1. Definitions and 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 with probability measure $\mathbb{P}$ on an underlying sample space Ω. Suppose 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 by

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

assuming that the sum is absolutely convergent (that is, assuming that the sum with x replaced by $|x|$ is finite). The assumption of absolute convergence is necessary to ensure that the sum in the expected value above does not depend on the order of the terms. Of course, if S is finite there are no convergence problems.

If X has a continuous distribution with probability density function f (and so S is typically an interval), then the **expected value** of X is defined by

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

assuming that the integral is absolutely convergent (that is, assuming that the integral with x replaced by $|x|$ is finite).

Finally, suppose that 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 by

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

assuming again that the sum and integral converge absolutely.

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.
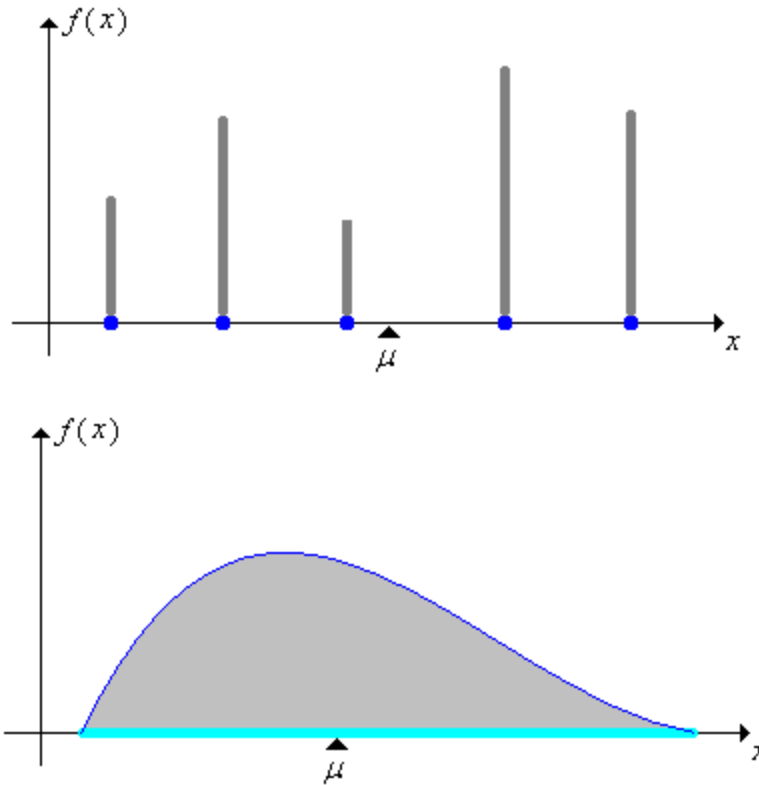
Please recall the other measures of the center of a distribution that we have studied:

- A mode is any value of x that maximizes $f(x)$.
- A median is any value of x that satisfies $\mathbb{P}(X < x) \leq \frac{1}{2}$ and $\mathbb{P}(X \leq x) \geq \frac{1}{2}$.

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 $(X_1, X_2, ...)$, 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

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

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.

Moments

If $a \in \mathbb{R}$ and $n > 0$, the **moment of X about a of order n** is defined to be

$$\mathbb{E}\big((X - a)^n\big)$$

The moments about 0 are simply referred to as **moments**. The moments about $\mu = \mathbb{E}(X)$ 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.

Conditional Expected Value

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 $B \subseteq \Omega$ for the experiment (with $\mathbb{P}(B) > 0$). The usual notation is $\mathbb{E}(X|B)$, and this expected value is computed by the definitions given above, except that the conditional probability density function $f(x|B)$ replaces the ordinary probability density function $f(x)$. 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.

Basic Results

The purpose of this section is to study some of the essential properties of expected value. Unless otherwise noted, we will assume that the indicated expected values exist.

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 $\mathbb{E}(r(X))$. 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.

1. Show that if X has a discrete distribution on a countable set S with probability density function f. then
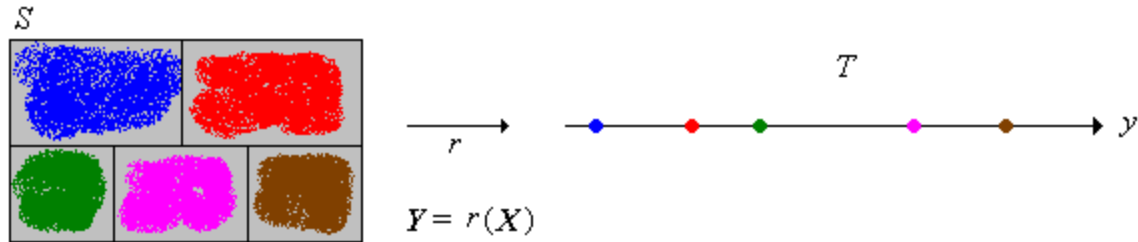$$\mathbb{E}(r(X)) = \sum_{x \in S} r(x) f(x)$$
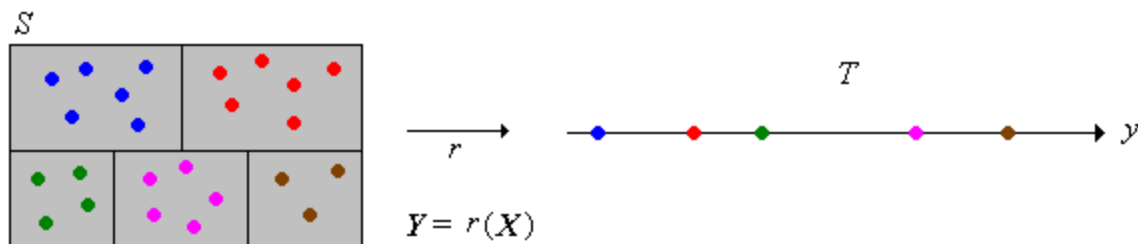
Similarly, if X has a continuous distribution on $S \subseteq \mathbb{R}^n$ with probability density function f. then

$$\mathbb{E}(r(X)) = \int_S r(x) f(x) dx$$

We will prove the continuous version in stages, through Exercise 2, Exercise 53, and Exercise 56.

2. Prove the version of the change of variables theorem when X has a continuous distribution on $S \subseteq \mathbb{R}^n$ and r is discrete (i.e., r has countable range).

The exercises below gives basic properties of expected value. These properties are true in general, but restrict your proofs to the discrete and continuous cases separately; the change of variables theorem is the main tool you will need. In these exercises X and Y are real-valued random variables for an experiment and c is a constant. We assume that the indicated expected values exist.

Linearity

3. Show that $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

4. Show that $\mathbb{E}(c X) = c \mathbb{E}(X)$

Suppose that $(X_1, X_2, ..., X_n)$ is a sequence of real-valued random variables for our experiment and that $(a_1, a_2, ..., a_n)$ is a sequence of constants. Then, as a consequence of the previous two results,

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

Thus, expected value is a **linear operation**. 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.

5. Suppose that $(X_1, X_2, ..., X_n)$ is a sequence of real-valued random variables, with common mean μ. If the random variables are also independent and identically distributed, then in statistical terms, the sequence is a **random sample** of size n from the common distribution.

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

Inequalities

The following exercises give some basic inequalities for expected value. The first is the most obvious, but is also the

main tool for proving the others.

6. Suppose that $\mathbb{P}(X \geq 0) = 1$. Prove the following results:

   a. $\mathbb{E}(X) \geq 0$
   b. $\mathbb{E}(X) = 0$ if and only if $\mathbb{P}(X = 0) = 1$

7. Suppose that $\mathbb{P}(X \leq Y) = 1$. Prove the following results:

   a. $\mathbb{E}(X) \leq \mathbb{E}(Y)$
   b. $\mathbb{E}(X) = \mathbb{E}(Y)$ if and only if $\mathbb{P}(X = Y) = 1$

Thus, expected value is an **increasing** operator. This is perhaps the second most important property of expected value, after linearity.

8. Show that

   a. $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$
   b. $|\mathbb{E}(X)| = \mathbb{E}(|X|)$ if and only if $\mathbb{P}(X \geq 0) = 1$ or $\mathbb{P}(X \leq 0) = 1$

9. Prove the following results: (Only in Lake Woebegone are all of the children above average.)

   a. $\mathbb{P}(X \geq \mathbb{E}(X)) > 0$
   b. $\mathbb{P}(X \leq \mathbb{E}(X)) > 0$

Symmetry

10. Suppose that X has a continuous distribution on $\mathbb{R}$ with a probability density f that is symmetric about a: $f(a + t) = f(a - t)$ for $t \in \mathbb{R}$. Show that if $\mathbb{E}(X)$ exists, then $\mathbb{E}(X) = a$.

Independence

11. Suppose that X and Y are independent real-valued random variables. Show that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

It follows from the last exercise that independent random variables are uncorrelated. 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, and that u and v are real-valued functions on S and T, respectively. Then $u(X)$ and $v(Y)$ are independent, real-valued random variables and hence

$$\mathbb{E}\big(u(X)v(Y)\big) = \mathbb{E}(u(X))\mathbb{E}(v(Y))$$

Examples and Applications

Constants and Indicator Variables

12. A constant c 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. Show that $\mathbb{E}(c) = c$

13. Let X be an indicator random variable (that is, a variable that takes only the values 0 and 1). Show that $\mathbb{E}(X) = \mathbb{P}(X = 1)$.

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 *Probability via Expectation*, by Peter Whittle.

Uniform Distributions

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

   a. Show that $\mathbb{E}(X)$ is the arithmetic average of the numbers in S.
   b. In particular, show that if X is uniformly distributed on the integer interval $S = \{m, m + 1, ..., n\}$ where $m \leq n$, then $\mathbb{E}(X) = \frac{m+n}{2}$.

15. Suppose that X has the continuous uniform distribution on an interval $[a, b]$.

   a. Show that the mean is the midpoint of the interval: $\mathbb{E}(X) = \frac{a+b}{2}$.
   b. Find a general formula for the moments of X.

16. Suppose that X is uniformly distributed on the interval $[a, b]$, and that g is an integrable function from $[a, b]$ into $\mathbb{R}$. Show that $\mathbb{E}(g(X))$ is the **average value** of g on $[a, b]$, as defined in calculus.

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

17. Find the average value of the sine function on the interval $[0, \pi]$.

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

   a. Find the probability density function of X^2.
   b. Find $\mathbb{E}(X^2)$ using the probability density function in (a).
   c. Find $\mathbb{E}(X^2)$ using the change of variables theorem.

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.

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

    a. $Y = X_1 + X_2$, the sum of the scores.
    b. $M = \frac{1}{2}(X_1 + X_2)$, the average of the scores.
    c. $Z = X_1 X_2$, the product of the scores.
    d. $U = \min\{X_1, X_2\}$, the minimum score
    e. $V = \max\{X_1, X_2\}$, the maximum score.

20. In the dice experiment, select two fair die. Note the shape of the density function and the location of the mean for the sum, minimum, and maximum variables. Run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the sample mean to the distribution mean for each of these variables.

21. Repeat Exercise 19 for ace-six flat dice.

22. Repeat Exercise 20 for ace-six flat dice.

Bernoulli Trials

Recall that a **Bernoulli trials process** is a sequence $(X_1, X_2, ...)$ 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 $p = \mathbb{P}(X_i = 1)$ is the basic parameter of the process. The process is named for James Bernoulli. A separate chapter on the Bernoulli Trials explores this process in detail.

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

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ \ k \in \{0, 1, ..., n\}$$

23. Show that $\mathbb{E}(Y) = n\,p$ in the following ways:

    a. From the definition
    b. From the representation as a sum of indicator variables

24. In the binomial coin experiment, vary n and p and note the shape of the density function and the location of the mean. For selected values of n and p, run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the sample mean to the distribution mean.

Now let W 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.

$$\mathbb{P}(W = n) = p\,(1-p)^{n-1}, \ \ n \in \mathbb{N}_+$$

✚ 25. Show that $\mathbb{E}(W) = \frac{1}{p}$.

✖ 26. In the negative binomial experiment, select $k = 1$ to get the geometric distribution. Vary p and note the shape of the density function and the location of the mean. For selected values of p, run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of 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. Let X_i denote the type of the ith object selected. Recall that $(X_1, X_2, ..., X_n)$ is a sequence of identically distributed (but *not* independent) indicator random variables. In fact the sequence is exchangeable.

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.

$$\mathbb{P}(Y = k) = \frac{\binom{r}{k} \binom{m - r}{n - k}}{\binom{m}{n}}, \quad k \in \{0, 1, ..., n\}$$

✚ 27. Show that $\mathbb{E}(Y) = n \frac{r}{m}$ in the following ways:

    a. Using the definition.
    b. Using the representation as a sum of indicator variables.

✖ 28. In the ball and urn experiment, vary n, r, and m and note the shape of the density function and the location of the mean. For selected values of n and p, run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the sample mean to the distribution mean.

The Poisson Distribution

Recall that the **Poisson distribution** has density function

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

where $a > 0$ 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.

✚ 29. Suppose that N has the Poisson distribution with parameter a. Show that $\mathbb{E}(N) = a$. Thus, the parameter of the Poisson distribution is the mean of the distribution.

✖ 30. In the Poisson experiment, the parameter is $a = r\,t$. Vary the parameter and note the shape of the density function and the location of the mean. For various values of the parameter, run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the sample mean to the distribution mean.

The Exponential Distribution

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

$$f(t) = r\,e^{-r\,t}, \ \ t \geq 0$$

where $r > 0$ 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.

31. Suppose that T has the exponential distribution with rate parameter r. Show that

   a. Show that $\mathbb{E}(T) = \frac{1}{r}$
   b. Show that the mode of T is 0.
   c. Show that the median of T is $\frac{\ln(2)}{r}$.
   d. Sketch the graph of f and show the location of the mean, median, and mode on the x-axis.

32. In the random variable experiment, select the gamma distribution. Set $k = 1$ to get the exponential distribution. Vary r with the scroll bar and note the position of the mean relative to the graph of the density function. Now with $r = 2$, run the experiment 1000 times updating every 10 runs. Note the apparent convergence of the sample mean to the distribution mean.

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

The Gamma Distribution

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

$$f(t) = r^{n}\,\frac{t^{n-1}}{(n-1)!}\,e^{-r\,t}, \ \ t \geq 0$$

where $n \in \mathbb{N}_{+}$ 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". The gamma distribution is studied in detail in the chapter on the Poisson Process. In particular, if $(X_1, X_2, ..., X_n)$ 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.

34. Suppose that T has the gamma distribution with shape parameter n and rate parameter r. Show that $\mathbb{E}(T) = \frac{n}{r}$ in two ways:

   a. Using the definition.
   b. Using the decomposition as a sum of exponential variables.

35. In the random variable experiment, select the gamma distribution. Vary the parameters and note the position of the

mean relative to the graph of the density function. For selected parameter values, run the experiment 1000 times updating every 10 runs. Note the apparent convergence of 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.

36. Suppose that X has probability density function $f(x) = 12\,x^2\,(1 - x),\ 0 \le x \le 1$.

   a. Find the mean of X.
   b. Find the mode of X
   c. Find the approximate median of X.
   d. Sketch the graph of f and show the location of the mean, median, and mode on the x-axis.

37. In the random variable experiment, select the beta distribution and set $a = 3$ and $b = 2$ to get the distribution in the last exercise. Run the experiment 1000 times, updating every 10 runs, and note the apparent convergence of the sample mean to the distribution mean.

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

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

39. Suppose that X has probability density function $f(x) = \dfrac{1}{\pi\,\sqrt{x\,(1-x)}},\ 0 < x < 1$. This particular beta distribution is also known as the **arcsine distribution**.

   a. Find the mean of X.
   b. Find median of X.
   c. Sketch the graph of f and show the location of the mean, median, and mode on the x-axis.

The Pareto Distribution

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

$$f(x) = \frac{a}{x^{a+1}},\ x \ge 1$$

where $a > 0$ is a parameter. The Pareto distribution is named for Vilfredo Pareto. It is a heavy-tailed distribution that is

widely used to model financial variables such as income. The Pareto distribution is studied in detail in the chapter on Special Distributions.

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

$$\mathbb{E}(X) = \begin{cases} \infty, & a \in (0, 1] \\ \dfrac{a}{a-1}, & a \in (1, \infty) \end{cases}$$

41. In the random variable experiment, select the Pareto distribution. For the following values of the shape parameter a, run the experiment 1000 times updating every 10 runs. Note the behavior of the empirical mean.

  a. $a = 1$
  b. $a = 2$
  c. $a = 3$

The Cauchy Distribution

Recall that the **Cauchy distribution** has probability density function

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

This distribution is named for Augustin Cauchy and is a member of the family of **student t distributions**. The t distributions are studied in detail in the chapter on Special Distributions.

42. Suppose that X has the Cauchy distribution.

  a. Sketch the graph of f.
  b. Show that $\mathbb{E}(X)$ does not exist.
  c. Show that the median of X is 0.
  d. Show that the mode of X is 0.

43. In the random variable experiment, select the student t distribution. Set $n = 1$ to get the Cauchy distribution. Run the simulation 1000 times, updating every 10 runs. Note the behavior of the empirical mean.

The Normal Distribution

Recall that the **standard normal distribution** is a continuous distribution with density function $\phi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2}$, $z \in \mathbb{R}$. 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.

44. Suppose that Z has the standard normal distribution.

  a. Sketch the graph of ϕ

b. Show that $\mathbb{E}(Z) = 0$.

c. Show that the mode of Z is 0.

d. Show that the median of Z is 0.

45. Suppose again that Z has the standard normal distribution and that $\mu \in (-\infty, \infty)$, $\sigma \in (0, \infty)$. Recall that $X = \mu + \sigma Z$ has the normal distribution with **location parameter** μ and **scale parameter** σ. Show that $\mathbb{E}(X) = \mu$, so that the location parameter is the mean.

46. In the random variable experiment, select the normal distribution. Vary the parameters and note the location of the mean. For selected parameter values, run the simulation 1000 times, updating every 10 runs, and note the apparent convergence of the empirical mean to the true mean.

Additional Exercises

47. Suppose that (X, Y) has probability density function $f(x, y) = x + y, \ 0 \le x \le 1, \ 0 \le y \le 1$. Find the following expected values:

a. $\mathbb{E}(X)$

b. $\mathbb{E}(X^2 Y)$

c. $\mathbb{E}(X^2 + Y^2)$

d. $\mathbb{E}(X Y | Y > X)$

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

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

50. 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. *Hint*: Express the number of ducks killed as a sum of indicator random variables.

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

Additional Properties

Special Results for Nonnegative Variables

51. Prove **Markov's inequality** (named after Andrei Markov): If X is a nonnegative random variable, then for $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$$

a. Let I_t denote the indicator variable of the event $\{X \geq t\}$. Show that $t\,I_t \leq X$.
b. Take expected values through the inequality in (a).

52. Let X be a nonnegative random variable, with either a discrete or continuous distribution. Show that
$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x)dx$$

Hint: In the representation above, express $\mathbb{P}(X > x)$ in terms of the probability density function of X, as a sum in the discrete case or an integral in the continuous case. Then interchange the integral and the sum (in the discrete case) or the two integrals (in the continuous case).

53. Use the result of Exercise 52 to prove the change of variables formula for expected value when the random variable X has a continuous distribution on $S \subseteq \mathbb{R}^n$ with probability density function f, and r is a nonnegative function on S.

The following result is similar to Exercise 52, but is specialized to nonnegative integer valued variables:

54. Suppose that N is a discrete random variable that takes values in $\mathbb{N}$. Show that
$$\mathbb{E}(N) = \sum_{n=0}^\infty \mathbb{P}(N > n) = \sum_{n=1}^\infty \mathbb{P}(N \geq n)$$

Hint: In the first representation, express $\mathbb{P}(N > n)$ as a sum in terms of the probability density function of N. Then interchange the two sums. The second representation can be obtained from the first by a change of variables in the summation index.

A General Definition of Expected Value

The result in Exercise 52 could be used as the basis of a general formulation of expected value that would work for discrete, continuous, or even mixed distributions, and would not require the assumption of the existence of density functions. First, the result in Exercise 52 is taken as the definition of $\mathbb{E}(X)$ if X is nonnegative. Next, for $x \in \mathbb{R}$, we define the **positive and negative parts** of x as follows

$$x^+ = \max\{x, 0\}, \quad x^- = \max\{0, -x\}$$

55. Show that

a. $x^+ \geq 0, \; x^- \geq 0$
b. $x = x^+ - x^-$
c. $|x| = x^+ + x^-$

Now, if X is a real-valued random variable, then X^+ and X^-, the positive and negative parts of X, are nonnegative random variables. Thus, assuming that $\mathbb{E}(X^+) < \infty$ or $\mathbb{E}(X^-) < \infty$ we would define (anticipating linearity)

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$$

The usual formulas for expected value in terms of the probability density function, for discrete, continuous, or mixed distributions, would now be proven as theorems. Essentially this would be Exercise 52 with the hypotheses and conclusions reversed.

⊞ 56. We can finally finish our proof of the change of variables formula for expected value when X has a continuous distribution on $S \subseteq \mathbb{R}^n$ with probability function f, and r is a real-valued function on S. *Hint:* Decompose r into its positive and negative parts, and then use the result in Exercise 53.

Jensens's Inequality

Our next sequence of exercises will establish an important inequality known as **Jensen's inequality**, named for Johan Jensen. First we need a definition. A real-valued function g defined on an interval $S \subseteq \mathbb{R}$ is said to be **convex** on S if for each $t \in S$, there exist numbers a and b (that may depend on t), such that

$$a\,t + b = g(t) \text{ and } a\,x + b \leq g(x) \text{ for } x \in S$$

The graph of $T(x) = a\,x + b$ is called a **supporting line** at t. Thus, a convex function has *at least one* supporting line at each point in the domain
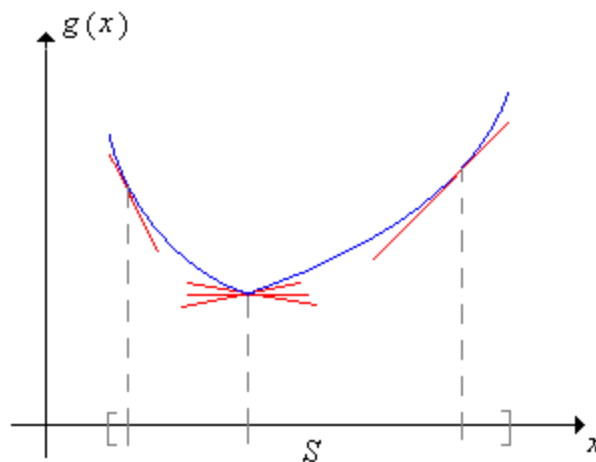

You may be more familiar with convexity in terms of the following theorem from calculus: If g has a continuous, non-negative second derivative on S, then g is convex on S (since the tangent line at t is a supporting line at t for each $t \in S$).

⊞ 57. Prove Jensen's inequality: If random variable X takes values in an interval S and g is convex on S, then
$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$$

Hint: In the definition of convexity given above, let $t = \mathbb{E}(X)$ and replace x with X. Then take expected values through the inequality.

Jensens's inequality extends easily to higher dimensions. The 2-dimensional version is particularly important, because it will be used to derive several special inequalities in the next section. First, a subset $S \subseteq \mathbb{R}^n$ is **convex** if for every pair of points in S, the line segment connecting those points also lies in S:
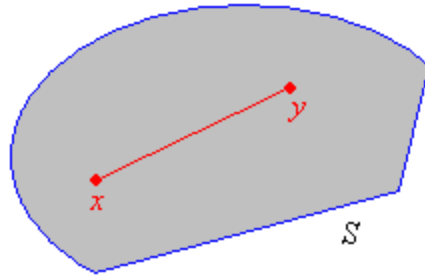
$$(x \in S) \text{ and } (y \in S) \text{ and } (p \in [0, 1]) \Rightarrow \left((1 - p)\, x + p\, y \in S\right)$$

Next, a real-valued function g on S is said to be **convex** if for each $t \in S$, there exist $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ (depending on t) such that

$$a \cdot t + b = g(t) \quad \text{and} \quad a \cdot x + b \leq g(x) \text{ for } x \in S$$

In $\mathbb{R}^2$, the graph of $T(x) = a \cdot x + b$ is called a **supporting plane** at t. From calculus, if g has continuous second derivatives on S and has a **positive non-definite** second derivative matrix, then g is convex on S.

■ 58. Suppose that $X = (X_1, X_2, ..., X_n)$ takes values in $S \subseteq \mathbb{R}^n$. Let $\mathbb{E}(X) = (\mathbb{E}(X_1), \mathbb{E}(X_2), ..., \mathbb{E}(X_n))$. Prove Jensen's inequality: if S is convex and g is a real-valued, convex function on S then

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$$

Hint: In the definition of convexity, let $t = \mathbb{E}(X)$ and let $x = X$. Then take expected values through the inequality. We will study the expected value of random vectors and matrices in more detail in a later section.

In both the one and n-dimensional cases, a function g is **concave** if the inequality in the definition is reversed. Jensen's inequality also reverses.

Exercises

■ 59. Suppose that X has probability density function $f(x) = r\, e^{-r\, x}$, $x \geq 0$ where $r > 0$. Thus, X has the exponential distribution with rate parameter r.

   a. Verify that $\mathbb{E}(X) = \frac{1}{r}$ using the formula in Exercise 52.
   b. Compute both sides of Markov's inequality.

?

■ 60. Suppose that W has probability density function $g(n) = p\,(1 - p)^{n-1}$, $n \in \mathbb{N}_+$ where $p \in (0, 1]$. Thus, W has the geometric distribution on $\mathbb{N}_+$ with success parameter p, and models the trial number of the first success in a sequence of Bernoulli trials.

   a. Verify that $\mathbb{E}(W) = \frac{1}{p}$ using the formula in Exercise 54.
   b. Compute both sides of Markov's inequality.
   c. Find $\mathbb{E}(W | W \text{ is even})$.

61. Suppose that X has probability density function $f(x) = \dfrac{a}{x^{a+1}}$, $x \geq 1$, where $a > 1$. Thus, X has the Pareto distribution with shape parameter a.

   a. Find $\mathbb{E}(X)$ using the formula in Exercise 52.

   b. Find $\mathbb{E}\left(\dfrac{1}{X}\right)$.

   c. Show that $g(x) = \dfrac{1}{x}$ is convex on $(0, \infty)$.

   d. Verify Jensen's inequality by comparing (b) and the reciprocal of (a).

62. Suppose that (X, Y) has probability density function $f(x, y) = 2(x + y)$, $0 \leq x \leq y \leq 1$

   a. Show that $g(x, y) = x^2 + y^2$ is convex on the domain of f.

   b. Compute $\mathbb{E}(X^2 + Y^2)$.

   c. Compute $\mathbb{E}(X)^2 + \mathbb{E}(Y)^2$.

   d. Verify Jensen's inequality by comparing (b) and (c).

63. Suppose that $\{x_1, x_2, \ldots, x_n\}$ is a set of positive numbers. Show that the **arithmetic mean** is at least as large as the **geometric mean**:

$$\left(x_1 \, x_2 \, \cdots \, x_n\right)^{1/n} \leq \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Hint: Let X be uniformly distributed on $\{x_1, x_2, \ldots, x_n\}$ and then use Jensen's inequality with $g(x) = \ln(x)$.