***17.** Prove Theorem 2, the extended form of Bayes' theorem. That is, suppose that E is an event from a sample space S and that $F_1, F_2, \ldots, F_n$ are mutually exclusive events such that $\bigcup_{i=1}^{n} F_i = S$. Assume that $p(E) \neq 0$ and $p(F_i) \neq 0$ for $i = 1, 2, \ldots, n$. Show that

$$p(F_j \mid E) = \frac{p(E \mid F_j)p(F_j)}{\sum_{i=1}^{n} p(E \mid F_i)p(F_i)}.$$

[*Hint:* Use the fact that $E = \bigcup_{i=1}^{n}(E \cap F_i)$.]

18. Suppose that a Bayesian spam filter is trained on a set of 500 spam messages and 200 messages that are not spam. The word "exciting" appears in 40 spam messages and in 25 messages that are not spam. Would an incoming message be rejected as spam if it contains the word "exciting" and the threshold for rejecting spam is 0.9?

19. Suppose that a Bayesian spam filter is trained on a set of 1000 spam messages and 400 messages that are not spam. The word "opportunity" appears in 175 spam messages and 20 messages that are not spam. Would an incoming message be rejected as spam if it contains the word "opportunity" and the threshold for rejecting a message is 0.9?

20. Would we reject a message as spam in Example 4

a) using just the fact that the word "undervalued" occurs in the message?

b) using just the fact that the word "stock" occurs in the message?

21. Suppose that a Bayesian spam filter is trained on a set of 10,000 spam messages and 5000 messages that are not spam. The word "enhancement" appears in 1500 spam messages and 20 messages that are not spam, while the word "herbal" appears in 800 spam messages and 200 messages that are not spam. Estimate the probability that a received message containing both the words "enhancement" and "herbal" is spam. Will the message be rejected as spam if the threshold for rejecting spam is 0.9?

22. Suppose that we have prior information concerning whether a random incoming message is spam. In particular, suppose that over a time period, we find that s spam messages arrive and h messages arrive that are not spam.

a) Use this information to estimate $p(S)$, the probability that an incoming message is spam, and $p(\overline{S})$, the probability an incoming message is not spam.

b) Use Bayes' theorem and part (a) to estimate the probability that an incoming message containing the word w is spam, where $p(w)$ is the probability that w occurs in a spam message and $q(w)$ is the probability that w occurs in a message that is not spam.

23. Suppose that E_1 and E_2 are the events that an incoming mail message contains the words w_1 and w_2, respectively. Assuming that E_1 and E_2 are independent events and that $E_1 \mid S$ and $E_2 \mid S$ are independent events, where S is the event that an incoming message is spam, and that we have no prior knowledge regarding whether or not the message is spam, show that

$$p(S \mid E_1 \cap E_2) = \frac{p(E_1 \mid S)p(E_2 \mid S)}{p(E_1 \mid S)p(E_2 \mid S) + p(E_1 \mid \overline{S})p(E_2 \mid \overline{S})}.$$

7.4 Expected Value and Variance

Introduction

The **expected value** of a random variable is the sum over all elements in a sample space of the product of the probability of the element and the value of the random variable at this element. Consequently, the expected value is a weighted average of the values of a random variable. The expected value of a random variable provides a central point for the distribution of values of this random variable. We can solve many problems using the notion of the expected value of a random variable, such as determining who has an advantage in gambling games and computing the average-case complexity of algorithms. Another useful measure of a random variable is its **variance**, which tells us how spread out the values of this random variable are. We can use the variance of a random variable to help us estimate the probability that a random variable takes values far removed from its expected value.

Expected Values

Links

Many questions can be formulated in terms of the value we expect a random variable to take, or more precisely, the average value of a random variable when an experiment is performed a large number of times. Questions of this kind include: How many heads are expected to appear when a coin is flipped 100 times? What is the expected number of comparisons used to find an element in a list using a linear search? To study such questions we introduce the concept of the expected value of a random variable.

DEFINITION 1 The *expected value*, also called the *expectation* or *mean*, of the random variable X on the sample space S is equal to

$$E(X) = \sum_{s \in S} p(s)X(s).$$

The *deviation* of X at $s \in S$ is $X(s) - E(X)$, the difference between the value of X and the mean of X.

Note that when the sample space S has n elements $S = \{x_1, x_2, \ldots, x_n\}$, $E(X) = \sum_{i=1}^{n} p(x_i)X(x_i)$.

Remark: When there are infinitely many elements of the sample space, the expectation is defined only when the infinite series in the definition is absolutely convergent. In particular, the expectation of a random variable on an infinite sample space is finite if it exists.

EXAMPLE 1 **Expected Value of a Die** Let X be the number that comes up when a fair die is rolled. What is the expected value of X?

Solution: The random variable X takes the values 1, 2, 3, 4, 5, or 6, each with probability $1/6$. It follows that

$$E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{21}{6} = \frac{7}{2}.$$

◀

EXAMPLE 2 A fair coin is flipped three times. Let S be the sample space of the eight possible outcomes, and let X be the random variable that assigns to an outcome the number of heads in this outcome. What is the expected value of X?

Extra Examples

Solution: In Example 10 of Section 7.2 we listed the values of X for the eight possible outcomes when a coin is flipped three times. Because the coin is fair and the flips are independent, the probability of each outcome is $1/8$. Consequently,

$$\begin{aligned} E(X) &= \frac{1}{8}[X(HHH) + X(HHT) + X(HTH) + X(THH) + X(TTH) \\ &\quad + X(THT) + X(HTT) + X(TTT)] \\ &= \frac{1}{8}(3 + 2 + 2 + 2 + 1 + 1 + 1 + 0) = \frac{12}{8} \\ &= \frac{3}{2}. \end{aligned}$$

Consequently, the expected number of heads that come up when a fair coin is flipped three times is $3/2$. ◀

When an experiment has relatively few outcomes, we can compute the expected value of a random variable directly from its definition, as was done in Example 2. However, when an experiment has a large number of outcomes, it may be inconvenient to compute the expected value of a random variable directly from its definition. Instead, we can find the expected value of a random variable by grouping together all outcomes assigned the same value by the random variable, as Theorem 1 shows.

THEOREM 1 If X is a random variable and $p(X = r)$ is the probability that $X = r$, so that $p(X = r) = \sum_{s \in S, X(s)=r} p(s)$, then

$$E(X) = \sum_{r \in X(S)} p(X = r)r.$$

Proof: Suppose that X is a random variable with range $X(S)$, and let $p(X = r)$ be the probability that the random variable X takes the value r. Consequently, $p(X = r)$ is the sum of the probabilities of the outcomes s such that $X(s) = r$. It follows that

$$E(X) = \sum_{r \in X(S)} p(X = r)r.$$

◁

Example 3 and the proof of Theorem 2 will illustrate the use of this formula. In Example 3 we will find the expected value of the sum of the numbers that appear on two fair dice when they are rolled. In Theorem 2 we will find the expected value of the number of successes when n Bernoulli trials are performed.

EXAMPLE 3 What is the expected value of the sum of the numbers that appear when a pair of fair dice is rolled?

Solution: Let X be the random variable equal to the sum of the numbers that appear when a pair of dice is rolled. In Example 12 of Section 7.2 we listed the value of X for the 36 outcomes of this experiment. The range of X is {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. By Example 12 of Section 7.2 we see that

$$\begin{aligned}
&p(X = 2) = p(X = 12) = 1/36,\\
&p(X = 3) = p(X = 11) = 2/36 = 1/18,\\
&p(X = 4) = p(X = 10) = 3/36 = 1/12,\\
&p(X = 5) = p(X = 9) = 4/36 = 1/9,\\
&p(X = 6) = p(X = 8) = 5/36,\\
&p(X = 7) = 6/36 = 1/6.
\end{aligned}$$

Substituting these values in the formula, we have

$$\begin{aligned}
E(X) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{1}{18} + 4 \cdot \frac{1}{12} + 5 \cdot \frac{1}{9} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{1}{6}\\
+ 8 \cdot \frac{5}{36} + 9 \cdot \frac{1}{9} + 10 \cdot \frac{1}{12} + 11 \cdot \frac{1}{18} + 12 \cdot \frac{1}{36}\\
= 7.
\end{aligned}$$

◀

THEOREM 2 The expected number of successes when n mutually independent Bernoulli trials are performed, where p is the probability of success on each trial, is np.

Proof: Let X be the random variable equal to the number of successes in n trials. By Theorem 2 of Section 7.2 we see that $p(X = k) = C(n, k)p^k q^{n-k}$. Hence, we have

$$\begin{aligned}
E(X) &= \sum_{k=1}^{n} kp(X = k) && \text{by Theorem 1}\\
&= \sum_{k=1}^{n} kC(n, k)p^k q^{n-k} && \text{by Theorem 2 in Section 7.2}\\
&= \sum_{k=1}^{n} nC(n-1, k-1)p^k q^{n-k} && \text{by Exercise 21 in Section 6.4}\\
&= np\sum_{k=1}^{n} C(n-1, k-1)p^{k-1} q^{n-k} && \text{factoring } np \text{ from each term}\\
&= np\sum_{j=0}^{n-1} C(n-1, j)p^{j} q^{n-1-j} && \text{shifting index of summation with } j = k-1\\
&= np(p+q)^{n-1} && \text{by the binomial theorem}\\
&= np. && \text{because } p+q = 1
\end{aligned}$$

This completes the proof because it shows that the expected number of successes in n mutually independent Bernoulli trials is np. ◁

We will also show that the hypothesis that the Bernoulli trials are mutually independent in Theorem 2 is not necessary.

Linearity of Expectations

Theorem 3 tells us that expected values are linear. For example, the expected value of the sum of random variables is the sum of their expected values. We will find this property exceedingly useful.

THEOREM 3

If X_i, $i = 1, 2, \ldots, n$ with n a positive integer, are random variables on S, and if a and b are real numbers, then

(i) $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$

(ii) $E(aX + b) = aE(X) + b$.

Proof: Part *(i)* follows for $n = 2$ directly from the definition of expected value, because

$$\begin{aligned}
E(X_1 + X_2) &= \sum_{s\in S} p(s)(X_1(s) + X_2(s))\\
&= \sum_{s\in S} p(s)X_1(s) + \sum_{s\in S} p(s)X_2(s)\\
&= E(X_1) + E(X_2).
\end{aligned}$$

The case for n random variables follows easily by mathematical induction using the case of two random variables. (We leave it to the reader to complete the proof.)

To prove part *(ii)*, note that

$$
\begin{aligned}
E(aX+b) &= \textstyle\sum_{s\in S} p(s)(aX(s)+b) \\
&= a\textstyle\sum_{s\in S} p(s)X(s) + b\textstyle\sum_{s\in S} p(s) \\
&= aE(X)+b \text{ because } \textstyle\sum_{s\in S} p(s) = 1.
\end{aligned}
$$

◁

Examples 4 and 5 illustrate how to use Theorem 3.

EXAMPLE 4 Use Theorem 3 to find the expected value of the sum of the numbers that appear when a pair of fair dice is rolled. (This was done in Example 3 without the benefit of this theorem.)

Solution: Let X_1 and X_2 be the random variables with $X_1((i,j)) = i$ and $X_2((i,j)) = j$, so that X_1 is the number appearing on the first die and X_2 is the number appearing on the second die. It is easy to see that $E(X_1) = E(X_2) = 7/2$ because both equal $(1+2+3+4+5+6)/6 = 21/6 = 7/2$. The sum of the two numbers that appear when the two dice are rolled is the sum $X_1 + X_2$. By Theorem 3, the expected value of the sum is $E(X_1+X_2) = E(X_1) + E(X_2) = 7/2 + 7/2 = 7$. ◀

EXAMPLE 5 In the proof of Theorem 2 we found the expected value of the number of successes when n independent Bernoulli trials are performed, where p is the probability of success on each trial by direct computation. Show how Theorem 3 can be used to derive this result where the Bernoulli trials are not necessarily independent.

Solution: Let X_i be the random variable with $X_i((t_1, t_2, \ldots, t_n)) = 1$ if t_i is a success and $X_i((t_1, t_2, \ldots, t_n)) = 0$ if t_i is a failure. The expected value of X_i is $E(X_i) = 1\cdot p + 0\cdot(1-p) = p$ for $i = 1, 2, \ldots, n$. Let $X = X_1 + X_2 + \cdots + X_n$, so that X counts the number of successes when these n Bernoulli trials are performed. Theorem 3, applied to the sum of n random variables, shows that $E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = np$. ◀

We can take advantage of the linearity of expectations to find the solutions of many seemingly difficult problems. The key step is to express a random variable whose expectation we wish to find as the sum of random variables whose expectations are easy to find. Examples 6 and 7 illustrate this technique.

EXAMPLE 6 **Expected Value in the Hatcheck Problem** A new employee checks the hats of n people at a restaurant, forgetting to put claim check numbers on the hats. When customers return for their hats, the checker gives them back hats chosen at random from the remaining hats. What is the expected number of hats that are returned correctly?

Solution: Let X be the random variable that equals the number of people who receive the correct hat from the checker. Let X_i be the random variable with $X_i = 1$ if the ith person receives the correct hat and $X_i = 0$ otherwise. It follows that

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

Because it is equally likely that the checker returns any of the hats to this person, it follows that the probability that the ith person receives the correct hat is $1/n$. Consequently, by Theorem 1, for all i we have

$$E(X_i) = 1\cdot p(X_i = 1) + 0\cdot p(X_i = 0) = 1\cdot 1/n + 0 = 1/n.$$

By the linearity of expectations (Theorem 3), it follows that

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = n \cdot 1/n = 1.$$

Consequently, the average number of people who receive the correct hat is exactly 1. Note that this answer is independent of the number of people who have checked their hats! (We will find an explicit formula for the probability that no one receives the correct hat in Example 4 of Section 8.6.) ◀

EXAMPLE 7 **Expected Number of Inversions in a Permutation** The ordered pair (i, j) is called an **inversion** in a permutation of the first n positive integers if $i < j$ but j precedes i in the permutation. For instance, there are six inversions in the permutation 3, 5, 1, 4, 2; these inversions are

$$(1, 3), (1, 5), (2, 3), (2, 4), (2, 5), (4, 5).$$

Let $I_{i,j}$ be the random variable on the set of all permutations of the first n positive integers with $I_{i,j} = 1$ if (i, j) is an inversion of the permutation and $I_{i,j} = 0$ otherwise. It follows that if X is the random variable equal to the number of inversions in the permutation, then

$$X = \sum_{1 \le i < j \le n} I_{i,j}.$$

Note that it is equally likely for i to precede j in a randomly chosen permutation as it is for j to precede i. (To see this, note that there are an equal number of permutations with each of these properties.) Consequently, for all pairs i and j we have

$$E(I_{i,j}) = 1 \cdot p(I_{i,j} = 1) + 0 \cdot p(I_{i,j} = 0) = 1 \cdot 1/2 + 0 = 1/2.$$

Because there are $\binom{n}{2}$ pairs i and j with $1 \le i < j \le n$ and by the linearity of expectations (Theorem 3), we have

$$E(X) = \sum_{1 \le i < j \le n} E(I_{i,j}) = \binom{n}{2} \cdot \frac{1}{2} = \frac{n(n-1)}{4}.$$

It follows that there are an average of $n(n-1)/4$ inversions in a permutation of the first n positive integers. ◀

Average-Case Computational Complexity

Links

Computing the average-case computational complexity of an algorithm can be interpreted as computing the expected value of a random variable. Let the sample space of an experiment be the set of possible inputs a_j, $j = 1, 2, \ldots, n$, and let X be the random variable that assigns to a_j the number of operations used by the algorithm when given a_j as input. Based on our knowledge of the input, we assign a probability $p(a_j)$ to each possible input value a_j. Then, the average-case complexity of the algorithm is

$$E(X) = \sum_{j=1}^{n} p(a_j)X(a_j).$$

This is the expected value of X.

Finding the average-case computational complexity of an algorithm is usually much more difficult than finding its worst-case computational complexity, and often involves the use of sophisticated methods. However, there are some algorithms for which the analysis required to find the average-case computational complexity is not difficult. For instance, in Example 8 we will illustrate how to find the average-case computational complexity of the linear search algorithm under different assumptions concerning the probability that the element for which we search is an element of the list.

EXAMPLE 8 **Average-Case Complexity of the Linear Search Algorithm** We are given a real number x and a list of n distinct real numbers. The linear search algorithm, described in Section 3.1, locates x by successively comparing it to each element in the list, terminating when x is located or when all the elements have been examined and it has been determined that x is not in the list. What is the average-case computational complexity of the linear search algorithm if the probability that x is in the list is p and it is equally likely that x is any of the n elements in the list? (There are $n+1$ possible types of input: one type for each of the n numbers in the list and a last type for numbers not in the list, which we treat as a single input.)

Solution: In Example 4 of Section 3.3 we showed that $2i+1$ comparisons are used if x equals the ith element of the list and, in Example 2 of Section 3.3, we showed that $2n+2$ comparisons are used if x is not in the list. The probability that x equals a_i, the ith element in the list, is p/n, and the probability that x is not in the list is $q = 1-p$. It follows that the average-case computational complexity of the linear search algorithm is

$$\begin{aligned}
E &= \frac{3p}{n} + \frac{5p}{n} + \cdots + \frac{(2n+1)p}{n} + (2n+2)q \\
&= \frac{p}{n}(3+5+\cdots+(2n+1)) + (2n+2)q \\
&= \frac{p}{n}((n+1)^2-1) + (2n+2)q \\
&= p(n+2) + (2n+2)q.
\end{aligned}$$

(The third equality follows from Example 2 of Section 5.1.) For instance, when x is guaranteed to be in the list, we have $p=1$ (so the probability that $x=a_i$ is $1/n$ for each i) and $q=0$. Then $E=n+2$, as we showed in Example 4 in Section 3.3.

When p, the probability that x is in the list, is $1/2$, it follows that $q=1-p=1/2$, so $E=(n+2)/2+n+1=(3n+4)/2$. Similarly, if the probability that x is in the list is $3/4$, we have $p=3/4$ and $q=1/4$, so $E=3(n+2)/4+(n+1)/2=(5n+8)/4$.

Finally, when x is guaranteed not to be in the list, we have $p=0$ and $q=1$. It follows that $E=2n+2$, which is not surprising because we have to search the entire list. ◀

Example 9 illustrates how the linearity of expectations can help us find the average-case complexity of a sorting algorithm, the insertion sort.

EXAMPLE 9 **Average-Case Complexity of the Insertion Sort** What is the average number of comparisons used by the insertion sort to sort n distinct elements?

Solution: We first suppose that X is the random variable equal to the number of comparisons used by the insertion sort (described in Section 3.1) to sort a list $a_1, a_2, \ldots, a_n$ of n distinct elements. Then $E(X)$ is the average number of comparisons used. (Recall that at step i for $i=2,\ldots,n$, the insertion sort inserts the ith element in the original list into the correct position in the sorted list of the first $i-1$ elements of the original list.)

We let X_i be the random variable equal to the number of comparisons used to insert a_i into the proper position after the first $i - 1$ elements $a_1, a_2, \ldots, a_{i-1}$ have been sorted. Because

$$X = X_2 + X_3 + \cdots + X_n,$$

we can use the linearity of expectations to conclude that

$$E(X) = E(X_2 + X_3 + \cdots + X_n) = E(X_2) + E(X_3) + \cdots + E(X_n).$$

To find $E(X_i)$ for $i = 2, 3, \ldots, n$, let $p_j(k)$ denote the probability that the largest of the first j elements in the list occurs at the kth position, that is, that $\max(a_1, a_2, \ldots, a_j) = a_k$, where $1 \le k \le j$. Because the elements of the list are randomly distributed, it is equally likely for the largest element among the first j elements to occur at any position. Consequently, $p_j(k) = 1/j$. If $X_i(k)$ equals the number of comparisons used by the insertion sort if a_i is inserted into the kth position in the list once $a_1, a_2, \ldots, a_{i-1}$ have been sorted, it follows that $X_i(k) = k$. Because it is possible that a_i is inserted in any of the first i positions, we find that

$$E(X_i) = \sum_{k=1}^{i} p_i(k) \cdot X_i(k) = \sum_{k=1}^{i} \frac{1}{i} \cdot k = \frac{1}{i} \cdot \sum_{k=1}^{i} k = \frac{1}{i} \cdot \frac{i(i+1)}{2} = \frac{i+1}{2}.$$

It follows that

$$\begin{aligned} E(X) &= \sum_{i=2}^{n} E(X_i) = \sum_{i=2}^{n} \frac{i+1}{2} = \frac{1}{2} \sum_{j=3}^{n+1} j \\ &= \frac{1}{2} \frac{(n+1)(n+2)}{2} - \frac{1}{2}(1+2) = \frac{n^2 + 3n - 4}{4}. \end{aligned}$$

To obtain the third of these equalities we shifted the index of summation, setting $j = i + 1$. To obtain the fourth equality, we used the formula $\sum_{k=1}^{m} k = m(m+1)/2$ (from Table 2 in Section 2.4) with $m = n + 1$, subtracting off the missing terms with $j = 1$ and $j = 2$. We conclude that the average number of comparisons used by the insertion sort to sort n elements equals $(n^2 + 3n - 4)/4$, which is $\Theta(n^2)$. ◀

The Geometric Distribution

We now turn our attention to a random variable with infinitely many possible outcomes.

EXAMPLE 10 Suppose that the probability that a coin comes up tails is p. This coin is flipped repeatedly until it comes up tails. What is the expected number of flips until this coin comes up tails?

Links

Solution: We first note that the sample space consists of all sequences that begin with any number of heads, denoted by H, followed by a tail, denoted by T. Therefore, the sample space is the set $\{T, HT, HHT, HHHT, HHHHT, \ldots\}$. Note that this is an infinite sample space. We can determine the probability of an element of the sample space by noting that the coin flips are independent and that the probability of a head is $1 - p$. Therefore, $p(T) = p$, $p(HT) = (1 - p)p$, $p(HHT) = (1 - p)^2 p$, and in general the probability that the coin is flipped n times before a tail comes up, that is, that $n - 1$ heads come up followed by a tail, is $(1 - p)^{n-1} p$. (Exercise 14 asks for a verification that the sum of the probabilities of the points in the sample space is 1.)

Now let X be the random variable equal to the number of flips in an element in the sample space. That is, $X(T) = 1$, $X(HT) = 2$, $X(HHT) = 3$, and so on. Note that $p(X = j) = (1 - p)^{j-1}p$. The expected number of flips until the coin comes up tails equals $E(X)$.

Using Theorem 1, we find that

$$E(X) = \sum_{j=1}^{\infty} j \cdot p(X = j) = \sum_{j=1}^{\infty} j(1-p)^{j-1}p = p\sum_{j=1}^{\infty} j(1-p)^{j-1} = p \cdot \frac{1}{p^2} = \frac{1}{p}.$$

[The third equality in this chain follows from Table 2 in Section 2.4, which tells us that $\sum_{j=1}^{\infty} j(1-p)^{j-1} = 1/(1-(1-p))^2 = 1/p^2$.] It follows that the expected number of times the coin is flipped until tails comes up is $1/p$. Note that when the coin is fair we have $p = 1/2$, so the expected number of flips until it comes up tails is $1/(1/2) = 2$. ◀

The random variable X that equals the number of flips expected before a coin comes up tails is an example of a random variable with a **geometric distribution**.

DEFINITION 2

A random variable X has a *geometric distribution with parameter* p if $p(X = k) = (1-p)^{k-1}p$ for $k = 1, 2, 3, \ldots$, where p is a real number with $0 \leq p \leq 1$.

Geometric distributions arise in many applications because they are used to study the time required before a particular event happens, such as the time required before we find an object with a certain property, the number of attempts before an experiment succeeds, the number of times a product can be used before it fails, and so on.

When we computed the expected value of the number of flips required before a coin comes up tails, we proved Theorem 4.

THEOREM 4

If the random variable X has the geometric distribution with parameter p, then $E(X) = 1/p$.

Independent Random Variables

We have already discussed independent events. We will now define what it means for two random variables to be independent.

DEFINITION 3

The random variables X and Y on a sample space S are *independent* if

$$p(X = r_1 \text{ and } Y = r_2) = p(X = r_1) \cdot p(Y = r_2),$$

or in words, if the probability that $X = r_1$ and $Y = r_2$ equals the product of the probabilities that $X = r_1$ and $Y = r_2$, for all real numbers r_1 and r_2.

EXAMPLE 11 Are the random variables X_1 and X_2 from Example 4 independent?

Extra Examples

Solution: Let $S = \{1, 2, 3, 4, 5, 6\}$, and let $i \in S$ and $j \in S$. Because there are 36 possible outcomes when the pair of dice is rolled and each is equally likely, we have

$$p(X_1 = i \text{ and } X_2 = j) = 1/36.$$

Furthermore, $p(X_1 = i) = 1/6$ and $p(X_2 = j) = 1/6$, because the probability that i appears on the first die and the probability that j appears on the second die are both $1/6$. It follows that

$$p(X_1 = i \text{ and } X_2 = j) = \frac{1}{36} \quad \text{and} \quad p(X_1 = i)p(X_2 = j) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36},$$

so X_1 and X_2 are independent. ◀

EXAMPLE 12 Show that the random variables X_1 and $X = X_1 + X_2$, where X_1 and X_2 are as defined in Example 4, are not independent.

Solution: Note that $p(X_1 = 1 \text{ and } X = 12) = 0$, because $X_1 = 1$ means the number appearing on the first die is 1, which implies that the sum of the numbers appearing on the two dice cannot equal 12. On the other hand, $p(X_1 = 1) = 1/6$ and $p(X = 12) = 1/36$. Hence $p(X_1 = 1 \text{ and } X = 12) \neq p(X_1 = 1) \cdot p(X = 12)$. This counterexample shows that X_1 and X are not independent. ◀

The expected value of the product of two independent random variables is the product of their expected values, as Theorem 5 shows.

THEOREM 5 If X and Y are independent random variables on a sample space S, then $E(XY) = E(X)E(Y)$.

Proof: To prove this formula, we use the key observation that the event $XY = r$ is the disjoint union of the events $X = r_1$ and $Y = r_2$ over all $r_1 \in X(S)$ and $r_2 \in Y(S)$ with $r = r_1 r_2$. We have

$$\begin{aligned}
E(XY) &= \sum_{r \in XY(S)} r \cdot p(XY = r) && \text{by Theorem 1} \\
&= \sum_{r_1 \in X(S), r_2 \in Y(S)} r_1 r_2 \cdot p(X = r_1 \text{ and } Y = r_2) && \text{expressing } XY = r \text{ as a disjoint union} \\
&= \sum_{r_1 \in X(S)} \sum_{r_2 \in Y(S)} r_1 r_2 \cdot p(X = r_1 \text{ and } Y = r_2) && \text{using a double sum to order the terms} \\
&= \sum_{r_1 \in X(S)} \sum_{r_2 \in Y(S)} r_1 r_2 \cdot p(X = r_1) \cdot p(Y = r_2) && \text{by the independence of } X \text{ and } Y \\
&= \sum_{r_1 \in X(S)} \Big(r_1 \cdot p(X = r_1) \cdot \sum_{r_2 \in Y(S)} r_2 \cdot p(Y = r_2)\Big) && \text{by factoring out } r_1 \cdot p(X = r_1) \\
&= \sum_{r_1 \in X(S)} r_1 \cdot p(X = r_1) \cdot E(Y) && \text{by the definition of } E(Y) \\
&= E(Y)\Big(\sum_{r_1 \in X(S)} r_1 \cdot p(X = r_1)\Big) && \text{by factoring out } E(Y) \\
&= E(Y)E(X) && \text{by the definition of } E(X)
\end{aligned}$$

We complete the proof by noting that $E(Y)E(X) = E(X)E(Y)$, which is a consequence of the commutative law for multiplication. ◁

Note that when X and Y are random variables that are not independent, we cannot conclude that $E(XY) = E(X)E(Y)$, as Example 13 shows.

EXAMPLE 13 Let X and Y be random variables that count the number of heads and the number of tails when a coin is flipped twice. Because $p(X = 2) = 1/4$, $p(X = 1) = 1/2$, and $p(X = 0) = 1/4$, by Theorem 1 we have

$$E(X) = 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4} = 1.$$

A similar computation shows that $E(Y) = 1$. We note that $XY = 0$ when either two heads and no tails or two tails and no heads come up and that $XY = 1$ when one head and one tail come up. Hence,

$$E(XY) = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}.$$

It follows that

$$E(XY) \neq E(X)E(Y).$$

This does not contradict Theorem 5 because X and Y are not independent, as the reader should verify (see Exercise 16). ◀

Variance

Links

The expected value of a random variable tells us its average value, but nothing about how widely its values are distributed. For example, if X and Y are the random variables on the set $S = \{1, 2, 3, 4, 5, 6\}$, with $X(s) = 0$ for all $s \in S$ and $Y(s) = -1$ if $s \in \{1, 2, 3\}$ and $Y(s) = 1$ if $s \in \{4, 5, 6\}$, then the expected values of X and Y are both zero. However, the random variable X never varies from 0, while the random variable Y always differs from 0 by 1. The variance of a random variable helps us characterize how widely a random variable is distributed. In particular, it provides a measure of how widely X is distributed about its expected value.

DEFINITION 4 Let X be a random variable on a sample space S. The *variance* of X, denoted by $V(X)$, is

$$V(X) = \sum_{s \in S} (X(s) - E(X))^2 p(s).$$

That is, $V(X)$ is the weighted average of the square of the deviation of X. The *standard deviation* of X, denoted $\sigma(X)$, is defined to be $\sqrt{V(X)}$.

Theorem 6 provides a useful simple expression for the variance of a random variable.

THEOREM 6 If X is a random variable on a sample space S, then $V(X) = E(X^2) - E(X)^2$.

Proof: Note that

$$\begin{aligned} V(X) &= \sum_{s \in S} (X(s) - E(X))^2 p(s) \\ &= \sum_{s \in S} X(s)^2 p(s) - 2E(X) \sum_{s \in S} X(s)p(s) + E(X)^2 \sum_{s \in S} p(s) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2. \end{aligned}$$

We have used the fact that $\sum_{s \in S} p(s) = 1$ in the next-to-last step. ◁

We can use Theorems 3 and 6 to derive an alternative formula for $V(X)$ that provides some insight into the meaning of the variance of a random variable.

COROLLARY 1 If X is a random variable on a sample space S and $E(X) = \mu$, then $V(X) = E((X - \mu)^2)$.

μ is the Greek letter mu.

Proof: If X is a random variable with $E(X) = \mu$, then

$$\begin{aligned}
E((X-\mu)^2) &= E(X^2 - 2\mu X + \mu^2) && \text{expanding } (X-\mu)^2 \\
&= E(X^2) - E(2\mu X) + E(\mu^2) && \text{by part } (i) \text{ of Theorem 3} \\
&= E(X^2) - 2\mu E(X) + E(\mu^2) && \text{by part } (ii) \text{ of Theorem 3, noting that } \mu \text{ is a constant} \\
&= E(X^2) - 2\mu E(X) + \mu^2 && \text{as } E(\mu^2) = \mu^2 \text{, because } \mu^2 \text{ is a constant} \\
&= E(X^2) - 2\mu^2 + \mu^2 && \text{because } E(X) = \mu \\
&= E(X^2) - \mu^2 && \text{simplifying} \\
&= V(X) && \text{by Theorem 6 and noting that } E(X) = \mu.
\end{aligned}$$

This completes the proof. ◁

Corollary 1 tells us that the variance of a random variable X is the expected value of the square of the difference between X and its own expected value. This is commonly expressed as saying that the variance of X is the mean of the square of its deviation. We also say that the standard deviation of X is the square root of the mean of the square of its deviation (often read as the “root mean square” of the deviation).

We now compute the variance of some random variables.

EXAMPLE 14 What is the variance of the random variable X with $X(t) = 1$ if a Bernoulli trial is a success and $X(t) = 0$ if it is a failure, where p is the probability of success and q is the probability of failure?

Extra Examples

Solution: Because X takes only the values 0 and 1, it follows that $X^2(t) = X(t)$. Hence,

$$V(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p) = pq.$$ ◀

EXAMPLE 15 **Variance of the Value of a Die** What is the variance of the random variable X, where X is the number that comes up when a fair die is rolled?

Solution: We have $V(X) = E(X^2) - E(X)^2$. By Example 1 we know that $E(X) = 7/2$. To find $E(X^2)$ note that X^2 takes the values i^2, $i = 1, 2, \ldots, 6$, each with probability $1/6$. It follows that

$$E(X^2) = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}.$$

We conclude that

$$V(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$ ◀

EXAMPLE 16 What is the variance of the random variable $X((i, j)) = 2i$, where i is the number appearing on the first die and j is the number appearing on the second die, when two fair dice are rolled?

Solution: We will use Theorem 6 to find the variance of X. To do so, we need to find the expected values of X and X^2. Note that because $p(X = k)$ is $1/6$ for $k = 2, 4, 6, 8, 10, 12$ and is 0 otherwise,

$$E(X) = (2 + 4 + 6 + 8 + 10 + 12)/6 = 7,$$

and

$$E(X^2) = (2^2 + 4^2 + 6^2 + 8^2 + 10^2 + 12^2)/6 = 182/3.$$

It follows from Theorem 6 that

$$V(X) = E(X^2) - E(X)^2 = 182/3 - 49 = 35/3.$$ ◀

Another useful property is that the variance of the sum of two or more independent random variables is the sum of their variances. The formula that expresses this property is known as **Bienaymé's formula**, after Irenée-Jules Bienaymé, the French mathematician who discovered it in 1853. Bienaymé's formula is useful for computing the variance of the result of n independent Bernoulli trials, for instance.

THEOREM 7 **BIENAYMÉ'S FORMULA** If X and Y are two independent random variables on a sample space S, then $V(X + Y) = V(X) + V(Y)$. Furthermore, if $X_i, i = 1, 2, \ldots, n$, with n a positive integer, are pairwise independent random variables on S, then $V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n)$.

Links

IRENÉE-JULES BIENAYMÉ (1796–1878) Bienaymé, born in Paris, moved with his family to Bruges in 1803 when his father became a government administrator. Bienaymé attended the Lycée impérial in Bruges, and when his family returned to Paris in 1811, the Lycée Louis-le-Grand. As a teenager, he helped defend Paris during the 1814 Napoleonic Wars; in 1815, he became a student at the École Polytechnique. In 1816 he joined the Ministry of Finances to help support his family. In 1819, he left the civil service, taking a job lecturing mathematics at the Académie militaire de Saint-Cyr. Unhappy with conditions there, he soon returned to the Ministry of Finances. He attained the position of inspector general, remaining until forced to retire in 1848 for political reasons. He was able to return as inspector general in 1850, but he retired a second time in 1852. In 1851 he briefly was professor at the Sorbonne and also served as an expert statistician for Napoleon III. Bienaymé was one of the founders of the Société Mathématique de France, and in 1875 was its president.

Bienaymé was noted for his ingenuity, but his papers frustrated readers by omitting important proofs. He published sparsely, often in obscure journals. However, he made important contributions to probability and statistics, and to their applications to the social sciences and to finance. Among his important contributions are the Bienaymé-Chebyshev inequality, which provides a simple proof of the law of large numbers, a generalization of Laplace's least square method, and Bienaymé's formula for the variance of a sum of random variables. He studied the extinction of aristocratic families, declining despite general population growth. Bienaymé was a skilled linguist; he translated the works of Chebyshev, a close friend, from Russian to French. It has been suggested that his relative obscurity results from his modesty, his lack of interest in asserting the priority of his discoveries, and the fact that his work was often ahead of its time. He and his brother married two sisters who were daughters of a family friend. Bienaymé and his wife had two sons and three daughters.

Proof: From Theorem 6, we have

$$V(X+Y) = E((X+Y)^2) - E(X+Y)^2.$$

It follows that

$$\begin{aligned} V(X+Y) &= E(X^2 + 2XY + Y^2) - (E(X) + E(Y))^2 \\ &= E(X^2) + 2E(XY) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2. \end{aligned}$$

Because X and Y are independent, by Theorem 5 we have $E(XY) = E(X)E(Y)$. It follows that

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

We leave the proof of the case for n pairwise independent random variables to the reader (Exercise 34). Such a proof can be constructed by generalizing the proof we have given for the case for two random variables. Note that it is not possible to use mathematical induction in a straightforward way to prove the general case (see Exercise 33). ◁

EXAMPLE 17 Find the variance and standard deviation of the random variable X whose value when two fair dice are rolled is $X((i, j)) = i + j$, where i is the number appearing on the first die and j is the number appearing on the second die.

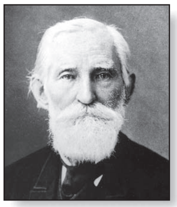
Solution: Let X_1 and X_2 be the random variables defined by $X_1((i, j)) = i$ and $X_2((i, j)) = j$ for a roll of the dice. Then $X = X_1 + X_2$, and X_1 and X_2 are independent, as Example 11 showed. From Theorem 7 it follows that $V(X) = V(X_1) + V(X_2)$. A simple computation as in Example 16, together with Exercise 29 in the Supplementary Exercises, tells us that $V(X_1) = V(X_2) = 35/12$. Hence, $V(X) = 35/12 + 35/12 = 35/6$ and $\sigma(X) = \sqrt{35/6}$. ◀

We will now find the variance of the random variable that counts the number of successes when n independent Bernoulli trials are carried out.

EXAMPLE 18 What is the variance of the number of successes when n independent Bernoulli trials are performed, where, on each trial, p is the probability of success and q is the probability of failure?

Solution: Let X_i be the random variable with $X_i((t_1, t_2, \ldots, t_n)) = 1$ if trial t_i is a success and $X_i((t_1, t_2, \ldots, t_n)) = 0$ if trial t_i is a failure. Let $X = X_1 + X_2 + \cdots + X_n$. Then X counts the number of successes in the n trials. From Theorem 7 it follows that $V(X) = V(X_1) + V(X_2) + \cdots + V(X_n)$. Using Example 14 we have $V(X_i) = pq$ for $i = 1, 2, \ldots, n$. It follows that $V(X) = npq$. ◀

Links



PAFNUTY LVOVICH CHEBYSHEV (1821–1894) Chebyshev was born into the gentry in Okatovo, Russia. His father was a retired army officer who had fought against Napoleon. In 1832 the family, with its nine children, moved to Moscow, where Pafnuty completed his high school education at home. He entered the Department of Physics and Mathematics at Moscow University. As a student, he developed a new method for approximating the roots of equations. He graduated from Moscow University in 1841 with a degree in mathematics, and he continued his studies, passing his master's exam in 1843 and completing his master's thesis in 1846.

Chebyshev was appointed in 1847 to a position as an assistant at the University of St. Petersburg. He wrote and defended a thesis in 1847. He became a professor at St. Petersburg in 1860, a position he held until 1882. His book on the theory of congruences written in 1849 was influential in the development of number theory. His work on the distribution of prime numbers was seminal. He proved Bertrand's conjecture that for every integer $n > 3$, there is a prime between n and $2n - 2$. Chebyshev helped develop ideas that were later used to prove the prime number theorem. Chebyshev's work on the approximation of functions using polynomials is used extensively when computers are used to find values of functions. Chebyshev was also interested in mechanics. He studied the conversion of rotary motion into rectilinear motion by mechanical coupling. The Chebyshev parallel motion is three linked bars approximating rectilinear motion.

Chebyshev's Inequality

How likely is it that a random variable takes a value far from its expected value? Theorem 8, called Chebyshev's inequality, helps answer this question by providing an upper bound on the probability that the value of a random variable differs from the expected value of the random variable by more than a specified amount.

THEOREM 8 **CHEBYSHEV'S INEQUALITY** Let X be a random variable on a sample space S with probability function p. If r is a positive real number, then

$$p(|X(s) - E(X)| \geq r) \leq V(X)/r^2.$$

Proof: Let A be the event

$$A = \{s \in S \mid |X(s) - E(X)| \geq r\}.$$

What we want to prove is that $p(A) \leq V(X)/r^2$. Note that

$$\begin{aligned} V(X) &= \sum_{s\in S}(X(s) - E(X))^2 p(s) \\ &= \sum_{s\in A}(X(s) - E(X))^2 p(s) + \sum_{s\notin A}(X(s) - E(X))^2 p(s). \end{aligned}$$

The second sum in this expression is nonnegative, because each of its summands is nonnegative. Also, because for each element s in A, $(X(s) - E(X))^2 \geq r^2$, the first sum in this expression is at least $\sum_{s\in A} r^2 p(s)$. Hence, $V(X) \geq \sum_{s\in A} r^2 p(s) = r^2 p(A)$. It follows that $V(X)/r^2 \geq p(A)$, so $p(A) \leq V(X)/r^2$, completing the proof. ◁

EXAMPLE 19 **Deviation from the Mean when Counting Tails** Suppose that X is the random variable that counts the number of tails when a fair coin is tossed n times. Note that X is the number of successes when n independent Bernoulli trials, each with probability of success $1/2$, are performed. It follows that $E(X) = n/2$ (by Theorem 2) and $V(X) = n/4$ (by Example 18). Applying Chebyshev's inequality with $r = \sqrt{n}$ shows that

$$p(|X(s) - n/2| \geq \sqrt{n}) \leq (n/4)/(\sqrt{n})^2 = 1/4.$$

Consequently, the probability is no more than $1/4$ that the number of tails that come up when a fair coin is tossed n times deviates from the mean by more than $\sqrt{n}$. ◀

Chebyshev's inequality, although applicable to any random variable, often fails to provide a practical estimate for the probability that the value of a random variable exceeds its mean by a large amount. This is illustrated by Example 20.

EXAMPLE 20 Let X be the random variable whose value is the number appearing when a fair die is rolled. We have $E(X) = 7/2$ (see Example 1) and $V(X) = 35/12$ (see Example 15). Because the only possible values of X are 1, 2, 3, 4, 5, and 6, X cannot take a value more than $5/2$ from its mean, $E(X) = 7/2$. Hence, $p(|X - 7/2| \geq r) = 0$ if $r > 5/2$. By Chebyshev's inequality we know that $p(|X - 7/2| \geq r) \leq (35/12)/r^2$.

For example, when $r = 3$, Chebyshev's inequality tells us that $p(|X - 7/2| \geq 3) \leq (35/12)/9 = 35/108 \approx 0.324$, which is a poor estimate, because $p(|X - 7/2| \geq 3) = 0$. ◀