4.1: Definitions and Basic Properties

Expected value is one of the most important concepts in probability. The expected value of a real-valued random variable gives the center of the distribution of the variable, in a special sense. Additionally, by computing expected values of various real transformations of a general random variable, we con extract a number of interesting characteristics of the distribution of the variable, including measures of spread, symmetry, and correlation. In a sense, expected value is a more general concept than probability itself.

Basic Concepts

Definitions

As usual, we start with a random experiment modeled by a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. So to review, Ω is the set of outcomes, $\mathscr{F}$ the collection of events and $\mathbb{P}$ the probability measure on the sample space $(\Omega, \mathscr{F})$. In the following definitions, we assume that X is a random variable for the experiment, taking values in $S \subseteq \mathbb{R}$.

If X has a discrete distribution with probability density function f (so that S is countable), then the *expected value* of X is defined as follows (assuming that the sum is well defined):

$$\mathbb{E}(X) = \sum_{x \in S} x f(x) \tag{4.1.1}$$

The sum defining the expected value makes sense if either the sum over the positive $x \in S$ is finite or the sum over the negative $x \in S$ is finite (or both). This ensures the that the entire sum exists (as an extended real number) and does not depend on the order of the terms. So as we will see, it's possible for $\mathbb{E}(X)$ to be a real number or ∞ or $-\infty$ or to simply not exist. Of course, if S is finite the expected value always exists as a real number.

If X has a continuous distribution with probability density function f (and so S is typically an interval or a union of disjoint intervals), then the *expected value* of X is defined as follows (assuming that the integral is well defined):

$$\mathbb{E}(X) = \int_S x f(x)\,dx \tag{4.1.2}$$

The probability density functions in basic applied probability that describe continuous distributions are piecewise continuous. So the integral above makes sense if the integral over positive $x \in S$ is finite or the integral over negative $x \in S$ is finite (or both). This ensures that the entire integral exists (as an extended real number). So as in the discrete case, it's possible for $\mathbb{E}(X)$ to exist as a real number or as ∞ or as $-\infty$ or to not exist at all. As you might guess, the definition for a mixed distribution is a combination of the definitions for the discrete and continuous cases.

If X has a mixed distribution, with partial discrete density g on D and partial continuous density h on C, where D and C are disjoint, D is countable, C is typically an interval, and $S = D \cup C$. The expected value of X is defined as follows (assuming that the expression on the right is well defined):

$$\mathbb{E}(X) = \sum_{x \in D} x g(x) + \int_C x h(x)\,dx \tag{4.1.3}$$

For the expected value above to make sense, the sum must be well defined, as in the discrete case, the integral must be well defined, as in the continuous case, and we must avoid the dreaded indeterminate form $\infty - \infty$. In the next section on additional properties, we will see that the various definitions given here can be unified into a single definition that works regardless of the type of distribution of X. An even more general definition is given in the advanced section on expected value as an integral.

Interpretation

The expected value of X is also called the *mean* of the distribution of X and is frequently denoted μ. The mean is the center of the probability distribution of X in a special sense. Indeed, if we think of the distribution as a mass distribution (with total mass 1), then the mean is the *center of mass* as defined in physics. The two pictures below show discrete and continuous probability density functions; in each case the mean μ is the center of mass, the balance point.

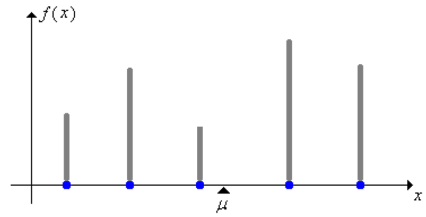

Figure 4.1.1: The mean μ as the center of mass of a discrete distribution.

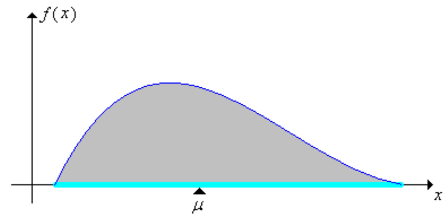

Figure 4.1.2: The mean μ as the center of mass of a continuous distribution.

Recall the other measures of the center of a distribution that we have studied:

- A mode is any $x \in S$ that maximizes f.
- A median is any $x \in \mathbb{R}$ that satisfies $\mathbb{P}(X < x) \le \frac{1}{2}$ and $\mathbb{P}(X \le x) \ge \frac{1}{2}$.

To understand expected value in a probabilistic way, suppose that we create a new, compound experiment by repeating the basic experiment over and over again. This gives a sequence of independent random variables $(X_1, X_2, \ldots)$, each with the same distribution as X. In statistical terms, we are sampling from the distribution of X. The average value, or *sample mean*, after n runs is

$$M_n = \frac{1}{n} \sum_{i=1}^n X_i \tag{4.1.4}$$

Note that M_n is a random variable in the compound experiment. The important fact is that the *average value* M_n converges to the *expected value* $\mathbb{E}(X)$ as $n \to \infty$. The precise statement of this is the law of large numbers, one of the fundamental theorems of probability. You will see the law of large numbers at work in many of the simulation exercises given below.

Extensions

If $a \in \mathbb{R}$ and $n \in \mathbb{N}$, the *moment of X about a of order n* is defined to be

$$\mathbb{E}\left[(X-a)^n\right] \tag{4.1.5}$$

(assuming of course that this expected value exists).

The moments about 0 are simply referred to as *moments* (or sometimes *raw moments*). The moments about μ are the *central moments*. The second central moment is particularly important, and is studied in detail in the section on variance. In some cases, if we know *all* of the moments of X, we can determine the entire distribution of X. This idea is explored in the section on generating functions.

The expected value of a random variable X is based, of course, on the probability measure $\mathbb{P}$ for the experiment. This probability measure could be a conditional probability measure, conditioned on a given event $A \in \mathscr{F}$ for the experiment (with $\mathbb{P}(A) > 0$). The usual notation is $\mathbb{E}(X \mid A)$, and this expected value is computed by the definitions given above, except that the conditional probability density function $x \mapsto f(x \mid A)$ replaces the ordinary probability density function f. It is very important to realize that, except for notation, no new concepts are involved. All results that we obtain for expected value in general have analogues for these conditional expected values. On the other hand, we will study a more general notion of conditional expected value in a later section.

Basic Properties

The purpose of this subsection is to study some of the essential properties of expected value. Unless otherwise noted, we will assume that the indicated expected values exist, and that the various sets and functions that we use are measurable. We start with two simple but still essential results.

Simple Variables

First, recall that a constant $c \in \mathbb{R}$ can be thought of as a random variable (on any probability space) that takes only the value c with probability 1. The corresponding distribution is sometimes called *point mass* at c.

If c is a constant random variable, then $\mathbb{E}(c) = c$.

Proof

As a random variable, c has a discrete distribution, so $\mathbb{E}(c) = c \cdot 1 = c$.

Next recall that an indicator variable is a random variable that takes only the values 0 and 1.

If X is an indicator variable then $\mathbb{E}(X) = \mathbb{P}(X = 1)$.

Proof

X is discrete so by definition, $\mathbb{E}(X) = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = \mathbb{P}(X = 1)$.

In particular, if $\mathbf{1}_A$ is the indicator variable of an event A, then $\mathbb{E}(\mathbf{1}_A) = \mathbb{P}(A)$, so in a sense, expected value subsumes probability. For a book that takes expected value, rather than probability, as the fundamental starting concept, see the book Probability via Expectation, by Peter Whittle.

Change of Variables Theorem

The expected value of a real-valued random variable gives the center of the distribution of the variable. This idea is much more powerful than might first appear. By finding expected values of various *functions* of a general random variable, we can measure many interesting features of its distribution.

Thus, suppose that X is a random variable taking values in a general set S, and suppose that r is a function from S into $\mathbb{R}$. Then $r(X)$ is a real-valued random variable, and so it makes sense to compute $\mathbb{E}[r(X)]$ (assuming as usual that this expected value exists). However, to compute this expected value from the definition would require that we know the probability density function of the transformed variable $r(X)$ (a difficult problem, in general). Fortunately, there is a much better way, given by the *change of variables theorem* for expected value. This theorem is sometimes referred to as the *law of the unconscious statistician*, presumably because it is so basic and natural that it is often used without the realization that it *is* a theorem, and not a definition.

If X has a discrete distribution on a countable set S with probability density function f. then

$$\mathbb{E}[r(X)] = \sum_{x \in S} r(x) f(x) \tag{4.1.6}$$

Proof

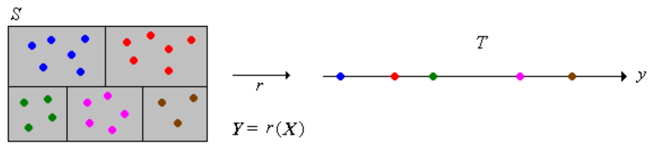

Figure 4.1.3: The change of variables theorem when X has a discrete distribution.

The next result is the change of variables theorem when X has a continuous distribution. We will prove the continuous version in stages, first when r has discrete range below and then in the next section in full generality. Even though the complete proof is delayed, however, we will use the change of variables theorem in the proofs of many of the other properties of expected value.

Suppose that X has a continuous distribution on $S \subseteq \mathbb{R}^n$ with probability density function f, and that $r: S \to \mathbb{R}$. Then

$$\mathbb{E}[r(X)] = \int_S r(x) f(x)\, dx \tag{4.1.7}$$

Proof when r has discrete range

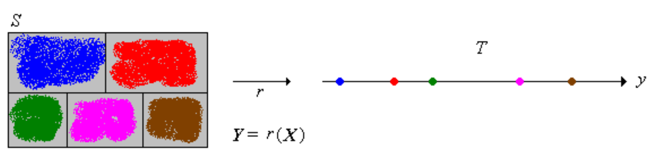

Figure 4.1.4: The change of variables theorem when X has a continuous distribution and r has countable range.

The results below gives basic properties of expected value. These properties are true in general, but we will restrict the proofs primarily to the continuous case. The proofs for the discrete case are analogous, with sums replacing integrals. The change of variables theorem is the main tool we will need. In these theorems X and Y are real-valued random variables for an experiment (that is, defined on an underlying probability space) and c is a constant. As usual, we assume that the indicated expected values exist. Be sure to try the proofs yourself before reading the ones in the text.

Linearity

Our first property is the *additive property*.

$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

Proof

We apply the change of variables theorem with the function $r(x, y) = x + y$. Suppose that (X, Y) has a continuous distribution with PDF f, and that X takes values in $S \subseteq \mathbb{R}$ and Y takes values in $T \subseteq \mathbb{R}$. Recall that X has PDF g given by $g(x) = \int_T f(x, y)\, dy$ for $x \in S$ and Y has PDF h given by $h(y) = \int_S f(x, y)\, dx$ for $y \in T$. Thus

$$\begin{aligned} \mathbb{E}(X+Y) &= \int_{S\times T} (x+y) f(x,y)\, d(x,y) = \int_{S\times T} x f(x,y)\, d(x,y) + \int_{S\times T} y f(x,y)\, d(x,y) && (4.1.8)\\ &= \int_S x \left(\int_T f(x,y)\, dy \right) dx + \int_T y \left(\int_S f(x,y)\, dx \right) dy = \int_S x g(x)\, dx + \int_T y h(y)\, dy = \mathbb{E}(X) + \mathbb{E}(Y) && (4.1.9) \end{aligned}$$

Writing the double integrals as iterated integrals is a special case of Fubini's theorem. The proof in the discrete case is the same, with sums replacing integrals.

Our next property is the *scaling property*.

$\mathbb{E}(cX) = c\,\mathbb{E}(X)$

Proof

We apply the change of variables formula with the function $r(x) = cx$. Suppose that X has a continuous distribution on $S \subseteq \mathbb{R}$ with PDF f. Then

$$\mathbb{E}(cX) = \int_S c\, x f(x)\, dx = c \int_S x f(x)\, dx = c\mathbb{E}(X) \tag{4.1.10}$$

Again, the proof in the discrete case is the same, with sums replacing integrals.

Here is the linearity of expected value in full generality. It's a simple corollary of the previous two results.

Suppose that $(X_1, X_2, \ldots)$ is a sequence of real-valued random variables defined on the underlying probability space and that $(a_1, a_2, \ldots, a_n)$ is a sequence of constants. Then

$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}(X_i) \tag{4.1.11}$$

Thus, expected value is a *linear operation* on the collection of real-valued random variables for the experiment. The linearity of expected value is so basic that it is important to understand this property on an intuitive level. Indeed, it is implied by the interpretation of expected value given in the law of large numbers.

Suppose that $(X_1, X_2, \ldots, X_n)$ is a sequence of real-valued random variables with common mean μ.

1. Let $Y = \sum_{i=1}^n X_i$, the sum of the variables. Then $\mathbb{E}(Y) = n\mu$.
2. Let $M = \frac{1}{n} \sum_{i=1}^n X_i$, the average of the variables. Then $\mathbb{E}(M) = \mu$.

Proof

1. By the additive property,

$$\mathbb{E}(Y) = \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}(X_i) = \sum_{i=1}^n \mu = n\mu \tag{4.1.12}$$

2. Note that $M = Y/n$. Hence from the scaling property and part (a), $\mathbb{E}(M) = \mathbb{E}(Y)/n = \mu$.

If the random variables in the previous result are also independent and identically distributed, then in statistical terms, the sequence is a *random sample* of size n from the common distribution, and M is the *sample mean*.

In several important cases, a random variable from a special distribution can be decomposed into a sum of simpler random variables, and then part (a) of the last theorem can be used to compute the expected value.

Inequalities

The following exercises give some basic inequalities for expected value. The first, known as the *positive property* is the most obvious, but is also the main tool for proving the others.

Suppose that $\mathbb{P}(X \ge 0) = 1$. Then

1. $\mathbb{E}(X) \ge 0$
2. If $\mathbb{P}(X > 0) > 0$ then $\mathbb{E}(X) > 0$.

Proof

1. This result follows from the definition, since we can take the set of values S of X to be a subset of $[0, \infty)$.
2. Suppose that $\mathbb{P}(X > 0) > 0$ (in addition to $\mathbb{P}(X \ge 0) = 1$). By the continuity theorem for increasing events, there exists $\epsilon > 0$ such that $\mathbb{P}(X \ge \epsilon) > 0$. Therefore $X - \epsilon \mathbf{1}(X \ge \epsilon) \ge 0$ (with probability 1). By part (a), linearity, and Theorem 2, $\mathbb{E}(X) - \epsilon \mathbb{P}(X \ge \epsilon) > 0$ so $\mathbb{E}(X) \ge \epsilon \mathbb{P}(X \ge \epsilon) > 0$.

Next is the *increasing property*, perhaps the most important property of expected value, after linearity.

Suppose that $\mathbb{P}(X \le Y) = 1$. Then

1. $\mathbb{E}(X) \le \mathbb{E}(Y)$
2. If $\mathbb{P}(X < Y) > 0$ then $\mathbb{E}(X) < \mathbb{E}(Y)$.

Proof

1. The assumption is equivalent to $\mathbb{P}(Y - X \ge 0) = 1$. Thus $\mathbb{E}(Y - X) \ge 0$ by part (a) of the positive property. But then $\mathbb{E}(Y) - \mathbb{E}(X) \ge 0$ by the linearity of expected value.
2. Similarly, this result follows from part (b) of the positive property.

Absolute value inequalities:

1. $|\mathbb{E}(X)| \le \mathbb{E}(|X|)$
2. If $\mathbb{P}(X > 0) > 0$ and $\mathbb{P}(X < 0) > 0$ then $|\mathbb{E}(X)| < \mathbb{E}(|X|)$.

Proof

1. Note that $-|X| \le X \le |X|$ (with probability 1) so by part (a) of the increasing property, $\mathbb{E}(-|X|) \le \mathbb{E}(X) \le \mathbb{E}(|X|)$. By linearity, $-\mathbb{E}(|X|) \le \mathbb{E}(X) \le \mathbb{E}(|X|)$ which implies $|\mathbb{E}(X)| \le \mathbb{E}(|X|)$.
2. If $\mathbb{P}(X > 0) > 0$ then $\mathbb{P}(-|X| < X) > 0$, and if $\mathbb{P}(X < 0) > 0$ then $\mathbb{P}(X < |X|) > 0$. Hence by part (b) of the increasing property, $-\mathbb{E}(|X|) < \mathbb{E}(X) < \mathbb{E}(|X|)$ and therefore $|\mathbb{E}(X)| < \mathbb{E}(|X|)$.

Only in Lake Woebegone are all of the children above average:

If $\mathbb{P}[X \ne \mathbb{E}(X)] > 0$ then

1. $\mathbb{P}[X > \mathbb{E}(X)] > 0$
2. $\mathbb{P}[X < \mathbb{E}(X)] > 0$

Proof

1. We prove the contrapositive. Thus suppose that $\mathbb{P}[X > \mathbb{E}(X)] = 0$ so that $\mathbb{P}[X \le \mathbb{E}(X)] = 1$. If $\mathbb{P}[X < \mathbb{E}(X)] > 0$ then by the increasing property we have $\mathbb{E}(X) < \mathbb{E}(X)$, a contradiction. Thus $\mathbb{P}[X = \mathbb{E}(X)] = 1$.
2. Similarly, if $\mathbb{P}[X < \mathbb{E}(X)] = 0$ then $\mathbb{P}[X = \mathbb{E}(X)] = 1$.

Thus, if X is not a constant (with probability 1), then X must take values greater than its mean with positive probability and values less than its mean with positive probability.

Symmetry

Again, suppose that X is a random variable taking values in $\mathbb{R}$. The distribution of X is *symmetric* about $a \in \mathbb{R}$ if the distribution of $a - X$ is the same as the distribution of $X - a$.

Suppose that the distribution of X is symmetric about $a \in \mathbb{R}$. If $\mathbb{E}(X)$ exists, then $\mathbb{E}(X) = a$.

Proof

By assumption, the distribution of $X - a$ is the same as the distribution of $a - X$. Since $\mathbb{E}(X)$ exists we have $\mathbb{E}(a - X) = \mathbb{E}(X - a)$ so by linearity $a - \mathbb{E}(X) = \mathbb{E}(X) - a$. Equivalently $2\mathbb{E}(X) = 2a$.

The previous result applies if X has a continuous distribution on $\mathbb{R}$ with a probability density f that is symmetric about a; that is, $f(a + x) = f(a - x)$ for $x \in \mathbb{R}$.

Independence

If X and Y are independent real-valued random variables then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Proof

Suppose that X has a continuous distribution on $S \subseteq \mathbb{R}$ with PDF g and that Y has a continuous distribution on $T \subseteq \mathbb{R}$ with PDF h. Then (X, Y) has PDF $f(x, y) = g(x)h(y)$ on $S \times T$. We apply the change of variables theorem with the function $r(x, y) = xy$.

$$\mathbb{E}(XY) = \int_{S \times T} xy f(x, y)\, d(x, y) = \int_{S \times T} xy g(x) h(y)\, d(x, y) = \int_S x g(x)\, dx \int_T y h(y)\, dy = \mathbb{E}(X)\mathbb{E}(Y) \tag{4.1.13}$$

The proof in the discrete case is similar with sums replacing integrals.

It follows from the last result that independent random variables are uncorrelated (a concept that we will study in a later section). Moreover, this result is more powerful than might first appear. Suppose that X and Y are independent random variables taking values in general spaces S and T respectively, and that $u : S \to \mathbb{R}$ and $v : T \to \mathbb{R}$. Then $u(X)$ and $v(Y)$ are independent, real-valued random variables and hence

$$\mathbb{E}\left[u(X)v(Y)\right] = \mathbb{E}\left[u(X)\right]\mathbb{E}\left[v(Y)\right] \tag{4.1.14}$$

Examples and Applications

As always, be sure to try the proofs and computations yourself before reading the proof and answers in the text.

Uniform Distributions

Discrete uniform distributions are widely used in combinatorial probability, and model a point chosen *at random* from a finite set.

Suppose that X has the discrete uniform distribution on a finite set $S \subseteq \mathbb{R}$.

1. $\mathbb{E}(X)$ is the arithmetic average of the numbers in S.
2. If the points in S are evenly spaced with endpoints a, b, then $\mathbb{E}(X) = \frac{a+b}{2}$, the average of the endpoints.

Proof

1. Let $n = \#(S)$, the number of points in S. Then X has PDF $f(x) = 1/n$ for $x \in S$ so
$$\mathbb{E}(X) = \sum_{x \in S} x \frac{1}{n} = \frac{1}{n} \sum_{x \in S} x \tag{4.1.15}$$
2. Suppose that $S = \{a, a + h, a + 2h, \ldots a + (n-1)h\}$ and let $b = a + (n-1)h$, the right endpoint. As in (a), S has n points so using (a) and the formula for the sum of the first $n - 1$ positive integers, we have
$$\mathbb{E}(X) = \frac{1}{n} \sum_{i=0}^{n-1} (a + ih) = \frac{1}{n}\left(na + h\frac{(n-1)n}{2}\right) = a + \frac{(n-1)h}{2} = \frac{a+b}{2} \tag{4.1.16}$$

The previous results are easy to see if we think of $\mathbb{E}(X)$ as the center of mass, since the discrete uniform distribution corresponds to a finite set of points with equal mass.

Open the special distribution simulator, and select the discrete uniform distribution. This is the uniform distribution on n points, starting at a, evenly spaced at distance h. Vary the parameters and note the location of the mean in relation to the probability density function. For selected values of the parameters, run the simulation 1000 times and compare the empirical mean to the distribution mean.

Next, recall that the continuous uniform distribution on a bounded interval corresponds to selecting a point *at random* from the interval. Continuous uniform distributions arise in geometric probability and a variety of other applied problems.

Suppose that X has the continuous uniform distribution on an interval $[a, b]$, where $a,\ b \in \mathbb{R}$ and $a < b$.

1. $\mathbb{E}(X) = \frac{a+b}{2}$, the midpoint of the interval.
2. $E\left(X^n\right) = \frac{1}{n+1}\left(a^n + a^{n-1}b + \cdots + ab^{n-1} + b^n\right) \quad$ for $n \in \mathbb{N}$.

Proof

1. Recall that X has PDF $f(x) = \frac{1}{b-a}$. Hence

$$\mathbb{E}(X) = \int_a^b x \frac{1}{b-a}\, dx = \frac{1}{b-a} \frac{b^2 - a^2}{2} = \frac{a+b}{2} \tag{4.1.17}$$

2. By the change of variables formula,

$$\mathbb{E}\left(X^n\right) = \int_a^b \frac{1}{b-a} x^n\, dx = \frac{b^{n+1} - a^{n+1}}{(n+1)`(b-a)} = \frac{1}{n+1}\left(a^n + a^{n-1}b + \cdots ab^{n-1} + b^n\right) \tag{4.1.18}$$

Part (a) is easy to see if we think of the mean as the center of mass, since the uniform distribution corresponds to a uniform distribution of mass on the interval.

Open the special distribution simulator, and select the continuous uniform distribution. This is the uniform distribution the interval $[a, a+w]$. Vary the parameters and note the location of the mean in relation to the probability density function. For selected values of the parameters, run the simulation 1000 times and compare the empirical mean to the distribution mean.

Next, the *average value* of a function on an interval, as defined in calculus, has a nice interpretation in terms of the uniform distribution.

Suppose that X is uniformly distributed on the interval $[a, b]$, and that g is an integrable function from $[a, b]$ into $\mathbb{R}$. Then $\mathbb{E}[g(X)]$ is the average value of g on $[a, b]$:

$$\mathbb{E}[g(X)] = \frac{1}{b-a}\int_a^b g(x) dx \tag{4.1.19}$$

Proof

This result follows immediately from the change of variables theorem, since X has PDF $f(x) = 1/(b-a)$ for $a \le x \le b$.

Find the average value of the following functions on the given intervals:

1. $f(x) = x$ on $[2, 4]$
2. $g(x) = x^2$ on $[0, 1]$
3. $h(x) = \sin(x)$ on $[0, \pi]$.

Answer

1. 3
2. $\frac{1}{3}$
3. $\frac{2}{\pi}$

The next exercise illustrates the value of the change of variables theorem in computing expected values.

Suppose that X is uniformly distributed on $[-1, 3]$.

1. Give the probability density function of X.
2. Find the probability density function of X^2.
3. Find $E\left(X^2\right)$ using the probability density function in (b).
4. Find $\mathbb{E}\left(X^2\right)$ using the change of variables theorem.

Answer

1. $f(x) = \frac{1}{4}$ for $-1 \le x \le 3$
2. $g(y) = \begin{cases} \frac{1}{4} y^{-1/2}, & 0 < y < 1 \\ \frac{1}{8} y^{-1/2}, & 1 < y < 9 \end{cases}$

3. $\int_0^9 y g(y)\,dy = \frac{7}{3}$
4. $\int_{-1}^3 x^2 f(x)\,dx = \frac{7}{3}$

The discrete uniform distribution and the continuous uniform distribution are studied in more detail in the chapter on Special Distributions.

Dice

Recall that a *standard die* is a six-sided die. A *fair* die is one in which the faces are equally likely. An *ace-six flat die* is a standard die in which faces 1 and 6 have probability $\frac{1}{4}$ each, and faces 2, 3, 4, and 5 have probability $\frac{1}{8}$ each.

Two standard, fair dice are thrown, and the scores (X_1, X_2) recorded. Find the expected value of each of the following variables.

1. $Y = X_1 + X_2$, the sum of the scores.
2. $M = \frac{1}{2}(X_1 + X_2)$, the average of the scores.
3. $Z = X_1 X_2$, the product of the scores.
4. $U = \min\{X_1, X_2\}$, the minimum score
5. $V = \max\{X_1, X_2\}$, the maximum score.

Answer

1. 7
2. $\frac{7}{2}$
3. $\frac{49}{4}$
4. $\frac{101}{36}$
5. $\frac{19}{4}$

In the dice experiment, select two fair die. Note the shape of the probability density function and the location of the mean for the sum, minimum, and maximum variables. Run the experiment 1000 times and compare the sample mean and the distribution mean for each of these variables.

Two standard, ace-six flat dice are thrown, and the scores (X_1, X_2) recorded. Find the expected value of each of the following variables.

1. $Y = X_1 + X_2$, the sum of the scores.
2. $M = \frac{1}{2}(X_1 + X_2)$, the average of the scores.
3. $Z = X_1 X_2$, the product of the scores.
4. $U = \min\{X_1, X_2\}$, the minimum score
5. $V = \max\{X_1, X_2\}$, the maximum score.

Answer

1. 7
2. $\frac{7}{2}$
3. $\frac{49}{4}$
4. $\frac{77}{32}$
5. $\frac{147}{32}$

In the dice experiment, select two ace-six flat die. Note the shape of the probability density function and the location of the mean for the sum, minimum, and maximum variables. Run the experiment 1000 times and compare the sample mean and the distribution mean for each of these variables.

Bernoulli Trials

Recall that a *Bernoulli trials process* is a sequence $\boldsymbol{X} = (X_1, X_2, \ldots)$ of independent, identically distributed indicator random variables. In the usual language of reliability, X_i denotes the outcome of trial i, where 1 denotes success and 0 denotes failure. The probability of success $p = \mathbb{P}(X_i = 1) \in [0, 1]$ is the basic parameter of the process. The process is named for Jacob Bernoulli. A separate chapter on the Bernoulli Trials explores this process in detail.

For $n \in \mathbb{N}_+$, the number of successes in the first n trials is $Y = \sum_{i=1}^n X_i$. Recall that this random variable has the binomial distribution with parameters n and p, and has probability density function f given by

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, \ldots, n\} \tag{4.1.20}$$

If Y has the binomial distribution with parameters n and p then $\mathbb{E}(Y) = np$

Proof from the definition

The critical tools that we need involve binomial coefficients: the identity $y\binom{n}{y} = n\binom{n-1}{y-1}$ for $y,\ n \in \mathbb{N}_+$, and the binomial theorem:

$$\mathbb{E}(Y) = \sum_{y=0}^{n} y \binom{n}{y} p^y (1-p)^{n-y} = \sum_{y=1}^{n} n \binom{n-1}{y-1} p^n (1-p)^{n-y} \tag{4.1.21}$$

$$= np \sum_{y=1}^{n-1} \binom{n-1}{y-1} p^{y-1} (1-p)^{(n-1)-(y-1)} = np[p + (1-p)]^{n-1} = np \tag{4.1.22}$$

Proof using the additive property

Since $Y = \sum_{i=1}^n X_i$, the result follows immediately from the expected value of an indicator variable and the additive property, since $\mathbb{E}(X_i) = p$ for each $i \in \mathbb{N}_+$.

Note the superiority of the second proof to the first. The result also makes intuitive sense: in n trials with success probability p, we expect np successes.

In the binomial coin experiment, vary n and p and note the shape of the probability density function and the location of the mean. For selected values of n and p, run the experiment 1000 times and compare the sample mean to the distribution mean.

Suppose that $p \in (0, 1]$, and let N denote the trial number of the first success. This random variable has the geometric distribution on $\mathbb{N}_+$ with parameter p, and has probability density function g given by

$$g(n) = p(1-p)^{n-1}, \quad n \in \mathbb{N}_+ \tag{4.1.23}$$

If N has the geometric distribution on $\mathbb{N}_+$ with parameter $p \in (0, 1]$ then $\mathbb{E}(N) = 1/p$.

Proof

The key is the formula for the deriviative of a geometric series:

$$\mathbb{E}(N) = \sum_{n=1}^{\infty} np(1-p)^{n-1} = -p \frac{d}{dp} \sum_{n=0}^{\infty} (1-p)^n = -p \frac{d}{dp} \frac{1}{p} = p \frac{1}{p^2} = \frac{1}{p} \tag{4.1.24}$$

Again, the result makes intuitive sense. Since p is the probability of success, we expect a success to occur after $1/p$ trials.

In the negative binomial experiment, select $k = 1$ to get the geometric distribution. Vary p and note the shape of the probability density function and the location of the mean. For selected values of p, run the experiment 1000 times and compare the sample mean to the distribution mean.

The Hypergeometric Distribution

Suppose that a population consists of m objects; r of the objects are type 1 and $m - r$ are type 0. A sample of n objects is chosen at random, without replacement. The parameters $m,\ r,\ n \in \mathbb{N}$ with $r \le m$ and $n \le m$. Let X_i denote the type of the ith object selected. Recall that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a sequence of identically distributed (but *not* independent) indicator random variable with $\mathbb{P}(X_i = 1) = r/m$ for each $i \in \{1, 2, \ldots, n\}$.

Let Y denote the number of type 1 objects in the sample, so that $Y = \sum_{i=1}^n X_i$. Recall that Y has the hypergeometric distribution, which has probability density function f given by

$$f(y) = \frac{\binom{r}{y}\binom{m-r}{n-y}}{\binom{m}{n}}, \quad y \in \{0, 1, \ldots, n\} \tag{4.1.25}$$

If Y has the hypergeometric distribution with parameters m, n, and r then $\mathbb{E}(Y) = n\frac{r}{m}$.

Proof from the definition

Using the hypergeometric PDF,

$$\mathbb{E}(Y)=\sum_{y=0}^{n} y \frac{\binom{r}{y}\binom{m-r}{n-y}}{\binom{m}{n}} \tag{4.1.26}$$

Note that the $y=0$ term is 0. For the other terms, we can use the identity $y\binom{r}{y}=r\binom{r-1}{y-1}$ to get

$$\mathbb{E}(Y)=\frac{r}{\binom{m}{n}} \sum_{y=1}^{n}\binom{r-1}{y-1}\binom{m-r}{n-y} \tag{4.1.27}$$

But substituting $k=y-1$ and using another fundamental identity,

$$\sum_{y=1}^{n}\binom{r-1}{y-1}\binom{m-r}{n-y}=\sum_{k=0}^{n-1}\binom{r-1}{k}\binom{m-r}{n-1-k}=\binom{m-1}{n-1} \tag{4.1.28}$$

So substituting and doing a bit of algebra gives $\mathbb{E}(Y)=n \frac{r}{m}$.

Proof using the additive property

A much better proof uses the additive property and the representation of Y as a sum of indicator variables. The result follows immediately since $\mathbb{E}(X_i)=r/m$ for each $i \in \{1,2,\ldots n\}$.

In the ball and urn experiment, vary n, r, and m and note the shape of the probability density function and the location of the mean. For selected values of the parameters, run the experiment 1000 times and compare the sample mean to the distribution mean.

Note that if we select the objects *with replacement*, then $\boldsymbol{X}$ would be a sequence of Bernoulli trials, and hence Y would have the binomial distribution with parameters n and $p=\frac{r}{m}$. Thus, the mean would still be $\mathbb{E}(Y)=n\frac{r}{m}$.

The Poisson Distribution

Recall that the *Poisson distribution* has probability density function f given by

$$f(n)=e^{-a} \frac{a^n}{n!}, \quad n \in \mathbb{N} \tag{4.1.29}$$

where $a \in (0,\infty)$ is a parameter. The Poisson distribution is named after Simeon Poisson and is widely used to model the number of “random points” in a region of time or space; the parameter a is proportional to the size of the region. The Poisson distribution is studied in detail in the chapter on the Poisson Process.

If N has the Poisson distribution with parameter a then $\mathbb{E}(N)=a$. Thus, the parameter of the Poisson distribution is the mean of the distribution.

Proof

The proof depends on the standard series for the exponential function

$$\mathbb{E}(N)=\sum_{n=0}^{\infty} n e^{-a} \frac{a^n}{n!}=e^{-a} \sum_{n=1}^{\infty} \frac{a^n}{(n-1)!}=e^{-a} a \sum_{n=1}^{\infty} \frac{a^{n-1}}{(n-1)!}=e^{-a} a e^{a}=a. \tag{4.1.30}$$

In the Poisson experiment, the parameter is $a=rt$. Vary the parameter and note the shape of the probability density function and the location of the mean. For various values of the parameter, run the experiment 1000 times and compare the sample mean to the distribution mean.

The Exponential Distribution

Recall that the *exponential distribution* is a continuous distribution with probability density function f given by

$$f(t)=r e^{-rt}, \quad t \in [0,\infty) \tag{4.1.31}$$

where $r \in (0,\infty)$ is the *rate parameter*. This distribution is widely used to model failure times and other “arrival times”; in particular, the distribution governs the time between arrivals in the Poisson model. The exponential distribution is studied in detail in the chapter on the Poisson Process.

Suppose that T has the exponential distribution with rate parameter r. Then $\mathbb{E}(T)=1/r$.

Proof

This result follows from the definition and an integration by parts:

$$\mathbb{E}(T) = \int_0^\infty t r e^{-rt}\,dt = -t e^{-rt}\Big|_0^\infty + \int_0^\infty e^{-rt}\,dt = 0 - \frac{1}{r} e^{-rt}\Big|_0^\infty = \frac{1}{r} \tag{4.1.32}$$

Recall that the mode of T is 0 and the median of T is $\ln 2/r$. Note how these measures of center are ordered: $0 < \ln 2/r < 1/r$

In the gamma experiment, set $n = 1$ to get the exponential distribution. This app simulates the first arrival in a Poisson process. Vary r with the scroll bar and note the position of the mean relative to the graph of the probability density function. For selected values of r, run the experiment 1000 times and compare the sample mean to the distribution mean.

Suppose again that T has the exponential distribution with rate parameter r and suppose that $t > 0$. Find $\mathbb{E}(T \mid T > t)$.

Answer

$t + \frac{1}{r}$

The Gamma Distribution

Recall that the *gamma distribution* is a continuous distribution with probability density function f given by

$$f(t) = r^n \frac{t^{n-1}}{(n-1)!} e^{-rt}, \quad t \in [0, \infty) \tag{4.1.33}$$

where $n \in N_+$ is the *shape parameter* and $r \in (0, \infty)$ is the *rate parameter*. This distribution is widely used to model failure times and other "arrival times", and in particular, models the nth arrival in the Poisson process. Thus it follows that if $(X_1, X_2, \ldots, X_n)$ is a sequence of independent random variables, each having the exponential distribution with rate parameter r, then $T = \sum_{i=1}^n X_i$ has the gamma distribution with shape parameter n and rate parameter r. The gamma distribution is studied in more generality, with non-integer shape parameters, in the chapter on the Special Distributions.

Suppose that T has the gamma distribution with shape parameter n and rate parameter r. Then $\mathbb{E}(T) = n/r$.

Proof from the definition

The proof is by induction on n, so let μ_n denote the mean when the shape parameter is $n \in \mathbb{N}_+$. When $n = 1$, we have the exponential distribution with rate parameter r, so we know $\mu_1 = 1/r$ by our result above. Suppose that $\mu_n = r/n$ for a given $n \in \mathbb{N}_+$. Then

$$\mu_{n+1} = \int_0^\infty t r^{n+1} \frac{t^n}{n!} e^{-rt}\,dt = \int_0^\infty r^{n+1} \frac{t^{n+1}}{n!} e^{-rt}\,dt \tag{4.1.34}$$

Integrate by parts with $u = \frac{t^{n+1}}{n!}$, $dv = r^{n+1} e^{-rt}\,dt$ so that $du = (n+1)\frac{t^n}{n!}\,dt$ and $v = -r^n e^{-rt}$. Then

$$\mu_{n+1} = (n+1) \int_0^\infty r^n \frac{t^n}{n!} e^{-rt}\,dt = \frac{n+1}{n} \int_0^\infty t r^n \frac{t^{n-1}}{(n-1)!}\,dt \tag{4.1.35}$$

But the last integral is μ_n, so by the induction hypothesis, $\mu_{n+1} = \frac{n+1}{n}\frac{n}{r} = \frac{n+1}{r}$.

Proof using the additive property

The result follows immediately from the additive property and the fact that T can be represented in the form $T = \sum_{i=1}^n X_i$ where X_i has the exponential distribution with parameter r for each $i \in \{1, 2, \ldots, n\}$.

Note again how much easier and more intuitive the second proof is than the first.

Open the gamma experiment, which simulates the arrival times in the Poisson process. Vary the parameters and note the position of the mean relative to the graph of the probability density function. For selected parameter values, run the experiment 1000 times and compare the sample mean to the distribution mean.

Beta Distributions

The distributions in this subsection belong to the family of *beta distributions*, which are widely used to model random proportions and probabilities. The beta distribution is studied in detail in the chapter on Special Distributions.

Suppose that X has probability density function f given by $f(x) = 3x^2$ for $x \in [0, 1]$.

1. Find the mean of X.

2. Find the mode of X.
3. Find the median of X.
4. Sketch the graph of f and show the location of the mean, median, and mode on the x-axis.

Answer

1. $\frac{3}{4}$
2. 1
3. $\left(\frac{1}{2}\right)^{1/3}$

In the special distribution simulator, select the beta distribution and set $a = 3$ and $b = 1$ to get the distribution in the last exercise. Run the experiment 1000 times and compare the sample mean to the distribution mean.

Suppose that a sphere has a random radius R with probability density function f given by $f(r) = 12r^2(1-r)$ for $r \in [0, 1]$. Find the expected value of each of the following:

1. The circumference $C = 2\pi R$
2. The surface area $A = 4\pi R^2$
3. The volume $V = \frac{4}{3}\pi R^3$

Answer

1. $\frac{6}{5}\pi$
2. $\frac{8}{5}\pi$
3. $\frac{8}{21}\pi$

Suppose that X has probability density function f given by $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$ for $x \in (0, 1)$.

1. Find the mean of X.
2. Find median of X.
3. Note that f is unbounded, so X does not have a mode.
4. Sketch the graph of f and show the location of the mean and median on the x-axis.

Answer

1. $\frac{1}{2}$
2. $\frac{1}{2}$

The particular beta distribution in the last exercise is also known as the (standard) *arcsine distribution*. It governs the last time that the Brownian motion process hits 0 during the time interval $[0, 1]$. The arcsine distribution is studied in more generality in the chapter on Special Distributions.

Open the Brownian motion experiment and select the last zero. Run the simulation 1000 times and compare the sample mean to the distribution mean.

Suppose that the grades on a test are described by the random variable $Y = 100X$ where X has the beta distribution with probability density function f given by $f(x) = 12x(1-x)^2$ for $x \in [0, 1]$. The grades are generally low, so the teacher decides to "curve" the grades using the transformation $Z = 10\sqrt{Y} = 100\sqrt{X}$. Find the expected value of each of the following variables

1. X
2. Y
3. Z

Answer

1. $\mathbb{E}(X) = \frac{2}{5}$
2. $\mathbb{E}(Y) = 40$
3. $\mathbb{E}(Z) = \frac{1280}{21} \approx 60.95$

The Pareto Distribution

Recall that the *Pareto distribution* is a continuous distribution with probability density function f given by

$$f(x) = \frac{a}{x^{a+1}}, \quad x \in [1, \infty) \tag{4.1.36}$$

where $a \in (0, \infty)$ is a parameter. The Pareto distribution is named for Vilfredo Pareto. It is a heavy-tailed distribution that is widely used to model certain financial variables. The Pareto distribution is studied in detail in the chapter on Special Distributions.

Suppose that X has the Pareto distribution with shape parameter a. Then

1. $\mathbb{E}(X) = \infty$ if $0 < a \le 1$
2. $\mathbb{E}(X) = \frac{a}{a-1}$ if $a > 1$

Proof

1. If $0 < a < 1$,

$$\mathbb{E}(X) = \int_1^\infty x \frac{a}{x^{a+1}} \, dx = \int_1^\infty \frac{a}{x^a} \, dx = \frac{a}{-a+1} x^{-a+1} \bigg|_1^\infty = \infty \tag{4.1.37}$$

since the exponent $-a+1 > 0$. If $a = 1$, $\mathbb{E}(X) = \int_1^\infty x \frac{1}{x^2} \, dx = \int_1^\infty \frac{1}{x} \, dx = \ln x \bigg|_1^\infty = \infty$.

2. If $a > 1$ then

$$\mathbb{E}(X) = \int_1^\infty x \frac{a}{x^{a+1}} \, dx = \int_1^\infty \frac{a}{x^a} \, dx = \frac{a}{-a+1} x^{-a+1} \bigg|_1^\infty = \frac{a}{a-1} \tag{4.1.38}$$

The previous exercise gives us our first example of a distribution whose mean is infinite.

In the special distribution simulator, select the Pareto distribution. Note the shape of the probability density function and the location of the mean. For the following values of the shape parameter a, run the experiment 1000 times and note the behavior of the empirical mean.

1. $a = 1$
2. $a = 2$
3. $a = 3$.

The Cauchy Distribution

Recall that the (standard) *Cauchy distribution* has probability density function f given by

$$f(x) = \frac{1}{\pi \left(1 + x^2\right)}, \quad x \in \mathbb{R} \tag{4.1.39}$$

This distribution is named for Augustin Cauchy. The Cauchy distributions is studied in detail in the chapter on Special Distributions.

If X has the Cauchy distribution then $\mathbb{E}(X)$ does not exist.

Proof

By definition,

$$\mathbb{E}(X) = \int_{-\infty}^\infty x \frac{1}{\pi(1+x^2)} \, dx = \frac{1}{2\pi} \ln\left(1 + x^2\right) \bigg|_{-\infty}^\infty \tag{4.1.40}$$

which evaluates to the meaningless expression $\infty - \infty$.

Note that the graph of f is symmetric about 0 and is unimodal. Thus, the mode and median of X are both 0. By the symmetry result, if X *had* a mean, the mean would be 0 also, but alas the mean does not exist. Moreover, the non-existence of the mean is not just a pedantic technicality. If we think of the probability distribution as a mass distribution, then the moment to the right of a is $\int_a^\infty (x - a) f(x) \, dx = \infty$ and the moment to the left of a is $\int_{-\infty}^a (x - a) f(x) \, dx = -\infty$ for every $a \in \mathbb{R}$. The center of mass simply does not exist. Probabilisitically, the law of large numbers fails, as you can see in the following simulation exercise:

In the Cauchy experiment (with the default parameter values), a light sources is 1 unit from position 0 on an infinite straight wall. The angle that the light makes with the perpendicular is uniformly distributed on the interval $\left(\frac{-\pi}{2}, \frac{\pi}{2}\right)$, so that the position of the light beam on the wall has the Cauchy distribution. Run the simulation 1000 times and note the behavior of the empirical mean.

The Normal Distribution

Recall that the *standard normal distribution* is a continuous distribution with density function ϕ given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad z \in \mathbb{R} \tag{4.1.41}$$

Normal distributions are widely used to model physical measurements subject to small, random errors and are studied in detail in the chapter on Special Distributions.

If Z has the standard normal distribution then $\mathbb{E}(X) = 0$.

Proof

Using a simple change of variables, we have

$$\mathbb{E}(Z) = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \, dz = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \Big|_{-\infty}^{\infty} = 0 - 0 \tag{4.1.42}$$

The standard normal distribution is unimodal and symmetric about 0. Thus, the median, mean, and mode all agree. More generally, for $\mu \in (-\infty, \infty)$ and $\sigma \in (0, \infty)$, recall that $X = \mu + \sigma Z$ has the normal distribution with *location parameter* μ and *scale parameter* σ. X has probability density function f given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad x \in \mathbb{R} \tag{4.1.43}$$

The location parameter is the mean of the distribution:

If X has the normal distribution with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma \in (0, \infty)$, then $\mathbb{E}(X) = \mu$

Proof

Of course we could use the definition, but a proof using linearity and the representation in terms of the standard normal distribution is trivial: $\mathbb{E}(X) = \mu + \sigma \mathbb{E}(Z) = \mu$.

In the special distribution simulator, select the normal distribution. Vary the parameters and note the location of the mean. For selected parameter values, run the simulation 1000 times and compare the sample mean to the distribution mean.

Additional Exercises

Suppose that (X, Y) has probability density function f given by $f(x, y) = x + y$ for $(x, y) \in [0, 1] \times [0, 1]$. Find the following expected values:

1. $\mathbb{E}(X)$
2. $\mathbb{E}(X^2 Y)$
3. $\mathbb{E}(X^2 + Y^2)$
4. $\mathbb{E}(XY \mid Y > X)$

Answer

1. $\frac{7}{12}$
2. $\frac{17}{72}$
3. $\frac{5}{6}$
4. $\frac{1}{3}$

Suppose that N has a discrete distribution with probability density function f given by $f(n) = \frac{1}{50} n^2 (5 - n)$ for $n \in \{1, 2, 3, 4\}$. Find each of the following:

1. The median of N.
2. The mode of N
3. $\mathbb{E}(N)$.
4. $\mathbb{E}(N^2)$
5. $\mathbb{E}(1/N)$.
6. $\mathbb{E}(1/N^2)$.

Answer

1. 3
2. 3
3. $\frac{73}{25}$
4. $\frac{47}{5}$
5. $\frac{2}{5}$
6. $\frac{1}{5}$

Suppose that X and Y are real-valued random variables with $\mathbb{E}(X) = 5$ and $\mathbb{E}(Y) = -2$. Find $\mathbb{E}(3X + 4Y - 7)$.

Answer

0

Suppose that X and Y are real-valued, independent random variables, and that $\mathbb{E}(X) = 5$ and $\mathbb{E}(Y) = -2$. Find $\mathbb{E}\left[(3X-4)(2Y+7)\right]$.

Answer

33

Suppose that there are 5 duck hunters, each a perfect shot. A flock of 10 ducks fly over, and each hunter selects one duck at random and shoots. Find the expected number of ducks killed.

Solution

Number the ducks from 1 to 10. For $k \in \{1, 2, \ldots, 10\}$ let X_k be the indicator variable that takes the value 1 if duck k is killed and 0 otherwise. Duck k is killed if at least one of the hunters selects her, so $\mathbb{E}(X_k) = \mathbb{P}(X_k = 1) = 1 - \left(\frac{9}{10}\right)^5$. The number of ducks killed is $N = \sum_{k=1}^{10} X_k$ so $\mathbb{E}(N) = 10\left[1 - \left(\frac{9}{10}\right)^5\right] = 4.095$

For a more complete analysis of the duck hunter problem, see The Number of Distinct Sample Values in the chapter on Finite Sampling Models.

Consider the following game: An urn initially contains one red and one green ball. A ball is selected at random, and if the ball is green, the game is over. If the ball is red, the ball is returned to the urn, another red ball is added, and the game continues. At each stage, a ball is selected at random, and if the ball is green, the game is over. If the ball is red, the ball is returned to the urn, another red ball is added, and the game continues. Let X denote the length of the game (that is, the number of selections required to obtain a green ball). Find $\mathbb{E}(X)$.

Solution

The probability density function f of X was found in the section on discrete distributions: $f(x) = \frac{1}{x(x+1)}$ for $x \in \mathbb{N}_+$. The expected length of the game is infinite:

$$\mathbb{E}(X) = \sum_{x=1}^{\infty} x \frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{x+1} = \infty \tag{4.1.44}$$