$, let j range over the possible values of X. Recall that the minimum value of j is $\max\{0, m-(N-r)\}$ and the maximum value of j is $\min\{r,m\}$. Now let $k = j-1$. This means that the maximum value for k is $\min\{r-1, m-1\}$. If the minimum value for j was $m-(N-r)$ then the minimum value for k is $m-(N-r)-1 = ((m-1)-((N-1)-(r-1)))$. If the minimum value for j was 0 then the minimum value for k is -1.

The key to the computation is to note that as j ranges over all of the values of X, the values of k cover all possible values of a HyperGeo$(N-1, m-1, r-1)$ distribution. In fact, the only possible value k may assume that is not in the range of such a distribution is if $k = -1$ as a minimum value. Now,

$$E[X] = \sum_j j \cdot \frac{\binom{r}{j}\binom{N-r}{m-j}}{\binom{N}{m}},$$

and if $j = 0$ is in the range of X, then that term of the sum is zero and it may be deleted without affecting the value. That is equivalent to deleting the $k = -1$ term, so the remaining values of

k exactly describe the range of a HyperGeo$(N-1, m-1, r-1)$ distriubtion. From there, the expected value may be calculated as

$$\begin{aligned} E[X] &= \sum_j j \cdot \frac{\binom{r}{j}\binom{N-r}{m-j}}{\binom{N}{m}} \\ &= \sum_j j \cdot \frac{\frac{r}{j}\binom{r-1}{j-1}\binom{(N-1)-(r-1)}{(m-1)-(j-1)}}{\frac{N}{m}\binom{N-1}{m-1}} \\ &= (\frac{rm}{N}) \cdot \sum_j \frac{\binom{r-1}{j-1}\binom{(N-1)-(r-1)}{(m-1)-(j-1)}}{\binom{N-1}{m-1}} \\ &= (\frac{rm}{N}) \cdot \sum_k \frac{\binom{r-1}{k}\binom{(N-1)-(r-1)}{(m-1)-k}}{\binom{N-1}{m-1}} \\ &= (\frac{rm}{N}) \cdot (1) = \frac{rm}{N}. \end{aligned}$$

This nearly completes the goal of calculating the expected values of hypergeometric distributions. The only remaining issues are the cases when $m = 0$ and $r = 0$. Since the hypergoemetric distribution was only defined when m and r were non-negative integers, and since the proof above requires the consideration of such a distribution for the values $m-1$ and $r-1$, the remaining cases must be handled separately. However, they are fairly easy and yield the same result, a fact we leave it to the reader to verify. ■

4.1.4 *Expected Value of* $f(X_1, X_2, \ldots, X_n)$

As we have seen previously, if X is a random variable and if f is a function defined on the possible outputs of X, then $f(X)$ is a random variable in its own right. The expected value of this new random variable may be computed in the usual way from the distribution of $f(X)$, but it is an extremely useful fact that it may also be computed from the distribution of X itself. The next example and theorems illustrate this fact.

EXAMPLE 4.1.18. Returning to a setting first seen in Example 3.3.1 we will let $X \sim \text{Uniform}(\{-2, -1, 0, 1, 2\})$, and let $f(x) = x^2$. How may $E[f(X)]$ be calculated?

We will demonstrate this in two ways – first by appealing directly to the definition, and then using the distribution of X instead of the distribution of $f(X)$. To use the definition of expected value, recall that $f(X) = X^2$ takes values in $\{0, 1, 4\}$ with the following probabilities: $P(f(X) = 0) = \frac{1}{5}$ while $P(f(X) = 1) = P(f(X) = 4) = \frac{2}{5}$. Therefore,

$$E[f(X)] = 0(\frac{1}{5}) + 1(\frac{2}{5}) + 4(\frac{2}{5}) = 2.$$

However, the values of $f(X)$ are completely determined from the values of X. For instance, the event $(f(X) = 4)$ had a probability of $\frac{2}{5}$ because it was the disjoint union of two other events $(X = 2) \cup (X = -2)$, each of which had probability $\frac{1}{5}$. So the term $4(\frac{2}{5})$ in the computation above could equally well have been thought of in two pieces

$$\begin{aligned} 4 \cdot P(f(X) = 4) &= 4 \cdot P((X = 2) \cup (X = -2)) \\ &= 4 \cdot (P(X = 2) + P(X = -2)) \\ &= 4 \cdot P(X = 2) + 4 \cdot P(X = -2) \\ &= 2^2 \cdot P(X = 2) + (-2)^2 \cdot P(X = -2), \end{aligned}$$

where the final expression emphasizes that the outcome of 4 resulted either from 2^2 or $(-2)^2$ depending on the value of X. Following a similar plan for the other values of $f(X)$ allows $E[f(X)]$ to be calcualted directly from the probabilities of X as

$$\begin{aligned} E[f(X)] &= (-2)^2 \cdot P(X=-2) + (-1)^2 \cdot P(X=-1) + 0^2 \cdot P(X=0) \\ &\qquad +1^2 \cdot P(X=1) + 2^2 \cdot P(X=2) \\ &= 4(\frac{1}{5}) + 1(\frac{1}{5}) + 0(\frac{1}{5}) + 1(\frac{1}{5}) + 4(\frac{1}{5}) \\ &= 2, \end{aligned}$$

which gives the same result as the previous computation. ■

The technique of the example above works for any functions as demonstrated by the next two theorems. We first state and prove a version for functions of a single random variable and then deal with the multivariate case.

THEOREM 4.1.19. *Let $X : S \to T$ be a discrete random variable and define a function $f : T \to U$. Then the expected value of $f(X)$ may be computed as*

$$E[f(X)] = \sum_{t \in T} f(t) \cdot P(X = t).$$

Proof - By definition $E[f(X)] = \sum_{u \in U} u \cdot P(f(X) = u)$. However, as in the previous example, the event $(f(X) = u)$ may be partitioned according to the input values of X which cause $f(X)$ to equal u. Recall that $f^{-1}(u)$ describes the set of values in T which, when input into the function f, produce the value u. That is, $f^{-1}(u) = \{t \in T : f(t) = u\}$. Therefore,

$$(f(X) = u) = \bigcup_{t \in f^{-1}(u)} (X = t), \quad \text{and so}$$

$$P(f(X) = u) = \sum_{t \in f^{-1}(u)} P(X = t).$$

Putting this together with the definition of $E[f(X)]$ shows

$$\begin{aligned} E[f(X)] &= \sum_{u \in U} u \cdot P(f(X) = u) \\ &= \sum_{u \in U} u \cdot \sum_{t \in f^{-1}(u)} P(X = t) \\ &= \sum_{u \in U} \sum_{t \in f^{-1}(u)} u \cdot P(X = t) \\ &= \sum_{u \in U} \sum_{t \in f^{-1}(u)} f(t) \cdot P(X = t) \\ &= \sum_{t \in T} f(t) \cdot P(X = t), \end{aligned}$$

where the final step is simply the fact that $T = f^{-1}(U)$ and so summing over the values of $t \in T$ is equivalent to grouping them together in the sets $f^{-1}(u)$ and summing over all values in U that may be achieved by $f(X)$. ■

THEOREM 4.1.20. *Let X_1, X_2, $\ldots X_n$ be random variables defined on a common sample space S. The X_j variables may have different ranges, so let $X_j : S \to T_j$. Let f be a function defined for all possible outputs of the X_j variables. Then*

$$E[f(X)] = \sum_{t_1 \in T_1, \ldots t_n \in T_n} f(t_1, \ldots, t_n) \cdot P(X_1 = t_1, \ldots, X_n = t_n).$$

The proof is nearly the same as for the one-variable case. The only diference is that $f^{-1}(u)$ is now a set of vectors of values $(t_1, \ldots, t_n)$, so that the event $(f(X) = u)$ decomposes into events of the form $(X_1 = t_1, \ldots, X_n = t_n)$. However, this change does not interfere with the logic of the proof. We leave the details to the reader.

EXERCISES

Ex. 4.1.1. Let X, Y be discrete random variables. Suppose $X \leq Y$ then show that $E[X] \leq E[Y]$.

Ex. 4.1.2. A lottery is held every day, and on any given day there is a 30% chance that someone will win, with each day independent of every other. Let X denote the random variable describing the number of times in the next five days that the lottery will be won.

(a) What type of random variable (with what parameter) is X?

(b) On average (expected value), how many times in the next five days will the lottery be won?

(c) When the lottery occurs for each of the next five days, what is the most likely number (mode) of days there will be a winner?

(d) How likely is it the lottery will be won in either one or two of the next five days?

Ex. 4.1.3. A game show contestant is asked a series of questions. She has a probability of 0.88 of knowing the answer to any given question, independently of every other. Let Y denote the random variable describing the number of questions asked until the contestant does not know the correct answer.

(a) What type of random variable (with what parameter) is Y?

(b) On average (expected value), how many questions will be asked until the first question for which the contestant does not know the answer?

(c) What is the most likely number of questions (mode) that will be asked until the contestant does not know a correct answer?

(d) If the contestant is able to answer twelve questions in a row, she will win the grand prize. How likely is it that she will know the answers to all twelve questions?

Ex. 4.1.4. Sonia sends out invitations to eleven of her friends to join her on a hike she's planning. She knows that each of her friends has a 59% chance of deciding to join her independently of each other. Let Z denote the number of friends who join her on the hike.

(a) What type of random variable (with what parameter) is Z?

(b) What is the average (expected value) number of her friends that will join her on the hike?

(c) What is the most likely number (mode) of her friends that will join her on the hike?

(d) How do your answers to (b) and (c) change if each friend has only a 41% chance of joining her?

Ex. 4.1.5. A player rolls three dice and earns \$1 for each die that shows a 6. How much should the player pay to make this a fair game?

Ex. 4.1.6. **("The St.Petersburg Paradox")** Suppose a game is played whereby a player begins flipping a fair coin and continues flipping it until it comes up heads. At that time the player wins a 2^n dollars where n is the total number of times he flipped the coin. Show that there is no amount of money the player could pay to make this a fair game. (Hint: See Example 4.1.4).

Ex. 4.1.7. Two different investment strategies have the following probabilities of return on \$10,000.

Strategy A has a 20% chance of returning \$14,000, a 35% chance of returning \$12,000, a 20% chance of returning \$10,000, a 15% chance of returning \$8,000, and a 10% chance of returning only \$6,000.

Strategy B has a 25% chance of returning \$12,000, a 35% chance of returning \$11,000, a 25% chance of returning \$10,000, and a 15% chance of returning \$9,000.

(a) Which strategy has the larger expected value of return?

(b) Which strategy is more likely to produce a positive return on investment?

(c) Is one strategy clearly preferable to the other? Explain your reasoning.

Ex. 4.1.8. Calculate the expected value of a Uniform$(\{1, 2, \ldots, n\})$ random variable by following the steps below.

(a) Prove the numerical fact that $\sum_{j=1}^{n} j = \frac{n^2+n}{2}$. (Hint: There are many methods to do this. One uses induction).

(b) Use (a) to show that if $X \sim$ Uniform$(\{1, 2, \ldots, n\})$, then $E[X] = \frac{n+1}{2}$.

Ex. 4.1.9. Use induction to extend the result of Theorem 4.1.7 by proving the following:

If $X_1, X_2, \ldots, X_n$ are random variables with finite expectation all defined on the same sample space S and if $a_1, a_2, \ldots a_n$ are real numbers, then

$$E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n].$$

Ex. 4.1.10. Suppose X and Y are random variables for which X has finitie expected value and Y has infinite expected value. Prove that $X + Y$ has infinite expected value.

Ex. 4.1.11. Suppose X and Y are random variables. Suppose $E[X] = \infty$ and $E[Y] = -\infty$.

(a) Provide an example to show that $E[X + Y] = \infty$ is possible.

(b) Provide an example to show that $E[X + Y] = -\infty$ is possible.

(c) Provide an example to show that $E[X + Y]$ may have finite expected value.

Ex. 4.1.12. Let $X \sim Poisson(\lambda)$.

(a) Write an expression for $E[X]$ as an infinite sum.

(b) Every non-zero term in your answer to (a) should have a λ in it. Factor this λ out and explain why the remaining sum equals 1. (Hint: One way to do this is through the use of infinite series. Another way is to use the idea from Example 4.1.17).

Ex. 4.1.13. A daily lottery is an event that many people play, but for which the likelihood of any given person winning is very small, making a Poisson approximation appropriate. Suppose a daily lottery has, on average, two winners every five weeks. Estimate the probability that next week there will be more than one winner.

4.2 VARIANCE AND STANDARD DEVIATION

As a single number, the average of a random variable may or may not be a good approximation of the values that variable is likely to produce. For example, let X be defined such that $P(X = 10) = 1$, let Y be defined so that $P(Y = 9) = P(Y = 11) = \frac{1}{2}$, and let Z be defined such that $P(Z = 0) = P(Z = 20) = \frac{1}{2}$. It is easy to check that all three of these random variables have an expected value of 10. However the number 10 exactly describes X, is always off from Y by an absolute value of 1 and is always off from Z by an absolute value of 10.

It is useful to be able to quantify how far away a random variable typically is from its average. Put another way, if we think of the expected value as somehow measuring the "center" of the random variable, we would like to find a way to measure the size of the "spread" of the variable about its center. Quantities useful for this are the variance and standard deviation.

DEFINITION 4.2.1. *Let X be a random variable with finite expected value. Then the variance of the random variable is written as $Var[X]$ and is defined as*

$$Var[X] = E[(X - E[X])^2]$$

The standard deviation of X is written as $SD[X]$ and is defined as

$$SD[X] = \sqrt{Var[X]}$$

Notice that $Var[X]$ is the average of the square distance of X from its expected value. So if X has a high probability of being far away from $E[X]$ the variance will tend to be large, while if X is very near $E[X]$ with high probability the variance will tend to be small. In either case the variance is the expected value of a squared quantity, and as such is always non-negative. Therefore $SD[X]$ is defined whenever $Var[X]$ is defined.

If we were to associate units with the random variable X (say $meters$), then the units of $Var[X]$ would be $meters^2$ and the units of $SD[X]$ would be $meters$. We will see that the standard deviation is more meaningful as a measure of the "spread" of a random variable while the variance tends to be a more useful quantity to consider when carrying out complex computations.

Informally we will view the standard deviation as a typical distance from average. So if X is a random variable and we calculate that $E[X] = 12$ and $SD[X] = 3$, we might say, "The variable X will typically take on values that are in or near the range $9 - 15$, one standard deviation either side of the average". A goal of this section is to make that language more precise, but at this point it will help with intuition to understand this informal view.

The variance and standard deviation are described in terms of the expected value. Therefore $Var[X]$ and $SD[X]$ can only be defined if $E[X]$ exists as a real number. However, it is possible that $Var[X]$ and $SD[X]$ could be infinite even if $E[X]$ is finite (see Exercises). In practical terms, if X has a finite expected value and infinite standard deviation, it means that the random variable

has a clear average, but is so spread out that any finite number underestimates the typical distance of the random variable from its average.

EXAMPLE 4.2.2. As above, let X be a constant varaible with $P(X = 10) = 1$. Let Y be such that $P(Y = 9) = P(Y = 11) = \frac{1}{2}$ and let Z be such that $P(Z = 0) = P(Z = 20) = \frac{1}{2}$.

Since X always equals $E[X]$, the quantity $(X - E[X])^2$ is always zero and we can conclude that $Var[X] = 0$ and $SD[X] = 0$. This makes sense given the view of $SD[X]$ as an estimate of how spread out the variable is. Since X is constant it is not at all spread out and so $SD[X] = 0$.

To calculate $Var[Y]$ we note that $(Y - E[Y])^2$ is always equal to 1. Therefore $Var[Y] = 1$ and $SD[Y] = 1$. Again this reaffirms the informal description of the standard deviation; the typical distance between Y and its average is 1.

Likewise $(Z - E[Z])^2$ is always equal to 100. Therefore $Var[Z] = 100$ and $SD[Z] = 10$. The typical distance between Z and its average is 10. ■

EXAMPLE 4.2.3. What are the variance and standard deviation of a die roll?

Before we carry out the calculation, let us use the informal idea of standard deviation to estimate an answer and help build intuition. We know the average of a die roll is 3.5. The closest a die could possibly be to this average is 0.5 (if it were to roll a 3 or a 4) and the furthest it could possibly be is 2.5 (if it were to roll a 1 or a 6). Therefore the standard deviation, a typical distance from average, should be somewhere between 0.5 and 2.5.

To calculate the quantity exactly, let X represent the roll of a die. By definition, $Var[X] = E[(X - 3.5)^2]$, and the values that $(X - 3.5)^2$ may assume are determined by the six values X may take on.

$$\begin{aligned} Var[X] &= E[(X - 3.5)^2] \\ &= \frac{1}{6}(2.5)^2 + \frac{1}{6}(1.5)^2 + \frac{1}{6}(0.5)^2 + \frac{1}{6}(-0.5)^2 + \frac{1}{6}(-1.5)^2 + \frac{1}{6}(-2.5)^2 \\ &= \frac{35}{12}. \end{aligned}$$

So, $SD[X] = \sqrt{\frac{35}{12}} \approx 1.71$ which is near the midpoint of the range of our estimate above. ■

4.2.1 *Properties of Variance and Standard Deviation*

THEOREM 4.2.4. *Let $a \in \mathbb{R}$ and let X be a random variable with finite variance (and thus, with finite expected value as well). Then,*

(a) $Var[aX] = a^2 \cdot Var[X]$;

(b) $SD[aX] = |a| \cdot SD[X]$;

(c) $Var[X + a] = Var[X]$; *and*

(d) $SD[X + a] = SD[X]$.

Proof of (a) and (b) - $Var[aX] = E[(aX - E[aX])^2]$. Using known properties of expected value this may be rewritten as

$$\begin{aligned} Var[aX] &= E[(aX - aE[X])^2] \\ &= E[a^2(X - E[X])^2] \\ &= a^2 E[(X - E[X])^2] \\ &= a^2 Var[X]. \end{aligned}$$

That concludes the proof of (a). The result from (b) follows by taking square roots of both sides of this equation.
Proof of (c) and (d) - (See Exercises) ■

The variance may also be computed using a different (but equivalent) formula if $E[X]$ and $E[X^2]$ are known.

THEOREM 4.2.5. *Let X be a random variable for which $E[X]$ and $E[X^2]$ are both finite. Then*

$$Var[X] = E[X^2] - (E[X])^2 .$$

Proof -

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + (E[X])^2] \\ &= E[X^2] - 2E[XE[X]] + E[(E[X])^2]. \end{aligned}$$

But $E[X]$ is a constant, so

$$\begin{aligned} Var[X] &= E[X^2] - 2E[XE[X]] + E[(E[X])^2] \\ &= E[X^2] - 2E[X]E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2 . \end{aligned}$$

■

In statistics we frequently want to consider the sum or average of many random variables. As such it is useful to know how the variance of a sum relates to the variances of each variable separately. Toward that goal we have

THEOREM 4.2.6. *If X and Y are independent random variables, both with finite expectation and finite variance, then*

(a) $Var[X+Y] = Var[X] + Var[Y]$*; and*

(b) $SD[X+Y] = \sqrt{(SD[X])^2 + (SD[Y])^2}$.

Proof - Using Theorem 4.2.5,

$$\begin{aligned} Var[X+Y] &= E[(X+Y)^2] - (E[X+Y])^2 \\ &= E[X^2 + 2XY + Y^2] - \left((E[X])^2 + 2E[X]E[Y] + (E[Y])^2\right) \\ &= E[X^2] + 2E[XY] + E[Y^2] - (E[X])^2 - 2E[X]E[Y] - (E[Y])^2. \end{aligned}$$

But by Theorem 4.1.10, $E[XY] = E[X]E[Y]$ since X and Y are independent. So,

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

Part (b) follows immediately after rewriting the variances in terms of standard deviations and taking square roots. As with expected values, this theorem may be generalized to a sum of any finite number of independent random variables using induction. The proof of that fact is left as Exercise 4.2.11. ■

EXAMPLE 4.2.7. What is the standard deviation of the sum of two dice?

We previously found that if X represents one die, then $Var[X] = \frac{35}{12}$. If X and Y are two independent dice, then $Var[X+Y] = Var[X] + Var[Y] = \frac{35}{12} + \frac{35}{12} = \frac{35}{6}$. Therefore $SD[X+Y] = \sqrt{\frac{35}{6}} \approx 2.42$. ■

4.2.2 *Variances of Common Distributions*

As with expected value, the variances of the common discrete random variables can be calculated from their corresponding distributions.

EXAMPLE 4.2.8. **(Variance of a Bernoulli(p))**
Let $X \sim Bernoulli(p)$. We have already calculated that $E[X] = p$. Since X only takes on the values 0 or 1 it is always true that $X^2 = X$. Therefore $E[X^2] = E[X] = p$.
So, $Var[X] = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$. ■

EXAMPLE 4.2.9. **(Variance of a Binomial(n,p))**
We will calculate the variance of a binomial using the fact that it may be viewed as the sum of n independent Bernoulli random variables. A strictly algebraic computation is also possible (see Exercises).

Let $X_1, X_2, \ldots, X_n$ be independent $Bernoulli(p)$ random variables. Therefore, if $Y = X_1 + X_2 + \cdots + X_n$ then $Y \sim \text{Binomial}(n, p)$ and

$$\begin{aligned} Var[Y] &= Var[X_1 + X_2 + \cdots + X_n] \\ &= Var[X_1] + Var[X_2] + \cdots + Var[X_n] \\ &= p(1-p) + p(1-p) + \cdots + p(1-p) \\ &= np(1-p). \end{aligned}$$

For an application of this computation we return to the idea of sampling from a population where some members of the population have a certain characteristic and others do not. The goal is to provide an estimate of the number of people in the sample that have the characteristic. For this example, suppose we were to randomly select 100 people from a large city in which 20% of the population works in a service industry. How many of the 100 people from our sample should we expect to be service industry workers?

If the sampling is done without replacement (so we cannot pick the same person twice), then strictly speaking the desired number would be described by a hypergeometric random variable. However, we have also seen that there is little difference between the binomial and hypergeometric distributions when the size of the sample is small relative to the size of the population. So since the sample is only 100 people from a "large city", we will assume this situation is modeled by a binomial random variable. Specifically, since 20% of the population consits of service workers, we will assume $X \sim Binomial(100, 0.2)$.

The simplest way to answer to the question of how many service industy workers to expect within the sample is to compute the expected value of X. In this case $E[X] = 100(0.2) = 20$, so we should expect around 20 of the 100 people in the sample to be service workers. However, this is an incomplete answer to the question since it only provides an average value; the actual number of service workers in the sample is probably not going to be exactly 20, it's only likely to be around 20 on average. A more complete answer to the question would give an estimate as to how far away from 20 the actual value is likely to be. But this is precisely what the standard deviation describes – an estimate of the likely difference between the actual result of the random variable and its expected value.

In this case $Var[X] = 100(0.2)(0.8) = 16$ and so $SD[X] = \sqrt{16} = 4$. This means that the actual number of service industry workers in the sample will typically be about 4 or so away from the expected value of 20, so a more complete answer to the question would be "The sample is likely to have around $16 - 24$ service workers in it". That is not to say that the actual number of service workers is guaranteed to fall in the that range, but the range provides s a sort of likely

error associated with the estimate of 20. Results in the $16-24$ range should be considered fairly common. Results far outside that range, while possible, should be considered fairly unusual. ■

Recall in Example 4.1.17 we calculated $E[X]$ using a technique in which the sum describing $E[X]$ was computed based on another sum which only involved the distribution of X directly. This second sum equalled 1 since it simply added up the probabilities that X assumed each of its possible values. In a similar fashion, it is sometimes possible to calculate a sum describing $E[X^2]$ in terms of a sum for $E[X]$ which is already known. From that point, Theorem 4.2.5 may be used to calculate the variance and standard deviation of X. This technique will be illustrated in the next example in which we calculate the spread associated with a geometric random variable.

EXAMPLE 4.2.10. **(Variance of a Geometric(p))**

Let $0 < p < 1$. $X \sim Geometric(p)$ for which we know $E[X] = \frac{1}{p}$. Then,

$$E[X^2] = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1}$$

To evaluate the sum of the series we will need to work the partial sums of the same. For any $n \geq 1$, let

$$\begin{aligned}
S_n = \sum_{k=1}^{n} k^2 p(1-p)^{k-1} &= \sum_{k=1}^{n} k^2 (1-(1-p))(1-p)^{k-1} \\
&= \sum_{k=1}^{n} k^2 (1-p)^{k-1} - \sum_{k=1}^{n} k^2 (1-p)^k \\
&= 1 + \sum_{k=2}^{n} (2k-1)(1-p)^{k-1} - n^2(1-p)^n \\
&= 1 - \sum_{k=2}^{n} (1-p)^{k-1} + 2\sum_{k=2}^{n} k(1-p)^{k-1} - n^2(1-p)^n \\
&= 2 - \sum_{k=1}^{n} (1-p)^{k-1} + 2(-1 + \sum_{k=1}^{n} k(1-p)^{k-1}) - n^2(1-p)^n \\
&= -\frac{1-(1-p)^n}{p} + \frac{2}{p}\sum_{k=1}^{n} kp(1-p)^{k-1} - n^2(1-p)^n
\end{aligned}$$

Using standard results from analysis and result from Example 4.1.15 we know that for $0 < p < 1$,

$$\lim_{n\to\infty} \sum_{k=1}^{n} kp(1-p)^{k-1} = \frac{1}{p}, \lim_{n\to\infty} (1-p)^n = 0, \text{ and } \lim_{n\to\infty} n^2(1-p)^n = 0.$$

Therefore $S_n \to -\frac{1}{p} + \frac{2}{p^2}$ as $n \to \infty$. Hence

$$E[X^2] = -\frac{1}{p} + \frac{2}{p^2}.$$

Using Theorem 4.2.5 the variance may then be calculated as

$$\begin{aligned}
Var[X] &= E[X^2] - (E[X])^2 \\
&= \frac{2}{p^2} - \frac{1}{p} - (\frac{1}{p})^2 \\
&= \frac{1}{p^2} - \frac{1}{p}
\end{aligned}$$

A similar technique may be used for calculating the variance of a Poisson random variable, a fact which is left as an exercise. We finish this subsection with a computation of the variance of a hypergeometric distribution using an idea similar to how we calculated its expected value in Example 4.1.17.

EXAMPLE 4.2.11. Let m and r be positive integers and let N be an integer with $N > \max\{m, r\}$ and let $X \sim \text{HyperGeo}(N, r, m)$. To calculate $E[X^2]$, as j ranges over the values of X,

$$\begin{aligned} E[X^2] &= \sum_j j^2 \cdot \frac{\binom{r}{j}\binom{N-r}{m-j}}{\binom{N}{m}} \\ &= \sum_j j^2 \cdot \frac{\frac{r}{j}\binom{r-1}{j-1}\binom{(N-1)-(r-1)}{(m-1)-(j-1)}}{\frac{N}{m}\binom{N-1}{m-1}} \\ &= \left(\frac{rm}{N}\right) \sum_j j \cdot \frac{\binom{r-1}{j-1}\binom{(N-1)-(r-1)}{(m-1)-(j-1)}}{\binom{N-1}{m-1}} \\ &= \left(\frac{rm}{N}\right) \cdot \sum_k (k+1) \frac{\binom{r-1}{k}\binom{(N-1)-(r-1)}{(m-1)-k}}{\binom{N-1}{m-1}} \end{aligned}$$

where k ranges over the values of $Y \sim \text{HyperGeo}(N-1, r-1, m-1)$. Therefore,

$$\begin{aligned} E[X^2] &= \left(\frac{rm}{N}\right) E[Y+1] \\ &= \left(\frac{rm}{N}\right)(E[Y]+1) \\ &= \left(\frac{rm}{N}\right)\left(\frac{(r-1)(m-1)}{(N-1)}+1\right). \end{aligned}$$

Now the variance may be easily computed as

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 \\ &= \left(\frac{rm}{N}\right)\left(\frac{(r-1)(m-1)}{(N-1)}+1\right) - \left(\frac{rm}{N}\right)^2 \\ &= \frac{N^2rm - Nrm^2 - Nr^2m + r^2m^2}{N^2(N-1)}. \end{aligned}$$

As with the computation of expected value, the cases of $m = 0$ and $r = 0$ must be handled separately, but yield the same result.

4.2.3 *Standardized Variables*

Many random variables may be rescaled into a standard format by shifting them so that they have an average of zero and then rescaling them so that they have a variance (and standard deviation) of one. We introduce this idea now, though its chief importance will not be realized until later.

DEFINITION 4.2.12. *A standardized random variable X is one for which*

$$E[X] = 0 \quad \textit{and} \quad Var[X] = 1.$$

THEOREM 4.2.13. *Let X be a discrete random variable with finite expected value and finite, non-zero variance. Then $Z = \frac{X-E[X]}{SD[X]}$ is a standardized random variable.*

Proof - The expected value value of Z is

$$\begin{aligned} E[Z] &= E[\frac{X - E[X]}{SD[X]}] \\ &= \frac{E[X - E[X]]}{SD[X]} \\ &= \frac{E[X] - E[X]}{SD[X]} = 0 \end{aligned}$$

while the variance of Z is

$$\begin{aligned} Var[Z] &= Var[\frac{X - E[X]}{SD[X]}] \\ &= \frac{Var[X - E[X]]}{(SD[X])^2} \\ &= \frac{Var[X]}{Var[X]} = 1. \end{aligned}$$

For easy reference we finish off this section by providing a chart of values associated with common discrete distributions.

Distribution	**Expected Value**	**Variance**
Bernoulli(p)	p	$p(1-p)$
Binomial(n, p)	np	$np(1-p)$
Geometric(p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$
HyperGeo(N, r, m)	$\frac{rm}{N}$	$\frac{N^2rm - Nrm^2 - Nr^2m + r^2m^2}{N^2(N-1)}$
Poisson(λ)	λ	λ
Uniform($\{1, 2, \ldots, n\}$)	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$

EXERCISES

Ex. 4.2.1. A random variable X has a probability mass function given by

$$P(X = 0) = 0.2, P(X = 1) = 0.5, P(X = 2) = 0.2, \text{ and } P(X = 3) = 0.1.$$

Calculate the expected value and standard deviation of this random variable. What is the probability this random variable will produce a result more than one standard deviation from its expected value?

Ex. 4.2.2. Answer the following questions about flips of a fair coin.

(a) Calculate the standard deviation of the number of heads that show up in 100 flips of a fair coin.

(b) Show that if the number of coins is quadrupled (to 400) the standard deviation only doubles.

Ex. 4.2.3. Suppose we begin rolling a die, and let X be the number of rolls needed before we see the first 3.

(a) Show that $E[X] = 6$.

(b) Calculate $SD[X]$.

(c) Viewing $SD[X]$ as a typical distance of X from its expected value, would it seem unusual to roll the die more than nine times before seeing a 3?

(d) Calculate the actual probability $P(X > 9)$.

(e) Calculate the probability X produces a result within one standard deviation of its expected value.

Ex. 4.2.4. A key issue in statistical sampling is the determination of how much a sample is likely to differ from the population it came from. This exercise explores some of these ideas.

(a) Suppose a large city is exactly 50% women and 50% men and suppose we randomly select 60 people from this city as part of a sample. Let X be the number of women in the sample. What are the expected value and standard deviation of X? Given these values, would it seem unusual if fewer than 45% of the individuals in the sample were women?

(b) Repeat part (a), but now assume that the sample consists of 600 people.

Ex. 4.2.5. Calculate the variance and standard deviation of the value of the lottery ticket from Example 3.1.4.

Ex. 4.2.6. Prove parts (c) and (d) of Theorem 4.2.4.

Ex. 4.2.7. Let $X \sim Binomial(n, p)$. Show that for $0 < p < 1$, this random variable has the largest standard deviation when $p = \frac{1}{2}$.

Ex. 4.2.8. Follow the steps below to calculate the variance of a random variable with a Uniform$(\{1, 2, \ldots, n\})$ distribution.

(a) Prove that $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$. (Induction is one way to do this).

(b) Let $X \sim$ Uniform$(\{1, 2, \ldots, n\})$. Use (a) to calculate $E[X^2]$.

(c) Use (b) and the fact that $E[X] = \frac{n+1}{2}$ to calculate $Var[X]$.

Ex. 4.2.9. This exercise provides an example of a random variable with finite expected value, but infinite variance. Let X be a random variable for which $P(X = \frac{2^n}{n(n+1)}) = \frac{1}{2^n}$ for all integers $n \geq 1$.

(a) Prove that X is a well-defined variable by showing $\sum_{n=1}^{\infty} P(X = \frac{2^n}{n(n+1)}) = 1$.

(b) Prove that $E[X] = 1$.

(c) Prove that $Var[X]$ is infinite.

Ex. 4.2.10. Recall that the hypergeometric distribution was first developed to answer questions about sampling without replacement. With that in mind, answer the following questions using the chart of expected values and variances.

(a) Use the formula in the chart to calculate the variance of a hypergeometric distribution if $m = 0$. Explain this result in the context of what it means in terms of sampling.

(b) Use the formula in the chart to calculate the variance of a hypergeometric distribution if $r = 0$. Explain this result in the context of what it means in terms of sampling.

(c) Though we only defined a hypergeometric distrbiution if $N > \max\{r, m\}$, the definition could be extended to $N = \max\{r, m\}$. Use the chart to calculate the variance of a hypergeometric distribution if $N = m$. Explain this result in the context of what it means in terms of sampling without replacement.

Ex. 4.2.11. Prove the following facts about independent random variables.

(a) Use Theorem 4.2.6 and induction to prove that if X_1, X_2, ..., X_n are independent, then

$$Var[X_1 + \cdots + X_n] = Var[X_1] + \cdots + Var[X_n].$$

(b) Suppose X_1, X_2, ..., X_n are i.i.d. Prove that

$$E[X_1 + \cdots + X_n] = n \cdot E[X_1] \quad \text{and} \quad SD[X_1 + \cdots + X_n] = \sqrt{n} \cdot SD[X_1].$$

(c) Suppose X_1, X_2, ..., X_n are mutually independent standardized random variables (not necessarilly identically distributed). Let

$$Y = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}.$$

Prove that Y is a standardized random variable.

Ex. 4.2.12. Let X be a discrete random variable which takes on only non-negative values. Show that if $E[X] = 0$ then $P(X = 0) = 1$.

Ex. 4.2.13. Suppose X is a discrete random variable with finite variance (and thus finite expected value as well) and suppose there are two different numbers $a, b \in \mathbb{R}$ for which $P(X = a)$ and $P(X = b)$ are both positive. Prove that $Var[X] > 0$.

Ex. 4.2.14. Let X be a discrete random variable with finite variance (and thus finite expected value as well).

(a) Prove that $E[X^2] \geq (E[X])^2$.

(b) Suppose there are two different numbers $a, b \in \mathbb{R}$ for which $P(X = a)$ and $P(X = b)$ are both positive. Prove that $E[X^2] > (E[X])^2$.

Ex. 4.2.15. Let $X \sim \text{Binomial}(n, p)$ for $n > 1$ and $0 < p < 1$. Using the steps below, provide an algebraic proof of the fact that $Var[X] = np(1-p)$ without appealing to the fact that such a variable is the sum of Bernoulli trials.

(a) Begin by writing an expression for $E[X^2]$ in summation form.

(b) Use (a) to show that $E[X^2] = np \cdot \sum_{k=0}^{n-1} (k+1)\binom{n-1}{k} p^k (1-p)^{(n-1)-k}$.

(c) Use (b) to explain why $E[X^2] = np \cdot E[Y + 1]$ where $Y \sim \text{Binomial}(n-1, p)$.

(d) Use (c) together with Theorem 4.2.5 to prove that $Var[X] = np(1-p)$.

4.3 STANDARD UNITS

When there is no confusion about what random variable is being discussed, it is usual to use the Greek letter μ in place of $E[X]$ and σ in place of $SD[X]$. When more than one variable is involved the same letters can be used with subscripts (μ_X and σ_X) to indicate which variable is being described.

In statistics one frequently measures results in terms of "standard units" – the number of standard deviations a result is from its expected value. For instance if $\mu = 12$ and $\sigma = 5$, then a result of $X = 20$ would be 1.6 standard units because $20 = \mu + 1.6\sigma$. That is, 20 is 1.6 standard deviations above expected value. Similarly a result of $X = 10$ would be -0.4 standard units because $10 = \mu - 0.4\sigma$.

Since the standard deviation measures a typical distance from average, results that are within one standard deviation from average (between -1 and $+1$ standard units) will tend to be fairly common, while results that are more than two standard deviations from average (less than -2 or greater than $+2$ in standard units) will usually be relatively rare. The likelihoods of some such events will be calculated in the next two examples. Notice that the event $(|X - \mu| \leq k\sigma)$ describes those outcomes of X that are within k standard deviations from average.

EXAMPLE 4.3.1. Let Y represent the sum of two dice. How likely is it that Y will be within one standard deviation of its average? How likely is it that Y will be more than two standard deviations from its average?

We can use our previous calculations that $\mu = 7$ and $\sigma = \sqrt{\frac{35}{6}} \approx 2.42$. The achievable values that are within one standard deviation of average are $5, 6, 7, 8$, and 9. So the probability that the sum of two dice will be within one standard deviation of average is

$$\begin{aligned} P(|Y - \mu| \leq \sigma) &= P(Y \in \{5, 6, 7, 8, 9\}) \\ &= \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} \\ &= \frac{2}{3}. \end{aligned}$$

There is about a 66.7% chance that a pair of dice will fall within one standard deviation of their expected value.

Two standard deviations is $2\sqrt{\frac{35}{6}} \approx 4.83$. Only the results 2 and 12 further than this distance from the expected value, so the probability that X will be more than two standard deviations from average is

$$\begin{aligned} P(|Y - \mu| > 2\sigma) &= P(Y \in \{2, 12\}) \\ &= \frac{2}{36} \approx 0.056. \end{aligned}$$

There is only about a 5.6% chance that a pair of dice will be more than two standard deviations from expected value. ■

EXAMPLE 4.3.2. If $X \sim Uniform\{(1, 2, \ldots, 100)\}$, what is the probability that X will be within one standard deviation of expected value? What is the probability it will be more than two standard deviations from expected value?

Again, based on earlier calculations we know that $\mu = \frac{101}{2} = 50.5$ and that $\sigma = \sqrt{\frac{9999}{12}} \approx 28.9$. Of the possible values that X can achieve, only the numbers $22, 23, \ldots, 79$ fall within one standard deviation of average. So the desired probability is

$$\begin{aligned} P(|X - \mu| \leq \sigma) &= P(X \in \{22, 23, \ldots, 79\}) \\ &= \frac{58}{100}. \end{aligned}$$

There is a 58% chance that this random variable will be within one standard deviation of expected value.

Similarly we can calculate that two standard deviations is $2\sqrt{\frac{9999}{12}} \approx 57.7$. Since $\mu = 50.5$ and since the minimal and maximal values of X are 1 and 100 respectively, results that are more than two or more standard deviations from average cannot happen at all for this random variable. In other words $P(|X - \mu| > 2\sigma) = 0$. ■

4.3.1 *Markov and Chebyshev Inequalities*

The examples of the previous section show that the exact probabilities a random variable will fall within a certain number of standard deviations of its expected value depend on the distribution of the random variable. However, there are some general results that apply to all random variables. To prove these results we will need to investigate some inequalities.

THEOREM 4.3.3. **(Markov's Inequality)** *Let X be a discrete random variable which takes on only non-negative values and suppose that X has a finite expected value. Then for any $c > 0$,*

$$P(X \geq c) \leq \frac{\mu}{c}.$$

Proof - Let T be the range of X, so T is a countable subset of the positive real numbers. By dividing T into those numbers smaller than c and those numbers that are at least as large as c we have

$$\begin{aligned} \mu &= \sum_{t \in T} t \cdot P(X = t) \\ &= \sum_{t \in T, t < c} t \cdot P(X = t) + \sum_{t \in T, t \geq c} t \cdot P(X = t). \end{aligned}$$

The first sum must be non-negative, since we assumed that T consisted of only non-negative numbers, so we only make the quantity smaller by deleting it. Likewise, for each term in the second sum, $t \geq c$ so we only make the quantity smaller by replacing t by c. This gives us

$$\begin{aligned} \mu &= \sum_{t \in T, t < c} t \cdot P(X = t) + \sum_{t \in T, t \geq c} t \cdot P(X = t) \\ &\geq \sum_{t \in T, t \geq c} c \cdot P(X = t) \\ &= c \cdot \sum_{t \in T, t \geq c} P(X = t). \end{aligned}$$

The events $(X = t)$ indexed over all values $t \in T$ for which $t \geq c$ are a countable collection of disjoint sets whose union is $(X \geq c)$. So,

$$\begin{aligned} \mu &\geq c \cdot \sum_{t \in T, t \geq c} P(X = t) \\ &= cP(X \geq c). \end{aligned}$$

Dividing by c gives the desired result.

Markov's theorem can be useful in its own right for producing an upper bound on the liklihood of certain events, but for now we will use it simply as a lemma to prove our next result.

THEOREM 4.3.4. **(Chebychev's Inequality)** *Let X be a discrete random variable with finite, non-zero variance. Then for any $k > 0$,*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof - The event $(|X - \mu| \geq k\sigma)$ is the same as the event $((X - \mu)^2 \geq k^2\sigma^2)$. The random variable $(X - \mu)^2$ is certainly non-negative and its expected value is the variance of X which we have assumed to be finite. Therefore we may apply Markov's inequality to $(X - \mu)^2$ to get

$$\begin{aligned} P(|X - \mu| \geq k\sigma) &= P((X - \mu)^2 \geq k^2\sigma^2) \\ &\leq \frac{E[(X - \mu)^2]}{k^2\sigma^2} \\ &= \frac{Var[X]}{k^2\sigma^2} \\ &= \frac{\sigma^2}{k^2\sigma^2} \\ &= \frac{1}{k^2}. \end{aligned}$$

Though the theorem is true for all $k > 0$, it doesn't give any useful information unless $k > 1$.

EXAMPLE 4.3.5. Let X be a discrete random variable. Find an upper bound on the likelihood that X will be more than two standard deviations from its expected value.

For the question to make sense we need to assume that X has finite variance to begin with. In which case we may apply Chebychev's inequality with $k = 2$ to find that

$$P(|X - \mu| > 2\sigma) \leq P(|X - \mu| \geq 2\sigma) \leq \frac{1}{4}.$$

There is at most a 25% chance that a random variable will be more than two standard deviations from its expected value. ■

EXERCISES

Ex. 4.3.1. Let $X \sim Binomial(10, \frac{1}{2})$.

(a) Calculate μ and σ.

(b) Calculate $P(|X - \mu| \leq \sigma)$, the probability that X will be within one standard deviation of average. Approximate your answer to the nearest tenth of a percent.

(c) Calculate $P(|X - \mu| > 2\sigma)$, the probability that X will be more than two standard deviations from average. Approximate your answer to the nearest tenth of a percent.

Ex. 4.3.2. Let $X \sim Geometric(\frac{1}{4})$.

(a) Calculate μ and σ.

(b) Calculate $P(|X - \mu| \leq \sigma)$, the probability that X will be within one standard deviation of average. Approximate your answer to the nearest tenth of a percent.

(c) Calculate $P(|X-\mu| > 2\sigma)$, the probability that X will be more than two standard deviations from average. Approximate your answer to the nearest tenth of a percent.

Ex. 4.3.3. Let $X \sim Poisson(3)$.

(a) Calculate μ and σ.

(b) Calculate $P(|X-\mu| \leq \sigma)$, the probability that X will be within one standard deviation of average. Approximate your answer to the nearest tenth of a percent.

(c) Calculate $P(|X-\mu| > 2\sigma)$, the probability that X will be more than two standard deviations from average. Approximate your answer to the nearest tenth of a percent.

Ex. 4.3.4. Let $X \sim Binomial(n, \frac{1}{2})$. Determine the smallest value of n for which $P(|X-\mu| > 4\sigma) > 0$. That is, what is the smallest n for which there is a positive probability that X will be more than four standard deviations from average.

Ex. 4.3.5. For $k \geq 1$ there are distributions for which Chebychev's inequality is an equality.

(a) Let X be a random variable with probability mass function $P(X = 1) = P(X = -1) = \frac{1}{2}$. Prove that Chebychev's inequality is an equality for this random variable when $k = 1$.

(b) Let X be a random variable with probability mass function $P(X = 1) = P(X = -1) = p$ and $P(X = 0) = 1 - 2p$. For any given value of $k > 1$, show that it is possible to select a value of p for which Chebychev's inequality is an equality when applied to this random variable.

Ex. 4.3.6. Let X be a discrete random variable with finite expected value μ and finite variance σ^2.

(a) Explain why $P(|X-\mu| > \sigma) = P((X-\mu)^2 > \sigma^2)$.

(b) Let T be the range of the random variable $(X-\mu)^2$.
Explain why $\sum_{t \in T} P((X-\mu)^2 = t) = 1$.

(c) Explain why $Var[X] = \sum_{t \in T} t \cdot P((X-\mu)^2 = t)$.

(d) Prove that if $P(|X-\mu| > \sigma) = 1$, then
$Var[X] > \sum_{t \in T} \sigma^2 \cdot P((X-\mu)^2 = t)$. (Hint: Use (a) to explain why replacing t by σ^2 in the sum from (c) will only make the quantity smaller).

(e) Use parts (b) and (d) to derive a contradiction. Note that this proves that the assumption that was made in part (d), namely that $P(|X-\mu| > \sigma) = 1$, cannot be true for any discrete random variable where μ and σ are finite quantities. In other words, no random variable can produce only values that are more than one standard deviation from average.

Ex. 4.3.7. Let X be a discrete random variable with finite expected value and finite variance.

(a) Prove $P(|X-\mu| \geq \sigma) = 1 \iff P(|X-\mu| = \sigma) = 1$. (A random variable that assumes only values one or more standard deviations from average must only produce values that are exactly one standard deviation from average).

(b) Prove that if $P(|X-\mu| > \sigma) > 0$ then $P(|X-\mu| < \sigma) > 0$. (If a random variable is able to produce values more one standard deviation from average, it must also be able to produce values that are less than one standard deviation from average).

4.4 CONDITIONAL EXPECTATION AND CONDITIONAL VARIANCE

In previous chapters we saw that information that a particular event had occurred could substantially change the probability associated with another event. That realization led us to the notion of conditional probability. It is also reasonable to ask how such information might affect the expected value or variance of a random variable.

DEFINITION 4.4.1. *Let $X : S \to T$ be a discrete random variable and let $A \subset S$ be an event for which $P(A) > 0$. The "conditional expected value" is defined from conditional probabilities in the same way the (ordinary) expected value is defined from (ordinary) probabilities. Likewise the "conditional variance" is described in terms of the conditional expected value in the same way the (ordinary) variance is described in terms of the (ordinary) expected value. Specificially, the "conditional expected value" of X given A is*

$$E[X|A] = \sum_{t \in T} t \cdot P(X = t|A),$$

and the "conditional variance" of X given A is

$$Var[X|A] = E[(X - E[X|A])^2|A].$$

EXAMPLE 4.4.2. A die is rolled. What are the expected value and variance of the result given that the roll was even?

Let X be the die roll. Then $X \sim \text{Uniform}(\{1, 2, 3, 4, 5, 6\})$, but conditioned on the event A that the roll was even, this changes so that

$$P(X = 1|A) = P(X = 3|A) = P(X = 5|A) = 0 \quad \text{while}$$

$$P(X = 2|A) = P(X = 4|A) = P(X = 6|A) = \frac{1}{3}.$$

Therefore,

$$E[X|A] = 2(\frac{1}{3}) + 4(\frac{1}{3}) + 6(\frac{1}{3}) = 4.$$

Note that the (unconditioned) expected value of a die roll is $E[X] = 3.5$, so the knowledge of event A slightly increases the expected value of the die roll.

The conditional variance is

$$Var[X|A] = (2-4)^2(\frac{1}{3}) + (4-4)^2(\frac{1}{3}) + (6-4)^2(\frac{1}{3}) = \frac{8}{3}.$$

This result is slightly less than $\frac{35}{12}$, the (unconditional) variance of a die roll. This means that knowledge of event A sligthly decreased the typical spread of the die roll results. ■

In many cases the event A on which an expected value is conditioned will be described in terms of another random variable. For instance $E[X|Y = y]$ is the conditional expectation of X given that variable Y has taken on the value y.

EXAMPLE 4.4.3. Cards are drawn from an ordinary deck of 52, one at a time, randomly and with replacement. Let X and Y denote the number of draws until the first ace and first king are drawn,

respectively. We are interested in say, $E[X|Y=3]$. When $Y=3$ an ace was seen of draw 3, but not on draws 1 or 2. Hence

$$P(\text{king on draw } n|Y=3) = \begin{cases} \frac{4}{48} & \text{if } n=1 \text{ or } 2 \\ 0 & \text{if } n=3 \\ \frac{4}{52} & \text{if } n>3 \end{cases}$$

so that

$$P(X=n|Y=5) = \begin{cases} \left(\frac{44}{48}\right)^{n-1}\frac{4}{48} & \text{if } n=1 \text{ or } 2 \\ 0 & \text{if } n=3 \\ \left(\frac{44}{48}\right)^{2}\left(\frac{48}{52}\right)^{n-4}\frac{4}{52} & \text{if } n>3 \end{cases}$$

For example, when $n>3$, in order to have $X=n$ a non-king must have been seen on draws 1 and 2 (each with probability $\frac{44}{48}$), a non-king must have resulted on draw 3 (which is automatic, since an ace was drawn), a non-king must have been seen on each of draws 4 through $n-1$ (each with probability $\frac{48}{52}$), and finally a king was produced on draw n (with probability $\frac{4}{52}$). Hence,

$$\begin{aligned} E[X|Y=3] &= \sum_{n=1}^{2} n\left(\frac{44}{48}\right)^{n-1}\frac{4}{48} + \sum_{n=4}^{\infty} n\left(\frac{44}{48}\right)^{2}\left(\frac{48}{52}\right)^{n-4}\frac{4}{52} \\ &= \sum_{n=1}^{2} n\left(\frac{44}{48}\right)^{n-1}\frac{4}{48} + \sum_{m=0}^{\infty} (m+4)\left(\frac{44}{48}\right)^{2}\left(\frac{48}{52}\right)^{m}\frac{4}{52}. \end{aligned}$$

But

$$\begin{aligned} \sum_{m=0}^{\infty}(m+4)r^m &= \sum_{m=0}^{\infty}\left(3r^m + \frac{d}{dr}r^{m+1}\right) \\ &= \frac{3}{1-r} + \frac{d}{dr}\left(\frac{r}{1-r}\right) \\ &= \frac{3}{1-r} + \frac{1}{(1-r)^2}, \end{aligned}$$

so

$$\begin{aligned} E[X|Y=3] &= \frac{4}{48} + 2\left(\frac{44}{48}\right)\left(\frac{4}{48}\right) + \left(\frac{44}{48}\right)^{2}\left(\frac{4}{52}\right)\left(\frac{3}{1-(48/52)} + \frac{1}{(1-(48/52))^2}\right) \\ &= \frac{4}{48} + 2\left(\frac{44}{48}\right)\left(\frac{4}{48}\right) + \left(\frac{44}{48}\right)^{2}\left(\frac{4}{52}\right)\left(\frac{3\times 52}{4} + \frac{52^2}{4^2}\right) \\ &= \frac{1}{12} + 2\left(\frac{11}{12}\right)\left(\frac{1}{12}\right) + 3\left(\frac{11}{12}\right)^{2} + \frac{52}{4}\left(\frac{11}{12}\right)^{2} \\ &= \frac{985}{72} \approx 13.68. \end{aligned}$$

Given that the first ace appeared on draw 3, it takes an average of between 13 and 14 draws until the first king appears. Compare this to the unconditional $E[X]$. Since $X \sim \text{Geometric}(\frac{4}{52})$ we know $E[X] = \frac{52}{4} = 13$. In other words, on average it takes 13 draws to observe the first king. But given that the first ace appeared on draw three, we should expect to need about 0.68 draws more (on average) to see the first king. ■

Recall how Theorem 1.3.2 described a way in which a non-conditional probability could be calculated in terms of conditional probabilities. There is an analogous theorem for expected value.

THEOREM 4.4.4. *Let $X : S \to T$ be a discrete random variable and let $\{B_i : i \geq 1\}$ be a disjoint collection of events for which $P(B_i) > 0$ for all i and such that $\bigcup_{i=1}^{\infty} B_i = S$. Suppose $P(B_i)$ and $E[X|B_i]$ are known. Then $E[X]$ may be computed as*

$$E[X] = \sum_{i=1}^{\infty} E[X|B_i]P(B_i).$$

Proof - Using Theorem 1.3.2 and the definition of conditional expectation,

$$\begin{aligned}\sum_{i=1}^{\infty} E[X|B_i]P(B_i) &= \sum_{i=1}^{\infty}\sum_{t\in T} t \cdot P(X = t|B_i)P(B_i) \\ &= \sum_{t\in T}\sum_{i=1}^{\infty} t \cdot P(X = t|B_i)P(B_i) \\ &= \sum_{t\in T} t \cdot P(X = t) = E[X].\end{aligned}$$

■

EXAMPLE 4.4.5. A venture capitalist estimates that regardless of whether the economy strengthens, weakens, or remains the same in the next fiscal quarter, a particular investment could either gain or lose money. However, he figures that if the economy strengthens, the investment should, on average, earn 3 million dollars. If the economy remains the same, he figures the expected gain on the investment will be 1 million dollars, while if the economy weakens, the investment will, on average, lose 1 million dollars. He also trusts economic forcasts which predict a 50% chance of a weaker economy, a 40% chance of a stagnant economy, and a 10% chance of a stronger economy. What should he calculate is the expected return on the investment?

Let X be the return on investment and let A, B, and C represent the events that the economy will be stronger, the same, and weaker in the next quarter, respectively. Then the estimates on return give the following information in millions:

$$E[X|A] = 3; \;\; E[X|B] = 1; \text{ and } E[X|C] = -1.$$

Therefore,

$$\begin{aligned}E[X] &= E[X|A]P(A) + E[X|B]P(B) + E[X|C]P(C) \\ &= 3(0.1) + 1(0.4) + (-1)(0.5) = 0.2\end{aligned}$$

The expected return on investment is \$200,000. ■

When the conditioning event is described in terms of outcomes of a random variable, Theorem 4.4.4 can be written in another useful way.

THEOREM 4.4.6. *Let X and Y be two discrete random variables on a sample space S with $Y : S \to T$. Let $g : T \to \mathbb{R}$ be defined as $g(y) = E[X|Y = y]$. Then*

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

It is common to use $E[X|Y]$ to denote $g(Y)$ after which the theorem may be expressed as $E[E[X|Y]] = E[X]$. This can be slightly confusing notation, but one must keep in mind that the exterior expected value in the expression $E[E[X|Y]]$ refers to the averge of $E[X|Y]$ viewed as a function of Y.

Proof - As y ranges over T, the events $(Y = y)$ are disjoint and cover all of S. Therefore, by Theorem 4.4.4,

$$\begin{aligned} E[g(Y)] &= \sum_{y\in T} g(y)P(Y=y) \\ &= \sum_{y\in T} E[X|Y=y]P(Y=y) \\ &= E[X]. \end{aligned}$$

■

EXAMPLE 4.4.7. Let $Y \sim \text{Uniform}(\{1, 2, \ldots, n\})$ and let X be the number of heads on Y flips of a coin. What is the expected value of X?

Without Theorem 4.4.6 this problem would require computing many complicated probabilities. However, it is made much simpler by noting that the distribution of X is given conditionally by $(X|Y = j) \sim \text{Binomial}(j, \frac{1}{2})$. Therefore we know $E[X|Y = j] = \frac{j}{2}$. Using the notation above, this may be written as $E[X|Y] = \frac{Y}{2}$ after which

$$E[X] = E[E[X|Y]] = E[\frac{Y}{2}] = \frac{1}{2}\frac{n+1}{2} = \frac{n+1}{4}.$$

■

Though it requires a somewhat more complicated formula, the variance of a random variable can be computed from conditional information.

THEOREM 4.4.8. *Let $X : S \to T$ be a discrete random variable and let $\{B_i : i \geq 1\}$ be a disjoint collection of events for which $P(B_i) > 0$ for all i and such that $\bigcup_{i=1}^{\infty} B_i = S$. Suppose $E[X|B_i]$ and $Var[X|B_i]$ are known. Then $Var[X]$ may be computed as*

$$Var[X] = \left(\sum_{i=1}^{\infty}(Var[X|B_i] + (E[X|B_i])^2)P(B_i)\right) - (E[X])^2.$$

Proof- First note that $Var[X|B_i] = E[X^2|B_i] - (E[X|B_i])^2$, and so

$$Var[X|B_i] + (E[X|B_i])^2 = E[X^2|B_i].$$

Therefore,

$$\sum_{i=1}^{\infty}(Var[X|B_i] + (E[X|B_i])^2)P(B_i) = \sum_{i=1}^{\infty} E[X^2|B_i]P(B_i),$$

but the right hand side of this equation is $E[X^2]$ from Theorem 4.4.4. The fact that $Var[X] = E[X^2] - (E[X])^2$ completes the proof of the theorem. ■

As with expected value, this formula may be rewritten in a different form if the conditioning events describe the outcomes of a random variable.

THEOREM 4.4.9. *Let X and $Y : S \to T$ be two discrete random variables on a sample space S. As in Theorem 4.4.6 let $g(y) = E[X|Y = y]$. Let $h(y) = Var[X|Y = y]$. Denoting $g(Y)$ by $E[X|Y]$ and denoting $h(Y)$ by $Var[X|Y]$, then*

$$Var[X] = E[Var[X|Y]] + Var[E[X|Y]].$$

Proof - First consdier the following three facts:

(1) $\sum_{t\in T} Var[X|Y=t]P(Y=t) = E[Var[X|Y]]$;

(2) $\sum_{t\in T} (E[X|Y=t])^2 P(Y=t) = E[(E[X|Y])^2]$; and

(3) $Var[E[X|Y]] = E[(E[X|Y])^2] - (E[E[X|Y]])^2 = E[(E[X|Y])^2] - (E[X])^2$.

Then from Theorem 4.4.8,

$$\begin{aligned} Var[\quad X \quad] &= \sum_{t\in T} (Var[X|Y=t] + (E[X|Y=t])^2)P(Y=t) - (E[X])^2 \\ &= \sum_{t\in T} Var[X|Y=t]P(Y=t) + \sum_{t\in T} (E[X|Y=t])^2 P(Y=t) - (E[X])^2 \\ &= E[Var[X|Y]] + E[(E[X|Y])^2] - (E[X])^2 \\ &= E[Var[X|Y]] + Var[E[X|Y]]. \end{aligned}$$

■

EXAMPLE 4.4.10. The number of eggs N found in nests of a certain species of turtles has a Poisson distribution with mean λ. Each egg has probability p of being viable and this event is independent from egg to egg. Find the mean and variance of the number of viable eggs per nest.

Let N be the total number of eggs in a nest and X the number of viable ones. Then if $N = n$, X has a binomial distribution with number of trials n and probability p of success for each trial. Thus, if $N = n$, X has mean np and variance $np(1-p)$. That is,

$$E[X|N=n] = np; \qquad Var[X|N=n] = np(1-p)$$

or

$$E[X|N] = pN; \qquad Var[X|N] = p(1-p)N.$$

Hence

$$E[X] = E[E[X|N]] = E[pN] = pE[N] = p\lambda$$

and

$$\begin{aligned} Var[X] &= E[Var[X|N]] + Var[E[X|N]] \\ &= E[p(1-p)N] + Var[pN] = p(1-p)E[N] + p^2 Var[N]. \end{aligned}$$

Since N is Poisson we know that $E[N] = Var[N] = \lambda$, so that

$$E[X] = p\lambda \qquad \text{and} \qquad Var[X] = p(1-p)\lambda + p^2\lambda = p\lambda.$$

■

EXERCISES

Ex. 4.4.1. Let $X \sim$ Geometric(p) and let A be event $(X \leq 3)$. Calculate $E[X|A]$ and $Var[X|A]$.

Ex. 4.4.2. Calculate the variance of the quantity X from Example 4.4.7.

Ex. 4.4.3. Return to Example 4.4.5. Suppose that, in addition to the estimates on average return, the investor had estimates on the standard deviations. If the economy strengthens or weakens, the estimated standard deviation is 3 million dollars, but if the economy stays the same, the estimated standard deviation is 2 million dollars. So, in millions of dollars,

$$SD[X|A] = 3; \;\; SD[X|B] = 2; \text{ and } SD[X|C] = 3.$$

Use this information, together with the conditional expectations from Example 4.4.5 to calculate $Var[X]$.

Ex. 4.4.4. A standard light bulb has an average lifetime of four years with a standard deviation of one year. A Super D-Lux lightbulb has an average lifetime of eight years with a standard devaition of three years. A box contains many bulbs – 90% of which are standard bulbs and 10% of which are Super D-Lux bulbs. A bulb is selected at random from the box. What are the average and standard deviation of the lifetime of the selected bulb?

Ex. 4.4.5. Let X and Y be described by the joint distribution

	$X=-1$	$X=0$	$X=1$
$Y=-1$	1/15	2/15	2/15
$Y=0$	2/15	1/15	2/15
$Y=1$	2/15	2/15	1/15

and answer the following questions.

(a) Calculate $E[X|Y=-1]$.

(b) Calculate $Var[X|Y=-1]$.

(c) Describe the distribution of $E[X|Y]$.

(d) Describe the distribution of $Var[X|Y]$.

Ex. 4.4.6. Let X and Y be discrete random variables. Let x be in the range of X and let y be in the range of Y.

(a) Suppose X and Y are independent. Show that $E[X|Y=y] = E[X]$ (and so $E[X|Y] = E[X]$).

(b) Show that $E[X|X=x] = x$ (and so $E[X|X] = X$). (From results in this section we know $E[X|Y]$ is always a random variable with expected value equal to $E[X]$. The results above in some sense show two extremes. When X and Y are independent, $E[X|Y]$ is a constant random variable $E[X]$. When X and Y are equal, $E[X|X]$ is just X itself).

Ex. 4.4.7. Let $X \sim$ Uniform $\{1, 2, \ldots, n\}$ be independent of $Y \sim$ Uniform $\{1, 2, \ldots, n\}$. Let $Z = \max(X, Y)$ and $W = \min(X, Y)$.

(a) Find the joint distribution of (Z, W).

(b) Fine $E[Z \mid W]$.

4.5 COVARIANCE AND CORRELATION

When faced with two different random variables, we are frequently interested in how the two different quantities relate to each other. Often the purpose of this is to predict something about one variable knowing information about the other. For instance, if rainfall amounts in July affect the quantity of corn harvested in August, then a farmer, or anyone else keenly interested in the supply and demand of the agriculture industry, would like to be able to use the July information to help make predictions about August costs.

4.5.1 *Covariance*

Just as we developed the concepts of expected value and standard deviation to summarize a single random variable, we would like to develop a number that describes something about how two different random variables X and Y relate to each other.

DEFINITION 4.5.1. **(Covariance of X and Y)** *Let X and Y be two discrete random variables on a sample space S. Then the "covariance of X and Y" is defined as*

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]. \tag{4.5.1}$$

Since it is defined in terms of an expected value, there is the possibility that the covariance may be infinite or not defined at all because the sum describing the expectation is divergent.

Notice that if X is larger than its average at the same time that Y is larger than its average (or if X is smaller than its average at the same time Y is smaller than its average) then $(X - E[X])(Y - E[Y])$ will contribute a positive result to the expected value describing the covariance. Conversely, if X is smaller than $E[X]$ while Y is larger than $E[Y]$ or vica versa, a negative result will be contributed toward the covariance. This means that when two variables tend to be both above average or both below average simultaneously, the covariance will typically be positive (and the variables are said to be positively correlated), but when one variable tends to be above average when the other is below average, the covariance will typically be negative (and the variables are said to be negatively correlated). When $Cov[X,Y] = 0$ the variables X and Y are said to be "uncorrelated".

For example, suppose X and Y are the height and weight, respectively, of an individual randomly selected from a large population. We might expect that $Cov[X,Y] > 0$ since people who are taller than average also tend to be heavier than average and people who are shorter than average tend to be lighter. Conversely suppose X and Y represent elevation and air density at a randomly selected point on Earth. We might expect $Cov[X,Y] < 0$ since locations at a higher elevation tend to have thinner air.

EXAMPLE 4.5.2. Consider a pair of random variables X and Y with joint distribution

	$X=-1$	$X=0$	$X=1$
$Y=-1$	1/15	2/15	2/15
$Y=0$	2/15	1/15	2/15
$Y=1$	2/15	2/15	1/15

By a routine calculation of the marginal distributions it can be shown that $X, Y \sim \text{Uniform}(\{-1,0,1\})$ and therefore that $E[X] = E[Y] = 0$. However, it is clear from the joint distribution that when

$X = -1$, then Y is more likely to be above average than below, while when $X = 1$, then Y is more likely to be below average than above. This suggests the two random variables should have a negative correlation. In fact, we can calculate

$$E[XY] = (-1)(\frac{4}{15}) + 0(\frac{9}{15}) + 1(\frac{2}{15}) = -\frac{2}{15},$$

and therefore $Cov[X, Y] = E[XY] - E[X]E[Y] = -\frac{2}{15}$. ■

As its name suggests, the covariance is closely related to the variance.

THEOREM 4.5.3. *Let X be a discrete random variable. Then*

$$Cov[X, X] = Var[X].$$

Proof - $Cov[X, X] = E[(X - E[X])(X - E[X])) = E[(X - E[X])^2) = Var[X]$. ■

With Theorem 4.2.5 it was shown that $Var[X] = E[X^2] - (E[X])^2$, which provided an alternate formula for the variance. There is an analogous alternate formula for the covariance.

THEOREM 4.5.4. *Let X and Y be discrete random variables with finite mean for which $E[XY]$ is also finite. Then*

$$Cov[X, Y] = E[XY] - E[X]E[Y].$$

Proof - Using the linearity properties of expected value,

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - E[X]Y + E[X]E[Y]] \\ &= E[XY] - E[XE[Y]] - E[E[X]Y] + E[E[X]E[Y]] \\ &= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

■

As with the expected value, the covariance is a linear quantity. It is also related to the concept of independence.

THEOREM 4.5.5. *Let X, Y, and Z be discrete random variables, and let $a, b \in \mathbb{R}$. Then,*

(a) $Cov[X, Y] = Cov[Y, X]$;

(b) $Cov[X, aY + bZ] = a \cdot Cov[X, Y] + b \cdot Cov[X, Z]$;

(c) $Cov[aX + bY, Z] = a \cdot Cov[X, Z] + b \cdot Cov[Y, Z]$; *and*

(d) If X and Y are independent with a finite covariance, then $Cov[X, Y] = 0$.

Proof of (1) - This follows immediately from the definition.

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[(Y - E[Y])(X - E[X])] = Cov[Y, X]. \end{aligned}$$

Therefore, reversing the roles of X and Y does not change the correlation.

Proof of (2) - This follows from linearity properties of expected value. Using Theorem 4.5.4

$$\begin{aligned} Cov[X, aY + bZ] &= E[X(aY + bZ)] - E[X]E[aY + bZ] \\ &= a \cdot E[XY] + b \cdot E[XZ] - a \cdot E[X]E[Y] - b \cdot E[X]E[Z] \\ &= a \cdot (E[XY] - E[X]E[Y]) + b \cdot (E[XZ] - E[X]E[Z]) \\ &= a \cdot Cov[X, Y] + b \cdot Cov[X, Z] \end{aligned}$$

Proof of (3) - This proof is essentially the same as that of (2) and is left as an exercise.
Poof of (4) - We have previously seen that if X and Y are independent, then $E[XY] = E[X]E[Y]$. Using Theorem 4.5.4 it follows that

$$Cov[X, Y] = E[XY] - E[X]E[Y] = 0.$$

Though independence of X and Y guarantees that they are uncorrelated, the converse is not true. It is possible that $Cov[X, Y] = 0$ and yet that X and Y are dependent, as the next example shows.

EXAMPLE 4.5.6. Let X, Y be two discrete random variables taking values $\{-1, 1\}$. Suppose their joint distribution $P(X = x, Y = y)$ is given by the table

	x=-1	x=1
y=-1	0.3	0.2
y=1	0.3	0.2

By summing the columns and rows respectively,

$$P(X = 1) = 0.4 \text{ and } P(X = -1) = 0.6, \text{ while}$$

$$P(Y = 1) = 0.5 \text{ and } P(Y = -1) = 0.5.$$

Moreover,

$$\begin{aligned} E[XY] &= (1)(-1)P(X = 1, Y = -1) + (-1)(1)P(X = -1, Y = 1) \\ &\quad +(1)(1)P(X = 1, Y = 1) + (-1)(-1)P(X = -1, Y = -1) \\ &= -0.3 - 0.2 + 0.2 + 0.3 = 0, \\ E[X] &= (1)0.4 + (-1)0.6 = -0.2, \\ E[Y] &= (1)0.5 + (-1)0.5 = 0, \end{aligned}$$

implying that $\text{Cov}[X, Y] = 0$. As

$$P(X = 1, Y = 1) = 0.2 \neq 0.1 = P(X = 1)P(Y = 1),$$

they are not independent random variables. ■

4.5.2 *Correlation*

The possible size of $Cov[X, Y]$ has upper and lower bounds based on the standard deviations of the two variables.

THEOREM 4.5.7. *Let X and Y be two discrete random variables both with finite variance. Then*

$$-\sigma_X\sigma_Y \leq Cov[X,Y] \leq \sigma_X\sigma_Y,$$

and therefore $-1 \leq \frac{Cov[X,Y]}{\sigma_X\sigma_Y} \leq 1$.

Proof - Standardize both variables and consider the expected value of their sum squared. Since this is the expected value of a non-negative quantity,

$$\begin{aligned} 0 &\leq E[(\frac{X-\mu_X}{\sigma_X} + \frac{Y-\mu_Y}{\sigma_Y})^2] \\ &= E[\frac{(X-\mu_X)^2}{\sigma_X^2} + 2\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(Y-\mu_Y)^2}{\sigma_Y^2}] \\ &= \frac{E[(X-\mu_X)^2]}{\sigma_X^2} + \frac{2E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} + \frac{E[(Y-\mu_Y)^2]}{\sigma_Y^2} \\ &= 1 + 2\frac{Cov[X,Y]}{\sigma_X\sigma_Y} + 1. \end{aligned}$$

Sovling the inequality for the covariance yields

$$Cov[X,Y] \geq -\sigma_X\sigma_Y.$$

A similar computation (see Exercises) for the expected value of the squared difference of the standardized variables shows

$$Cov[X,Y] \leq \sigma_X\sigma_Y.$$

Putting both inequalities together proves the theorem. ■

DEFINITION 4.5.8. *The quantity $\frac{Cov[X,Y]}{\sigma_X\sigma_Y}$ from Theorem 4.5.7 is known as the"correlation" of X and Y and is often denoted as $\rho[X,Y]$. Thinking in terms of dimensional analysis, both the numerator and denominator include the units of X and the units of Y. The correlation, therefore, has no units associated with it. It is thus a dimensionless rescaling of the covariance and is frequently used as an absolute measure of trends between the two variables.*

EXERCISES

Ex. 4.5.1. Consider the experiment of flipping two coins. Let X be the number of heads among the coins and let Y be the number of tails among the coins.

(a) Should you expect X and Y to be posivitely correlated, negatively correlated, or uncorrelated? Why?

(b) Calculate $Cov[X,Y]$ to confirm your answer to (a).

Ex. 4.5.2. Let $X \sim$ Uniform$(\{0,1,2\})$ and let Y be the number of heads in X flips of a coin.

(a) Should you expect X and Y to be positively correlated, negatively correlated, or uncorrelated? Why?

(b) Calculate $Cov[X, Y]$ to confirm your answer to (a).

Ex. 4.5.3. Prove part (3) of Theorem 4.5.5.

Ex. 4.5.4. Prove the missing inequality from the proof of Theorem 4.5.7. Specifically, use the inequality

$$0 \leq E[(\frac{X - \mu_X}{\sigma_X} - \frac{Y - \mu_Y}{\sigma_Y})^2]$$

to prove that $Cov[X, Y] \leq \sigma_X \sigma_Y$.

Ex. 4.5.5. Prove that the inequality of Theorem 4.5.7 is an equality if and only if there are $a, b \in \mathbb{R}$ with $a \neq 0$ for which $P(Y = aX + b) = 1$. (Put another way, the correlation of X and Y is ± 1 exactly when Y can be expressed as a non-trivial linear function of X).

Ex. 4.5.6. In previous sections it was shown that if X and Y are independent, then $Var[X + Y] = Var[X] + Var[Y]$. If X and Y are dependent, the result is typically not true, but the covariance provides a way relate the variances of X and Y to the variance of their sum.

(a) Show that for any discrete random variables X and Y,

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y].$$

(b) Use (a) to conclude that when X and Y are positively correlated, then $Var[X + Y] > Var[X] + Var[Y]$, while when X and Y are negatively correlated, $Var[X + Y] < Var[X] + Var[Y]$.

(c) Suppose X_i $1 \leq i \leq n$ are discrete random variables with finite variance and covariances. Use induction and (a) to conclude that

$$Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i] + 2 \sum_{1 \leq i < j \leq n} Cov[X_i, X_j].$$

4.6 EXCHANGEABLE RANDOM VARIABLES

We conclude this section with a discussion on exchangeable random variables. In brief we say that a collection of random variables is exchangeable if the joint probability mass function of $(X_1, X_2, \ldots, X_n)$ is a symmetric function. In other words, the distribution of $(X_1, X_2, \ldots, X_n)$ is independent of the order in which the X_i's appear. In particular any collection of mutually independent random variables is exchangeable.

DEFINITION 4.6.1. *Let $n \geq 2$ and $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ be a bijection. We say that a subset T of $\mathbb{R}^n$ is symmetric if*

$$(x_1, x_2, \ldots, x_n) \in T \iff (x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}) \in T$$

for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. For any symmetric set T, a function $f : T \to \mathbb{R}$ is symmetric if

$$f(x_1, x_2, \ldots, x_n) = f(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)})$$

for all $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

A bijection $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ is often referred to as a permutation of $\{1, 2, \ldots, n\}$. When $n = 2$ the function f would be symmetric if $f(x, y) = f(y, x)$ for all $x, y \in \mathbb{R}$.

DEFINITION 4.6.2. *Let $n \geq 1$ and $X_1, X_2, \ldots, X_n$ be discrete random variables. We say that $X_1, X_2, \ldots, X_n$ is a collection of exchangeable random variables if the joint probability mass function given by*

$$f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, \ldots X_n = x_n)$$

is a symmetric function.

In particular, $X_1, X_2, \ldots, X_n$ are exchangeable then for any one of the possible $n!$ permutations, σ, of $\{1, 2, \ldots, n\}$, $X_1, X_2, \ldots, X_n$ and $X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)}$ have the same distribution.

EXAMPLE 4.6.3. Suppose we have an urn of m distinct objects labelled $\{1, 2, \ldots, m\}$. Objects are drawn at random from the urn without replacements till the urn is empty. Let X_i be the label of the i-th object that is drawn. Then $X_1, X_2, \ldots, X_m$ is a particular ordering of the objects in the urn. Since each ordering is equally likely and there are $m!$ possible orderings we have that the joint probability mass function

$$f(x_1, x_2, \ldots, x_m) = P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = \frac{1}{m!},$$

whenever $x_i \in \{1, 2, \ldots m\}$ with $x_i \neq x_j$. As the function is a constant function on the symmetric set $\{1, 2, \ldots, m\}$, it is clearly symmetric. So the random variables $X_1, X_2, \ldots, X_m$ are exchangeable. ■

THEOREM 4.6.4. *Let $X_1, X_2, \ldots, X_n$ be a collection of exchangeable random variables on a sample space S. Then for any $i, j \in \{1, 2, \ldots, n\}$, X_i and X_j have the same marginal distribution.*

Proof - The random variables $(X_1, X_2, \ldots, X_n)$ are exchangeable. Then we have for any permutation σ and $x_i \in \text{Range}(X_i)$

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{\sigma(1)} = x_1, X_{\sigma(2)} = x_2, \ldots, X_{\sigma(n)} = x_n).$$

As this is true for all permutations σ all the random variables must have same range. Otherwise if any two of them differ the we could get a contradiction by choosing an appropriate permutation.

Let T denote the common range. Let $i \in \{2, \ldots, n\}, a, b \in T$. Let

$$A = \{x_j \in T : 1 \leq j \neq 1, i \leq n\}$$

By using the exchangeable property with the permutation σ that is given by $\sigma(i) = 1, \sigma(1) = i$ and $\sigma(j) = j$ for all $j \neq 1, i$. We have that for any $x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \in A$

$$\begin{aligned} &P(X_1 = a, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}, X_i = b, X_{i+1} = x_{i+1}, \ldots X_n = x_n) \\ &= P(X_1 = b, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}, X_i = a, X_{i+1} = x_{i+1}, \ldots X_n = x_n). \end{aligned}$$

Therefore,

$$\begin{aligned}
P(X_1 = a) &= P(\bigcup_{b\in T} X_1 = a, X_i = b) \\
&= \sum_{b\in T} P(X_1 = a, X_i = b) \\
&= \sum_{b\in T} P(\bigcup_{x_j\in A} X_1 = a, X_2 = x_2, \dots X_{i-1} = x_{i-1}, X_i = b, X_{i+1} = x_{i+1}, \dots X_n = x_n) \\
&= \sum_{b\in T}\sum_{x_j\in A} P(X_1 = a, X_2 = x_2, \dots, X_i = b, \dots X_n = x_n) \\
&= \sum_{b\in T}\sum_{x_j\in A} P(X_1 = b, X_2 = x_2, \dots, X_i = a, \dots X_n = x_n) \\
&= \sum_{b\in T} P(\bigcup_{x_j\in A} X_1 = b, X_2 = x_2, \dots X_{i-1} = x_{i-1}, X_i = a, X_{i+1} = x_{i+1}, \dots X_n = x_n) \\
&= \sum_{b\in T} P(X_1 = b, X_i = a) \\
&= P(\bigcup_{b\in T} X_1 = b, X_i = a) \\
&= P(X_i = a)
\end{aligned}$$

So the distribution of X_i is the same as the distribution of X_1 and hence all of them have the same distribution. ■

EXAMPLE 4.6.5. **(Sampling without Replacement)** An urn contains b black balls and r red balls. A ball is drawn at random and its colour noted. This procedure is repeated n times. Assume that $n \le b+r$. Let $\max 0, n-r \le k \le \min(n, b)$. In this example we examine the random variables X_i given by

$$X_i = \begin{cases} 1 & \text{if } i\text{-th ball drawn is black} \\ 0 & \text{otherwise} \end{cases}$$

We have already seen that (See Theorem 2.3.2 and Example 2.3.1)

$$P(k \text{ black balls are drawn in } n \text{ draws}) = \binom{n}{k} \frac{\prod_{i=0}^{k-1}(b-i)\prod_{i=0}^{m-k-1}(r-i)}{\prod_{i=0}^{m-1}(r+b-i)}.$$

Using the same proof we see that the joint probability mass function of $(X_1, X_2, \dots, X_n)$ is given by

$$f(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2 \dots X_n = x_n) = \frac{\prod_{i=0}^{\sum_{i=1}^n x_i - 1}(b-i)\prod_{i=0}^{\sum_{i=1}^n x_i - k - 1}(r-i)}{\prod_{i=0}^{\sum_{i=1}^n x_i - 1}(r+b-i)},$$

where $x_i \in \{0, 1\}$. It is clear from the right hand side of the above that the function f depends only on the $\sum_{i=1}^n x_i$. Hence any permutation of the x_i's will not change the value of f. So f is a symmetric function and the random variables are exchangeable. Therefore, by Theorem 4.6.4 we know that for any $1 \le i \le n$,

$$P(X_i = 1) = P(X_1 = 1) = \frac{b}{b+r}.$$

So we can conclude that they are all identically distributed as Bernoulli $(\frac{b}{b+r})$ and the probability of choosing a black ball in the i-th draw is $\frac{b}{b+r}$ (See Exercise 4.6.4 for a similar result). Further for any i, j

$$\begin{aligned} Cov[X_i, X_j] &= E[X_i X_j] - E[X_i]E[X_j] \\ &= E[X_1 X_2] - \left(\frac{b}{b+r}\right)^2 \\ &= \frac{b(b-1)}{(b+r)(b+r-1)} - \left(\frac{b}{b+r}\right)^2 \\ &= \frac{-br}{(b+r)^2(b+r-1)} \end{aligned}$$

Finally, we observe that $Y = \sum_{i=1}^{n} X_i$ is a Hypergeometric $(b+r, b, m)$. Exchangeability thus provides another alternative way to compute the mean and variance of Y. Using the linearity of expectation provided by Theorem 4.1.7, we have

$$E[Y] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = n\frac{b}{b+r}.$$

and by Exercise 4.5.6,

$$\begin{aligned} Var[Y] &= Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i] + 2 \sum_{1 \le i < j \le n}^{n} Cov[X_i, X_j] \\ &= nVar[X_1] + n(n-1)Cov[X_1, X_2] \\ &= n\frac{br}{(b+r)^2} + n(n-1)(\frac{-br}{(b+r)^2(b+r-1)}) \\ &= n\frac{br}{(b+r)^2}\frac{b+r-n}{b+r-1}. \end{aligned}$$

■

EXERCISES

Ex. 4.6.1. Suppose $X_1, X_2, \ldots, X_n$ are exchangeable random variables. For any $2 \le m < n$, show that $X_1, X_2, \ldots, X_m$ are also a collection of exchangeable random variables.

Ex. 4.6.2. Suppose $X_1, X_2, \ldots, X_n$ are exchangeable random variables. Let T denote their common range. Suppose $b : T \to \mathbb{R}$. Show that $b(X_1), b(X_2), \ldots, b(X_n)$ is also a collection of exchangeable random variables.

Ex. 4.6.3. Suppose n cards are drawn from a standard pack of 52 cards without replacement (so we will assume $n \le 52$). For $1 \le i \le n$, let X_i be random variables given by

$$X_i = \begin{cases} 1 & \text{if } i\text{-th card drawn is black in colour} \\ 0 & \text{otherwise} \end{cases}$$

(a) Suppose $n = 52$. Using Example 4.6.3 and the Exercise 4.6.2 show that $(X_1, X_2, X_3, \ldots X_n)$ are exchangeable.

(b) Show that $(X_1, X_2, X_3, \ldots, X_n)$ are exchangeable for any $2 \le n \le 52$. *Hint: If $n < 52$ extend the sample to exhause the deck of cards. Use (a) and Exercise 4.6.1*

(c) Find the probability that the second and fourth card drawn have the same colour.

Ex. 4.6.4. **(Polya Urn Scheme)** An urn contains b black balls and r red balls. A ball is drawn at random and its colour noted. Then it is replaced along with $c \geq 0$ balls of the same colour. This procedure is repeated n times.

(a) Let $1 \leq k \leq m \leq n$. Show that

$$P(k \text{ black balls are drawn in } m \text{ draws}) = \binom{m}{k} \frac{\prod_{i=0}^{k-1}(b+ci)\prod_{i=0}^{m-k-1}(r+ci)}{\prod_{i=0}^{m-1}(r+b+ci)}$$

(b) Let $1 \leq i \leq n$ and random variables X_i be given by

$$X_i = \begin{cases} 1 & \text{if } i\text{-th ball drawn is black} \\ 0 & \text{otherwise} \end{cases}$$

Show that the collection of random variables is exchangeable.

(c) Let $1 \leq m \leq n$. Let B_m be the event that the m-th ball drawn is black. Show that

$$P(B_m) = \frac{b}{b+r}.$$