5 Probability

5.1 Modelling Uncertainty

Probability theory is the mathematical discipline concerned with modeling situations in which the outcome is uncertain. For example, in random sampling, we do not know which sample of inviduals from the population that we might actually get in our sample. The basic notion is that of a probability.

Definition 5.1.1 (A probability)**.** A **probability** is a number meant to measure the likelihood of the occurrence of some uncertain event (in the future).

Definition 5.1.2 (probability)**.** Probability (or the **theory of probability**) is the mathematical discipline that

1. constructs mathematical models for "real-world" situations that enable the computation of probabilities ("applied" probability)
2. develops the theoretical structure that undergirds these models ("theoretical" or "pure" probability).

The setting in which we make probability computations is that of a random process. (What we call a random process is usually called a random experiment in the literature but we use process here so as not to get the concept confused with that of randomized experiment.) A random process has three key characteristics:

Characteristics of a Random Process:

1. A random process is something that is to happen in the future (not in the past). We can only make probability statements about things that have not yet happened.
2. The outcome of the process could be any one of a number of outcomes and which outcome will obtain is uncertain.
3. The process could be repeated indefinitely (under essentially the same circumstances), at least in theory.

Historically, some of the basic random processes that were used to develop the theory of probability were those originating in games of chance. Tossing a coin or dealing a poker hand from a well-shuffled deck are examples of such processes. For our purposes the two most important random processes are producing a random sample from a population and assigning subjects randomly to the treatments of a randomized comparative experiment. (It is clear that each of these scenarios has all three characteristics of a random process.) Essentially all the probability statements that we want to make in statistics come from these two situations (and their cousins).

The first step in understanding a random process is to identify what might happen.

Definition 5.1.3 (sample space, event)**.** Given a random process, the **sample space** is the set (collection) of all possible outcomes of the process. An **event** of the random process is any subset of the sample space.

The next example lists several random processes, their sample spaces, and a typical event for each.

Example 5.1.1

1. A fair die is tossed. The sample space can be described as the set $S = \{1, 2, 3, 4, 5, 6\}$. A typical event might be $E = \{2, 4, 6\}$; i.e., the event that an even number is rolled.
2. A card is chosen from a well-shuffled standard deck of playing cards. There are 52 outcomes in the sample space. A typical event might be "A heart is chosen" which is a subset consisting of 13 of the possible outcomes.
3. Twenty-nine students are in a certain statistics class. It is decided to choose a simple random sample of 5 of the students. There are a boatload of possible outcomes. (It can be shown that there are 118,755 different samples of 5 students out of 29.) One event of interest is the collection of all outcomes in which all 5 of the students are male. Suppose that 25 of the students in the class are male. Then it can be shown that 53,130 of the outcomes comprise this event.

We often have some choice as to what we call outcomes of a random process. For example, in Example 5.1.1(3), we might consider two samples different outcomes if the students in the sample are chosen in a different order, even if the same five students appear in the samples. Or we might call such samples the same outcome. To some extent, what we call an outcome depends on the way in which we are going to use the results of the random process.

Given a random process, our goal is to assign to each event E a number $P(E)$ (called the **probability of** E) such that $P(E)$ measures in some way the likelihood of E. In order to assign such numbers however, we need to understand what they are intended to measure.

Interpreting probability computations is fraught with all sorts of philosophical issues but it is not too great a simplification at this stage to distinguish between two different interpretations of probability statements.

The frequentist interpretation.
The probability of an event E, $P(E)$, is the limit of the relative frequency that E occurs in repeated trials of the process as the number of trials approaches infinity.

In other words, if the event E occurs e_n many times in the first n trials, then on the frequentist interpretation, $P(E) = \lim_{n\to\infty} e_n/n$.

The subjectivist interpretation.
The probability of an event E, $P(E)$, is an expression of how confident the assignor is that the event will happen in the next trial of the process.

It is easy to think of examples of probability statements in the real world that are more naturally interpreted using either of these interpretations rather than the other. In this text, we will usually phrase our interpretations of probability statements using the frequentist interpretation. Mathematics cannot tell us which of these two interpretations is right or indeed how to assign probabilities in any particular situation. But mathematicians have developed some basic axioms to constrain our choice of probabilities. The three fundamental axioms of probability are

Axiom 5.1.4. For all events A, $P(A) \geq 0$.

Axiom 5.1.5. $P(\mathcal{S}) = 1$.

Axiom 5.1.6. If A_1 and A_2 are disjoint events (i.e., have no outcomes in common) then

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

If one interprets probabilities as limiting relative frequencies, it is easy to see that these three axioms should be true.

The axioms do not tell us how to assign the probabilities in any particular case. They only provide some minimal constraints on this assignment. There are two important methods for assigning that we will use extensively.

The equally likely outcomes model

In some cases, we can list all the outcomes in such a way that is plausible to suppose that each outcome is equally likely. For example, the very definition of choosing a random sample of size 5 from a class of 29 requires us to develop a method so that each sample of size 5 is equally likely to occur. In this case, it is easy to compute the probability of an event E. If there are N equally likely outcomes, the probability of each outcome should be $1/N$. The probability of an event E is k/N where there are k outcomes in the event.

Example 5.1.2

A six-sided die is rolled. Then one of six possible outcomes occurs. From the symmetry of the die it is reasonable to assume that the six outcomes are equally likely. Therefore, the probability of each outcome is $1/6$. If E is the event that is described by "the die comes up 1 or 2" then $P(E) = 2/6 = 1/3$.

In a more interesting and more useful example in Example **??** there are 118,755 possible different samples of five students from 29 and by the definition of simple random sample these samples are equally likely to occur. Since 53,130 of these comprise the event E of getting all males in the sample, the probability of this event is 53130/118755 = 44.7%.

Example 5.1.3

Perhaps the canonical historical example of a random process for which it is possible to generate a list of equally likely outcomes is the process in which two dice are thrown and the number on each face is recorded. It is easy to see that there are 36 equally likely outcomes (list the pairs (i, j) of numbers where i is the number on the first die, j is the number on the second die and i and j range from 1 to 6). One event related to this process is the event E that the throw results in a sum of 7 on the two dice. It is easy to see that there are 6 outcomes in E so that $P(E) = 6/36 = 1/6$.

Past performance as an indicator of the future

In some cases, we have data on many previous trials of the process. In this case we may estimate the probability of each outcome by the relative frequency with which it occurred in the previous trials. This method is used extensively in the insurance industry. For example the probability that a male alive on his 55th birthday dies before his 56th is currently estimated to be 0.0081 or slightly less than 1% based on the recent history of 55 year old males.

Example 5.1.4

In the 2007 baseball season, Manny Ramirez came to the plate 569 times. Of those 569 times, he had 89 singles, 33 doubles, 1 triple, 20 homeruns, 78 walks (and hit by pitch),

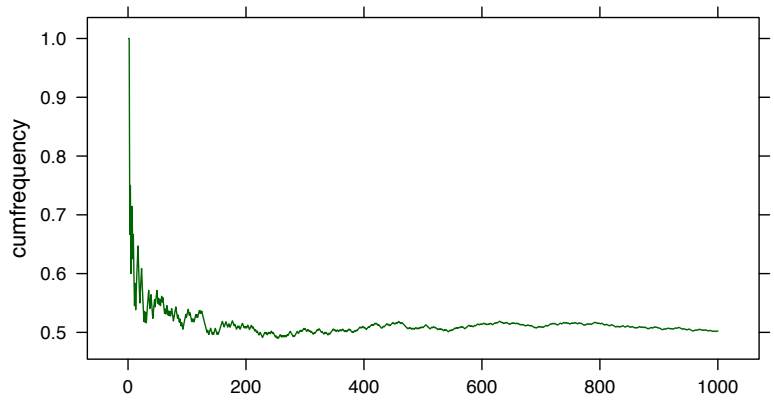
and 348 outs. We might estimate that the probability Ramirez will hit a homerun in his next plate appearance to be 20/569 = .035.

For the purpose of investigating how random processes work, it is very useful to use R. In the following example, we simulate one, and then five, of Manny Ramirez's plate appearances.

```
> outcomes=c('Out','Single','Double','Triple','Homerun','Walk')
> ramirez=c(348,89,33,1,20,78)/569
> sum(ramirez)
[1] 1
> ramirez
[1] 0.611599297 0.156414763 0.057996485 0.001757469 0.035149385 0.137082601
> sample(outcomes,1,prob=ramirez)
[1] "Double"
> sample(outcomes,5,prob=ramirez,replace=T)
[1] "Out"    "Double" "Out"    "Out"    "Walk"
```

In the next example, we simulate the tossing of a coin 1,000 times. The graph provides some evidence that the limiting relative frequency of "Heads" is 0.5.

```
> coins=sample(c('H','T'),1000,replace=T)
> cumfrequency = cumsum(coins=='H')/c(1:1000)
> plot(cumfrequency,type='l')
```



5.2 Discrete Random Variables

5.2.1 Random Variables

If the outcomes of a random process are numbers, we will call the random process a **random variable.** Since non-numerical outcomes can always be coded with numbers, restricting our attention to random variables results in no loss of generality. We will use upper-case letters to name random variables (X, Y, etc.) and the corresponding lower-case letters (x, y, etc.) to denote the possible values of the random variable. Then we can describe events by equalities and inequalities so that we can write such things as $P(X = 3)$, $P(Y = y)$ and $P(Z \leq z)$. Some examples of random variables include

1. Choose a random sample of size 12 from 250 boxes of Raisin Bran. Let X be the random variable that counts the number of underweight boxes and let Y be the random variable that is the average weight of the 12 boxes.
2. Choose a Calvin senior at random. Let Z be the GPA of that student and let U be the composite ACT score of that student.
3. Assign 12 chicks at random to two groups of six and feed each group a different feed. Let D be the difference in average weight between the two groups.
4. Throw a fair die until all six numbers have appeared. Let T be the number of throws necessary.

We will consider two types of random variables, discrete and continuous.

Definition 5.2.1 (discrete random variable)**.** A random variable X is **discrete** if its possible values can be listed $x_1, x_2, x_3, \ldots$.

In the example above, the random variables X, U, and T are discrete random variables. Note that the possible values for X are $0, 1, \ldots, 12$ but that T has infinitely many possible values $1, 2, 3, \ldots$. The random variables Y, Z, and D above are not discrete. The random variable Z (GPA) for example can take on all values between 0.00 and 4.00. (We should make the following caveat here however. All variables are discrete in the sense that there are only finitely many different measurements possible to us. Each measurement device that we use has divisions only down to a certain tolerance. Nevertheless it is usually more helpful to view these measurements as on a continuous scale rather than a discrete one. We learned that in calculus.)

The following definition is not quite right—it omits some technicalities. But it is close enough for our purposes.

Definition 5.2.2 (continuous random variable)**.** A random variable X is **continuous** if its possible values are all x in some interval of real numbers.

In this section, we focus on properties of *discrete* random variables.

If X is a discrete random variable, we will be able to compute the probability of any event defined in terms of X if we know all the possible values of X and the probability $P(X = x)$ for each such value x.

Definition 5.2.3 (probability mass function)**.** The **probability mass function** (pmf) of a discrete random variable X is the function f such that for all x, $f(x) = P(X = x)$. We will sometimes write f_X to denote the probability mass function of X when we want to make it clear which random variable is in question.

The word mass is not arbitrary. It is convenient to think of probability as a unit mass that is divided into point masses at each possible outcome. The mass of each point is its probability. Note that mass obeys the probability axioms.

Example 5.2.1

Two dice are thrown and the sum X of the numbers appearing on their faces is recorded. X is a random variable with possible values $2, 3, \ldots, 12$. By using the method of equally likely outcomes, we can see that the pmf f of X is given by the following table:

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

We can now compute such probabilities as $P(X \leq 5) = 5/18$ by adding the appropriate values of f.

Example 5.2.2

We can think of a categorical variable as a discrete random variable by coding. Suppose that a student is chosen at random from the Calvin student body. We will code the class of the student by $1, 2, 3, 4$ for the four standard classes and 5 for other. The coded class is a random variable. Referring to Table 4.6, we see that the probability mass function of X is given by $f(1) = 0.27$, $f(2) = 0.24$, $f(3) = 0.21$, $f(4) = 0.25$, $f(5) = 0.03$, and $f(x) = 0$ otherwise.

One useful way of picturing a probability mass function is by a probability histogram. For the mass function in Example 5.2.2, we have the corresponding histogram in Figure 5.2.1.

On the frequentist interpretation of probability, if we repeat the random process many times, the histogram of the results of those trials should approximate the probability histogram. The probability histogram is not a histogram of data from many trials however. It is a representation of what might happen in the next trial. We will often use this idea to work in reverse. In other words, given a histogram of data obtained from successive trials

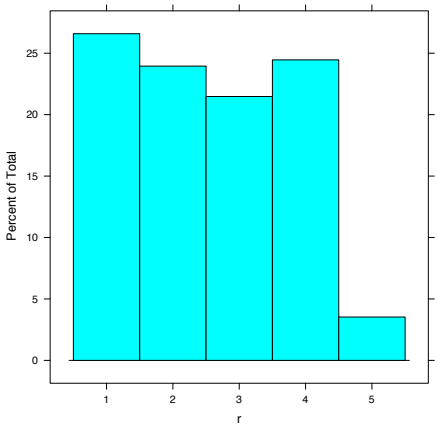


Figure 5.1: The probability histogram for the Calvin class random variable.

of a random process, we will choose the pmf to fit the data. Of course we might not ask for a perfect fit but instead we will choose the pmf f to fit the data approximately but so that f has some simple form.

Several families of random variables are particularly important to us and provide models for many real-world situations. We examine two of such families here. Each arises from a common kind of random process that will be important for statistical inference. The second of these arises from the very important case of simple random sampling from a population. We will first study a somewhat different case (which, among other uses, can be used to study sampling with replacement).

5.2.2 The Binomial Distribution

A **binomial process** is a process characterized by the following conditions:

1. The process consists of a sequence of finitely many (n) *trials* of some simpler process.
2. Each trial results in one of two possible outcomes, usually called success (S) and failure (F).
3. The probability of success on each trial is a constant denoted by π.
4. The trials are independent one from another - that is the outcome of one trial does not affect the outcome of any other.

Thus a binomial process is characterized by two **parameters**, n and π. Given a binomial process, the natural random variable to observe is the number of successes.

Definition 5.2.4 (binomial random variable)**.** Given a binomial process, the **binomial random variable** X associated with this process is defined by X is the number of successes in the n trials of the process. If X is a binomial random variable with parameters n and π, we write $X \sim \mathsf{Binom}(n, \pi)$.

The symbol $\sim$ can be read as "has the distribution" or something to that effect. Our use of the word distribution is consistent with its meaning defined earlier. Here to specify a distribution is to specify the possible values of the random variable and the probability that the random variable attains any particular value.

Example 5.2.3

The following are all natural examples of binomial random variables.

1. A fair coin is tossed $n = 10$ times with the probability of a HEAD (success) being $\pi = .5$. X is the number of heads.
2. A basketball player shoots $n = 25$ freethrows with the probability of making each freethrow being $\pi = .70$. Y is the number of made freethrows.
3. A quality control inspector tests the next $n = 12$ widgets off the assembly line each of which has a probability of 0.10 of being defective. Z is the number of defective widgets.
4. Ten Calvin students are randomly sampled with replacement. W is the number of males in the sample.

The probability mass function for a binomial distribution is given in the following theorem.

Theorem 5.2.5 (The Binomial Distribution)**.** Suppose that X is a binomial random variable with parameters n and π. The pmf of X is given by

$$f_X(x; n, \pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x} = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x}, \qquad x = 0, 1, 2, \ldots, n .$$

Note the use of the semicolon in the defintion of f_X in the theorem. We will use a semicolon to separate the possible values of the random variable (x) from the parameters (n, π). For any particular binomial experiment, n and π are fixed. If n and π are understood, we might write $f_X(x)$ for $f_X(x; n, \pi)$.

For all but very small n, computing f by hand is tedious. We will use `R` to do this. Besides computing the mass function, `R` can be used to compute the cumulative distribution function F_X which is the useful function defined in the next definition.

Definition 5.2.6 (cumulative distribution function)**.** If X is any random variable, the **cumulative distribution function** of X (cdf) is the function F_X given by

$$F_X(x) = P(X \le x) = \sum_{y \le x} f_X(y)$$

We will usually use the convention that the pmf of X is named by a lower-case letter (usually f_X) and the cdf by the corresponding upper-case letter (usually F_X). The `R` functions to compute the cdf, pdf, and also to simulate binomial processes are as follows if $X \sim \mathsf{Binom}(n, \pi)$.

function (& parameters)	explanation
`rbinom(n,size,prob)`	makes `n` random draws of the random variable X and returns them in a vector.
`dbinom(x,size,prob)`	returns P(X = `x`) (the pmf).
`pbinom(q,size,prob)`	returns P($X \le$ `q`) (the cdf).

Example 5.2.4

Suppose that a manufacturing process produces defective parts with probability $\pi = .1$. If we take a random sample of size 10 and count the number of defectives X, we might assume that $X \sim \mathsf{Binom}(10, 0.1)$. Some examples of `R` related to this situation are as follows.

```
> defectives=rbinom(n=30, size=10,prob=0.1)
> defectives
 [1] 2 0 2 0 0 0 0 2 0 1 1 1 0 0 2 2 3 1 1 2 1 1 0 2 0 1 1 0 1 1
> table(defectives)
defectives
 0  1  2  3
11 11  7  1
> dbinom(c(0:4),size=10,prob=0.1)
[1] 0.34867844 0.38742049 0.19371024 0.05739563 0.01116026
> dbinom(c(0:4),size=10,prob=0.1)*30                    # pretty close to table
[1] 10.4603532 11.6226147  5.8113073  1.7218688  0.3348078
> pbinom(c(0:5),size=10,prob=0.1)                       # same as cumsum(dbinom(...))
[1] 0.3486784 0.7360989 0.9298092 0.9872048 0.9983651 0.9998531
>
```

It is important to note that

- `R` uses `size` for the number of trials (what we have called n) and `n` for the number of random draws.

- `pbinom()` gives the cdf not the pdf. Reasons for this naming convention will become clearer later.

- There are similar functions in `R` for many of the distributions we will encounter, and they all follow a similar naming scheme. We simply replace `binom` with the `R`-name for a different distribution.

5.2.3 The Hypergeometric Distribution

The hypergeometric distribution arises from considering the situation of random sampling from a population in which there are just two types of individuals. (That is there is a categorical variable defined on the population with just two levels.) It is traditional to describe the distribution in terms of the urn model. Suppose that we have an urn with two different colors of balls. There are m white balls and n black balls. Suppose we choose k balls from the urn in such a way that every set of k balls is equally likely to be chosen (i.e., a random sample of balls) and count the number X of white balls. We say that X has the **hypergeometric distribution** with parameters m, n, and k and write $X \sim \mathsf{Hyper}(m, n, k)$. A simple example shows how we can compute probabilities in this case.

Example 5.2.5

Suppose the urn has 2 white and 3 black balls and that we choose 2 balls at random without replacement. If X is the number of white balls, we have $X \sim Hyper(2, 3, 2)$. Notice that in this case there are 10 different possible choices of two balls. If we label the balls W1, W2, B1, B2, B3, we have the following:

2 whites	(W1,W2)
1 white	(W1,B1),(W1,B2), (W1,B3), (W2,B1), (W2,B2), (W2,B3)
0 whites	(B1,B2), (B1,B3), (B2,B3)

Since the 10 different pairs are equally likely, we have $P(X = 0) = 3/10$, $P(X = 1) = 6/10$, and $P(X = 2) = 1/10$.

The systematic counting of Example 5.2.5 can generalized to yield a simple formula for the pmf of any hypergeometric random variable.

Theorem 5.2.7. Suppose that $X \sim \mathsf{Hyper}(m, n, k)$. Then the pmf f of X is given by

$$f_X(x; m, n, k) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}, \qquad x \le \min(k, m) .$$

`R` knows the hypergeometric distribution and the syntax is exactly the same as for the binomial distribution (except that the names of the parameters have changed).

function (& parameters)	explanation
`rhyper(nn,m,n,k)`	makes `nn` random draws of the random variable X and returns them in a vector.
`dhyper(x,m,n,k)`	returns $P(X = x)$ (the pmf).
`phyper(q,m,n,k)`	returns $P(X \leq q)$ (the cdf).

Example 5.2.6

Suppose that a statistics class has 29 students, 25 of whom are male. Let's call the females the white balls and the males the black balls. suppose that we choose 5 of these students at random and without replacement, i.e., a random sample of size 5. Let X be the number of females in our sample. Then $X \sim \mathsf{Hyper}(4, 25, 5)$. Some interesting questions related to this random variable are answered by the R output below.

```
> dhyper(x=c(0:5),m=4,n=25,k=5)
[1] 0.4473916888 0.4260873226 0.1162056334 0.0101048377 0.0002105175
[6] 0.0000000000
> dhyper(x=c(0:5),k=5,m=4,n=25)            # order of named arguments does not matter
[1] 0.4473916888 0.4260873226 0.1162056334 0.0101048377 0.0002105175
[6] 0.0000000000
> phyper(q=c(0:5),m=4,n=25,k=5)
[1] 0.4473917 0.8734790 0.9896846 0.9997895 1.0000000 1.0000000
> rhyper(nn=30,m=4,n=25,k=5)              # note nn for number of random outcomes
 [1] 2 1 1 1 1 2 2 2 1 1 1 0 1 0 0 0 1 1 0 0 1 1 0 1 1 1 2 0 0 0
> dhyper(0:5,4,25,5)                      # default order of unnamed arguments
[1] 0.4473916888 0.4260873226 0.1162056334 0.0101048377 0.0002105175
[6] 0.0000000000
>
```

5.3 Continuous Random Variables

Recall that a continuous random variable X is one that can take on all values in an interval of real numbers. For example, the height of a randomly chosen Calvin student in inches could be any real number between, say, 36 and 80. Of course all continuous random variables are idealizations. If we measure heights to the nearest quarter inch, there are only finitely many possibilities for this random variable and we could, in principle, treat it as discrete. We know from calculus however that treating measurements as continuous valued functions often simplifies rather than complicates our techniques. In order to understand what kinds of probability statements that we would like to make about continuous random variables, it is helpful to keep in mind this idea of the finite precision of our measurements however. For example, a statement that a randomly chosen individual is 72 inches tall is

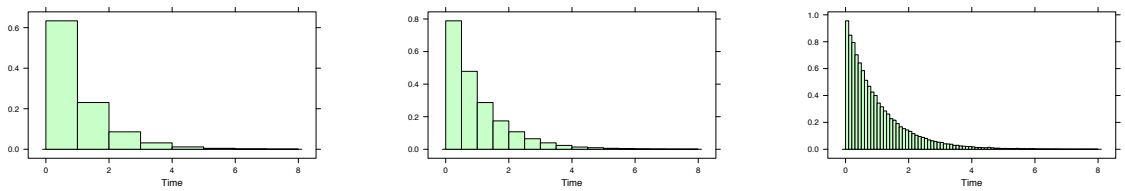


Figure 5.2: Discretized pmf for T.

not a claim that the individual is exactly 72 inches tall but rather a claim that the height of the individual is in some small interval (maybe $71\frac{3}{4}$ to $72\frac{1}{4}$ if we are measuring to the nearest half inch). So probabilities of the form $P(X = x)$ are not meaningful. Rather the appropriate probability statements will be of the form $P(a \leq X \leq b)$.

5.3.1 pdfs and cdfs

Recall the analogy of probability and mass. In the case of discrete random variables, we represented the probability $\mathrm{P}(X = x)$ by a point of mass $\mathrm{P}(X = x)$ at the point x and had total mass 1. In this case mass is continuous and the appropriate weighting of mass is a density function. In the following example, we can see how this works.

Example 5.3.1

A Geiger counter emits a beep when a radioactive particle is detected. The rate of beeping determines how radioactive the source is. Suppose that we record the time T to the next beep. It turns out that T behaves like a random variable. Suppose that we measured T with increasing precision. We might get histograms that look like those in Figure 5.2 for the pmf of T. It's pretty obvious that we want to replace these histograms by a smooth curve. In fact the pictures should remind us of the pictures drawn for the Riemann sums that define the integral.

The analogue to a probability mass function for a continuous variable is a probability density function.

Definition 5.3.1 (probability density function, continuous random variable)**.** A **probability density function** (pdf) is a function f such that

- $f(x) \geq 0$ for all real numbers x, and
- $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

The continuous random variable X defined by the pdf f satisfies

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\, dx\,, \qquad \text{for any real numbers } a \leq b\,.$$

The following simple lemma demonstrates one way in which continuous random variables are very different from discrete random variables.

Lemma 5.3.2. Let X be a continuous random variable with pdf f. Then for any $a \in \mathbb{R}$,

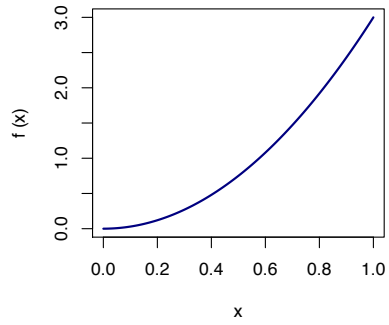
1. $P(X = a) = 0$,
2. $P(X < a) = P(X \leq a)$, and
3. $P(X > a) = P(X \geq a)$.

Proof. $\int_a^a f(x)\,dx = 0$. And $P(X \leq a) = P(X < a) + P(X = a) = P(X < a)$. □

Example 5.3.2

Q. Consider the function $f(x) = \begin{cases} 3x^2 & x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$ Show that f is a pdf and calculate $P(X \leq 1/2)$.

A. Let's begin be looking at a plot of the pdf.



The rectangular region of the plot has an area of 3, so it is plausible that the area under the graph of the pdf is 1. We can verify this by integration.

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 3x^2\,dx = x^3\Big|_0^1 = 1\,,$$

so f is a pdf and $P(X \leq 1/2) = \int_0^{1/2} 3x^2\,dx = x^3\Big|_0^{1/2} = 1/8$.

The cdf of a continuous random variable is defined the same way as it was for a discrete random variable, but we use an integral rather than a sum to get the cdf from the pdf in this case.

Definition 5.3.3 (cumulative distribution function)**.** Let X be a continuous random variable with pdf f, then the **cumulative distribution function** (cdf) for X is

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(t)\, dt\ .$$

Example 5.3.3

Q. Determine the cdf of the random variable from Example 5.3.2.

A. For any $x \in [0, 1]$,

$$F_X(x) = \mathrm{P}(X \leq x) = \int_0^x 3t^2\, dt = t^3\Big|_0^x = x^3\ .$$

So

$$F_X(x) = \begin{cases} 0 & x \in [-\infty, 0) \\ x^3 & x \in [0, 1] \\ 1 & x \in (1, \infty)\ . \end{cases}$$

Notice that the cdf F_X is an antiderivative of the pdf f_X. This follows immediately from the Fundamental Theorem of Calculus. Notice also that $\mathrm{P}(a \leq X \leq b) = F(b) - F(a)$.

Lemma 5.3.4. Let F_X be the cdf of a continuous random variable X. Then the pdf f_X satisfies

$$f_X(x) = \frac{d}{dx} F_X(x)\ .$$ □

Just as the binomial and hypergeometric distributions were important families of discrete random variables, there are several important families of continuous random variables that are often used as models of real-world situations. We investigate a few of these in the next three subsections.

5.3.2 Uniform Distributions

The continuous uniform distribution has a pdf that is constant on some interval.

Definition 5.3.5 (uniform random variable)**.** A **continuous uniform random variable** on the interval $[a, b]$ is the random variable with pdf given by

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to confirm that this function is indeed a pdf. We could integrate, or we could simply use geometry. The region under the graph of the uniform pdf is a rectangle with width $b - a$ and height $\frac{1}{b-a}$, so the area is 1.

Example 5.3.4

Q. Let X be uniform on $[0, 10]$. What is $P(X > 7)$? What is $P(3 \leq X < 7)$?

A. Again we argue geometrically. $P(X > 7)$ is represented by a rectangle with base from 7 to 10 along the x-axis and a height of .1, so $P(X > 7) = 3 \cdot 0.1 = 0.3$.

Similarly $P(3 \leq X < 7) = 0.4$. In fact, for any interval of width w contained in $[0, 10]$, the probability that X falls in that particular interval is $w/10$.

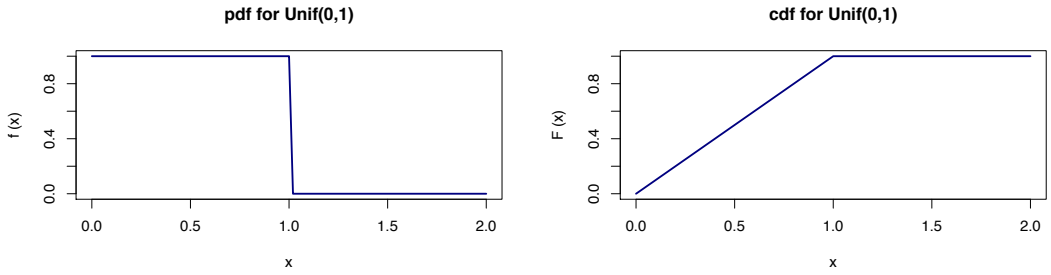
We could also compute these results by integrating, but this would be silly.

Example 5.3.5

Q. Let X be uniform on the interval $[0, 1]$ (which we denote $X \sim \mathsf{Unif}(0, 1)$) what is the cdf for X?

A. For $x \in [0, 1]$, $F_X(x) = \int_0^x 1 \, dx = x$, so

$$F_X(x) = \begin{cases} 0 & x \in (\infty, 0) \\ x & x \in [0, 1] \\ 1 & x \in (1, \infty) \end{cases}.$$



Although it has a very simple pdf and cdf, this random variable actually has several important uses. One such use is related to random number generation. Computers are not able to generate truly random numbers. Algorithms that attempt to simulate randomness are called pseudo-random number generators. $X \sim \mathsf{Unif}(0, 1)$ is a model for an idealized random number generator. Computer scientists compare the behavior of a pseudo-random number generator with the behavior that would be expected for X to test the quality of the pseudo-random number generator.

There are `R` functions for computing the pdf and cdf of a uniform random variable as well as a function to return random numbers. An additional function computes the

quantiles of the uniform distribution. If $X \sim \mathsf{Unif}(\texttt{min}, \texttt{max})$ the following functions can be used.

function (& parameters)	explanation
`runif(n,min,max)`	makes `n` random draws of the random variable X and returns them in a vector.
`dunif(x,min,max)`	returns $f_X(x)$, (the pdf).
`punif(q,min,max)`	returns $\mathrm{P}(X \leq \texttt{q})$ (the cdf).
`qunif(p,min,max)`	returns x such that $\mathrm{P}(X \leq x) = p$.

Here are examples of computations for $X \sim \mathsf{Unif}(0, 10)$.

```
> runif(6,0,10)    # 6 random values on [0,10]
[1] 5.449745 4.124461 3.029500 5.384229 7.771744 8.571396
> dunif(5,0,10)    # pdf is 1/10
[1] 0.1
> punif(5,0,10)    # half the distribution is below 5
[1] 0.5
> qunif(.25,0,10)  # 1/4 of the distribution is below 2.5
[1] 2.5
```

5.3.3 Exponential Distributions

In Example 5.3.1 we considered a "waiting time" random variable, namely the waiting time until the next radioactive event. Waiting times are important random variables in reliability studies. For example, a common characteristic of a manufactured object is MTF or mean time to failure. The model often used for the Geiger counter random variable is the exponential distribution. Note that a waiting time can be any x in the range $0 \leq x < \infty$.

Definition 5.3.6 (The exponential distribution)**.** The random variable X has the **exponential distribution** with parameter $\lambda > 0$ ($X \sim \mathsf{Exp}(\lambda)$) if X has the pdf

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0\,. \end{cases}$$

It is easy to see that the function f_X of the previous definition is a pdf for any positive value of λ. `R` refers to the value of λ as the `rate` so the appropriate functions in `R` are `rexp(n,rate)`, `dexp(x,rate)`, `pexp(x,rate)`, and `qexp(p,rate)`. We will see later that `rate` is an apt name for λ as λ will be the rate per unit time if X is a waiting time random variable.

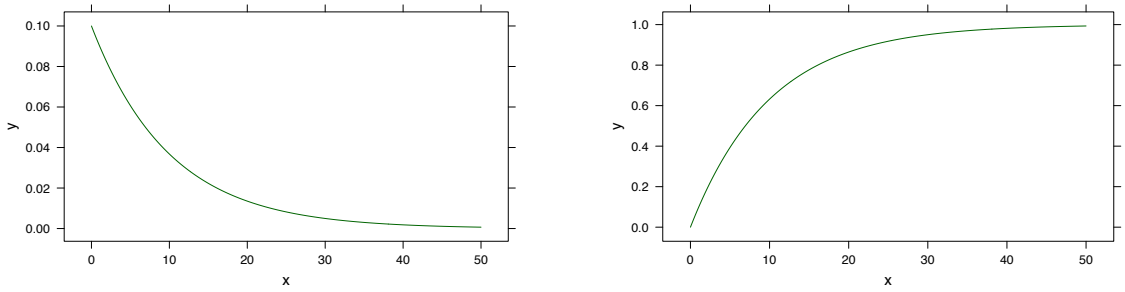


Figure 5.3: The pdf and cdf of the random variable $T \sim \mathsf{Exp}(0.1)$.

Example 5.3.6

Suppose that a random variable T measures the time until the next radioactive event is recorded at a Geiger counter (time measured since the last event). For a particular radioactive material, a plausible models for T is $T \sim \mathsf{Exp}(0.1)$ where time is measured in seconds. Then the following R session computes some important values related to T.

```
> pexp(q=0.1,rate=.1)   #  probability waiting time less than .1
[1] 0.009950166
> pexp(q=1,rate=.1)     #  probability waiting time less than 1
[1] 0.09516258
> pexp(q=10,rate=.1)
[1] 0.6321206
> pexp(q=20,rate=.1)
[1] 0.8646647
> pexp(100,rate=.1)
[1] 0.9999546
> pexp(30,rate=.1)-pexp(5,rate=.1)   # probability waiting time between 5 and 30
[1] 0.5567436
> qexp(p=.5,rate=.1)    # probability is .5 that T is less than 6.93
[1] 6.931472
```

The graphs in Figure 5.3 are graphs of the pdf and cdf of this random variable. All exponential distributions look the same except for the scale. The rate of 0.1 here means that we can expect that in the long run this process will average 0.1 counts per second.

We pause to note that, when given some random variable (such as the waiting time to a geiger counter event), we are not handed its pdf as well. The pdf represents a *model* of the situation. In choosing the model, we really are faced with two decisions.

1. Which family (e.g., uniform, exponential, etc.) of distributions best models the situation?

2. What particular values of the parameters should we use for the pdf?

Sometimes we can begin to answer Question 1 even before we collect data. Each of the distributions that we have met has certain properties which we check against our process. For example, it is often apparent whether the properties of a binomial process should apply to a certain process we are examining. Of course it is always useful to check our answer to Question 1 by collecting data and verifying that the shape of the distribution of the data collected is consistent with the distribution we are using. The only reasonable way to answer Question 2, however, is to collect data. In Example 5.3.6, for instance, we saw that if $X \sim \mathsf{Exp}(0.1)$ that $P(X \leq 6.93) = .5$. Therefore if about half of our data are less than 6.93, we would say that the data are consistent with the hypothesis that $X \sim \mathsf{Exp}(0.1)$ but if almost all the data are less than 5, we would probably doubt that X has this distribution. The problems of choosing the appropriate distribution and the appropriate values of the parameters is an important one which, unfortunately, cannot be adequately addressed in this course.

5.3.4 Weibull Distributions

A very important generalization of the exponential distributions are the Weibull distributions. They are often used by engineers to model phenomena such as failure, manufacturing or delivery times. They have also been used for applications as diverse as fading in wireless communications channels and wind velocity. The Weibull is a two-parameter family of distributions. The two parameters are a shape parameter α and a scale parameter λ.

Definition 5.3.7 (The Weibull distributions)**.** The random variable X has a **Weibull distribution** with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$ ($X \sim \mathsf{Weib}(\alpha, \beta)$) if the pdf of X is

$$f_X(x; \alpha, \beta) = \begin{cases} \left(\frac{\alpha}{\beta^\alpha}\right) x^{\alpha-1} e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Notice that if $X \sim \mathsf{Weib}(1, \lambda)$ then $X \sim \mathsf{Exp}(1/\lambda)$. Varying α in the Weibull distribution changes the shape of the distribution while changing β changes the scale. The effect of fixing β ($\beta = 5$) and changing α ($\alpha = 1, 2, 3$) is illustrated by the first graph in Figure 5.4 while the second graph shows the effect of changing β ($\beta = 1, 3, 5$) with α fixed at $\alpha = 2$. The appropriate `R` functions to compute with the Weibull distribution are `dweibull(x,shape,scale)`, `pweibull(q,shape,scale)`, etc.

Example 5.3.7

The Weibull distribution is sometimes used to model the maximum wind velocity measured during a 24 hour period at a specific location. The dataset http://www.calvin.edu/~stob/data/wind.csv gives the maximum wind velocity at the

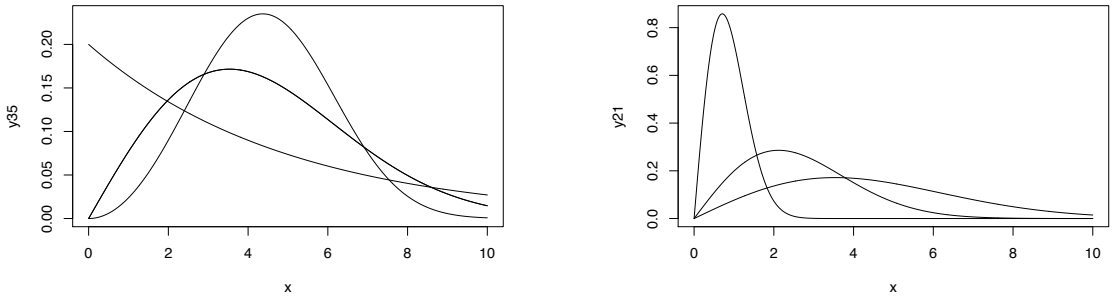


Figure 5.4: Left: fixed β. Right: fixed α.

San Diego airport on each of 6,209 consecutive days. It is claimed that the maximum wind velocity measured on a day behaves like a random variable W that has a Weibull distribution with $\alpha = 3.46$ and $\beta = 16.90$. The R code below investigates that model using this past data. (In fact, this model is not a very good one although the output below suggests that it might be plausible.)

```
> w$Wind
   [1] 14 11 10 13 11 11 26 21 14 13 10 10 13 10 13 13 12 12 13 17 11 11 13 25 15
  [26] 18 13 17 12 14 15 10 16 17 17 13 18 14 12 20 11 14 20 16 12 14 18 17 13 16
  [51] 13 16 11 13 11 15 13 15 16 18 14 15 15 14 14 16 15 18 14 16 14 10 17 14 12
.............
> cutpts=c(0,5,10,15,20,25,30)
> table(cut(w$Wind,cutpts))

  (0,5]  (5,10] (10,15] (15,20] (20,25] (25,30]
      2     434    3303    1910     409      95
> length(w$Wind[w$Wind<12.5])/6209
[1] 0.2728298                              #  27.3% days with max windspeed less than 12.
> pweibull(12.5,3.46,16.9)
[1] 0.2968784                              #  29.7% predicted by Weibull model
> length(w$Wind[w$Wind<22.5])/6209
[1] 0.951361
> pweibull(22.5,3.46,16.9)
[1] 0.9322498
> simulation=rweibull(100000,3.46,16.9)    # 100,000 simulated days
> mean(simulation)                         # simulated days have mean about the same as a
[1] 15.18883
> mean(w$Wind)
[1] 15.32405
> sd(simulation)                           # simulated days have greater variation
[1] 4.85144
```

```
> sd(w$Wind)
[1] 4.239603
>
```

5.4 Mean and Variance of a Random Variable

Just as numerical summaries of a data set can help us understand our data, numerical summaries of the distribution of a random variable can help us understand the behavior of that random variable. In this section we develop two of the most important numerical summaries of random variables: mean and variance. Just as the concepts with the same names we discussed earlier measure center and spread for data, so the ones we discuss here do for the distribution of a random variable. In each case, we will use our experience with data to help us develop a definition.

5.4.1 The Mean of a Discrete Random Variable

Example 5.4.1

Q. Let's begin with a motivating example. Suppose a student has taken 10 courses and received 5 A's, 4 B's and 1 C. Using the traditional numerical scale where an A is worth 4, a B is worth 3 and a C is worth 2, what is this student's GPA (grade point average)?

A. The first thing to notice is that $\frac{4+3+2}{3} = 3$ is not correct. We cannot simply add up the values and divide by the number of values. Clearly this student should have GPA that is higher than 3.0, since there were more A's than C's.

Consider now a correct way to do this calculation and some algebraic reformulations of it.

$$\begin{aligned}\text{GPA} = \frac{4+4+4+4+4+3+3+3+3+2}{10} &= \frac{5\cdot 4 + 4\cdot 3 + 1\cdot 2}{10} \\ &= \frac{5}{10}\cdot 4 + \frac{4}{10}\cdot 3 + \frac{1}{10}\cdot 2 \\ &= 4\cdot\frac{5}{10} + 3\cdot\frac{4}{10} + 2\cdot\frac{1}{10} \\ &= 3.4\end{aligned}$$

Our definition of the mean of a random variable follows the example above. Notice that we can think of the GPA as a sum of terms of the form

(grade)(proportion of students getting that grade) .

Since the limiting proportion of outcomes that have a particular value is the probability of that value, we are led to the following definition.

Definition 5.4.1 (mean)**.** Let X be a discrete random variable with pmf f. The **mean** (also called **expected value**) of X is denoted as μ_X or $\mathrm{E}(X)$ and defined by

$$\mu_X = \mathrm{E}(X) = \sum_x x \cdot f(x) \, .$$

The sum is taken over all possible values of X.

Example 5.4.2

Q. If we flip four fair coins and let X count the number of heads, what is $\mathrm{E}(X)$?

A. If we flip four fair coins and let X count the number of heads, then the distribution of X is described by the following table. (Note that $X \sim \mathsf{Binom}(4, .5)$.)

value of X	0	1	2	3	4
probability	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

So the expected value is

$$0 \cdot \frac{1}{16} + 1 \cdot \frac{4}{16} + 2 \cdot \frac{6}{16} + 3 \cdot \frac{4}{16} + 4 \cdot \frac{1}{16} = 2$$

On average we get 2 heads in 4 tosses. This is certainly in keeping with our informal understanding of the word average.

More generally, the mean of a binomial random variable is found by the following Theorem.

Theorem 5.4.2. Let $X \sim \mathsf{Binom}(n, \pi)$. Then $\mathrm{E}(X) = n\pi$.

Similarly, the mean of a hypergeometric random variable is just what we think it should be.

Theorem 5.4.3. Let $X \sim \mathsf{Hyper}(m, n, k)$. Then $\mathrm{E}(X) = km/(m+n)$.

The following example illustrates the computation of the mean for a hypergeometric random variable.

```
> x=c(0:5)
> p=dhyper(x,m=4,n=25,k=5)
> sum(x*p)
[1] 0.6896552
> 4/29 * 5
[1] 0.6896552
```

5.4.2 The Mean of a Continuous Random Variable

If we think of probability as mass, then the expected value for a discrete random variable X is the center of mass of a system of point masses where a mass $f_X(x)$ is placed at each possible value of X. The expected value of a continuous random variable should also be the center of mass where the pdf is now interpreted as density.

Definition 5.4.4 (mean)**.** Let X be a continuous random variable with pdf f. The **mean** of X is defined by

$$\mu_X = \mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx \,.$$

Example 5.4.3

Recall the pdf in Example 5.3.2: $f(x) = \begin{cases} 3x^2 & x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$. Then

$$\mathrm{E}(X) = \int_0^1 x \cdot 3x^2\, dx = 3/4 \,.$$

The value 3/4 seems plausible from the graph of f.

We compute the mean of two of our favorite continuous random variables in the next Theorem.

Theorem 5.4.5.

1. If $X \sim \mathsf{Unif}(a, b)$ then $\mathrm{E}(X) = (a + b)/2$.
2. If $X \sim \mathsf{Exp}(\lambda)$ then $\mathrm{E}(X) = 1/\lambda$.

Proof. The proof of each of these is a simple integral. These are left to the reader. □

Our intuition tells us that in a large sequence of trials of the random process described by X, the sample mean of the observations should be usually be close the mean of X. This is in fact true and is known as the **Law of Large Numbers**. We will not state that law precisely here but we will illustrate it using several simulations in R.

```
> r=rexp(100000,rate=1)
> mean(r)                          # should be 1
[1] 0.9959467
> r=runif(100000,min=0,max=10)
> mean(r)
```

```
[1] 5.003549                          # should be 5
> r=rbinom(100000,size=100,p=.1)
> mean(r)
[1] 9.99755                           # should be 10
> r=rhyper(100000,m=10,n=20,k=6)
> mean(r)
[1] 1.99868                           # should be 2
```

5.4.3 Transformations of Random Variables

After collecting data, we often transform it. That is we apply some function to all the data. For example, we saw the value of using a logarithmic transformation to linearize some bivariate relationships. Now consider the notion of transforming a random variable.

Definition 5.4.6 (transformation)**.** Suppose that t is a function defined on all the possible values of the random variable X. Then the random variable $t(X)$ is the random variable that has outcome $t(x)$ whenever x is the outcome of X.

If the random variable Y is defined by $Y = t(X)$, then Y itself has an expected value. To find the expected value of Y, we would need to find the pmf or pdf of Y, $f_Y(y)$, and then use the definition of $\mathrm{E}(Y)$ to compute $\mathrm{E}(Y)$. Occasionally, this is easy to do, particularly in the case of a discrete random variable X.

Example 5.4.4

Suppose that X is the random variable that results when a single die is rolled and the number on its face recorded. The pdm of X is $f(x) = 1/6$, $x = 1, 2, 3, 4, 5, 6$, and $\mathrm{E}(X) = 3.5$. Now suppose that for a certain game, the value $Y = X^2$ is interesting. Then the pdm of Y is easily seen to be $f(y) = 1/6$, $y = 1, 4, 9, 16, 25, 36$, and $E(Y) = 15.2$. Note that to find $\mathrm{E}(Y)$ we first found the pdm of Y and then found $\mathrm{E}(Y)$ using the usual method. Note that $\mathrm{E}(Y) \neq [\mathrm{E}(X)]^2$!

It turns out that there is a way to compute $\mathrm{E}(t(X))$ that does not require us to first find f_Y. This is especially useful in the case that X is continuous.

Lemma 5.4.7. If X is a random variable (discrete or continuous) and t a function defined on the values of X, then if $Y = t(X)$ and X has pdf (pmf) f_X

$$\mathrm{E}(Y) = \begin{cases} \sum_x t(x) f_X(x) \,, & \text{if } X \text{ is discrete} \,, \\ \int_{-\infty}^{\infty} t(x) f(x)\, dx \,, & \text{if } X \text{ is continuous} \,. \end{cases}$$

We will not give the proof but it is easy to see that this lemma should be so (at least for the discrete case) by looking at an example.

Example 5.4.5

Let X be the result of tossing a fair die. X has possible outcomes $1, 2, 3, 4, 5, 6$. Let Y be the random variable $|X - 2|$. Then the lemma gives

$$\mathrm{E}(Y) = \sum_{x=1}^{6} |x-2| \cdot \frac{1}{6} = 1 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} = \frac{11}{6} .$$

But if we can also compute $\mathrm{E}(Y)$ directly from the definition. Noting that the possible values of Y are $0, 1, 2, 3, 4$, we have

$$\mathrm{E}(Y) = \sum_{y=0}^{4} y f_Y(y) = 0 \cdot \frac{1}{6} + 1 \cdot \frac{2}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} = \frac{11}{6} .$$

The sum that computes $\mathrm{E}(Y)$ is clearly the same sum as $\mathrm{E}(X)$ but in a "different order" and with some terms combined since there are more than one x that produce a given value of Y.

Example 5.4.6

Suppose that $X \sim \mathsf{Unif}(0, 1)$ and that $Y = X^2$. Then

$$\mathrm{E}(Y) = \int_0^1 x^2 \cdot 1 \, dx = 1/3 .$$

This is consistent with the following simulation.

```
> x=runif(1000,0,1)
> y=x^2
> mean(y)
[1] 0.326449
```

While it is not necessarily the case that $\mathrm{E}(t(X)) = t(\mathrm{E}(X))$ (see problem 5.32), the next proposition shows that the expectation function is a "linear operator."

Lemma 5.4.8. If a and b are real numbers, then $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$.

5.4.4 The Variance of a Random Variable

We are now in a position to define the variance of a random variable. Recall that the variance of a set of n data points $x_1, \ldots, x_n$ is almost the average of the squared-deviation from the sample mean.

$$\text{Var}(x) = \sum (x_i - \overline{x})^2/(n-1)$$

The natural analogue for random variables is the following.

Definition 5.4.9 (variance, standard deviation of a random variable)**.** Let X be a random variable. The **variance** of X is defined by

$$\sigma_X^2 = \text{Var}(X) = \text{E}((X - \mu_X)^2) \, .$$

The **standard deviation** is the square root of the variance and is denoted σ_X.

The following lemma records the variance of several of our favorite random variables.

Lemma 5.4.10.

1. If $X \sim \mathsf{Binom}(n, \pi)$ then $\text{Var}(X) = n\pi(1-\pi)$.
2. If $X \sim \mathsf{Hyper}(m, n, k)$ then $\text{Var}(X) = k\left(\frac{m}{m+n}\right)\left(\frac{n}{m+n}\right)\left(\frac{m+n-k}{m+n-1}\right)$.
3. If $X \sim \mathsf{Unif}(a, b)$ then $\text{Var}(X) = (b-a)^2/12$.
4. If $X \sim \mathsf{Exp}(\lambda)$ then $\text{Var}(X) = 1/\lambda^2$.

It is instructive to compare the variances of the binomial and the hypergeometric distribution. We do that in the next example.

Example 5.4.7

Suppose that a population has 10,000 voters and that 4,000 of them plan to vote for a certain candidate. We select 100 voters at random and ask them if they favor this candidate. Obviously, the number of voters X that favor this candidate has the distribution $\mathsf{Hyper}(4000, 6000, 100)$. This distribution has mean 40 and variance 100(.4)(.6)(.99). On the other hand, were we to treat this situation as sampling with replacement so that $X \sim \mathsf{Binom}(100, .4)$, X would have mean 40 and variance 100(.4)(.6). The only difference in the two expressions for the variance is the factor $\frac{m+n-k}{m+n-1}$ which is sometimes called the **finite population correction factor**. It should really be called the **sampling without replacement correction factor**.

■

The following lemma sometimes helps us to compute the variance of X. It also is useful in understanding the properties of the variance.

Lemma 5.4.11. Suppose that the random variable X is either discrete or continuous with mean μ_X. Then

$$\sigma_X^2 = \mathrm{E}(X^2) - \mu_X^2 \, .$$

Proof. We have

$$\sigma_X^2 = \mathrm{E}((X-\mu_X)^2) = \mathrm{E}(X^2 - 2\mu_X X + \mu_X^2) = \mathrm{E}(X^2) - 2\mu_X E(X) + \mu_X^2 = \mathrm{E}(X^2) - \mu_X^2 \, .$$

Note that we have used the linearity of E and also that $E(c) = c$ if c is a constant. □

5.5 The Normal Distribution

The most important distribution in statistics is called the normal distribution.

Definition 5.5.1 (normal distribution)**.** A random variable X has the **normal distribution** with parameters μ and σ if X has pdf

$$f_X(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \, , \qquad -\infty < x < \infty \, .$$

We write $X \sim \mathsf{Norm}(\mu,\sigma)$ in this case.

The mean and variance of a normal distribution are μ and σ^2 so that the parameters are aptly, rather than confusingly, named. `R` functions `dnorm(x,mean,sd)`, `pnorm(q,mean,sd)`, `rnorm(n,mean,sd)`, and `qnorm(p,mean,sd)` compute the relevant values.

If $\mu = 0$ and sd $= 1$ we say that X has a **standard normal distribution.** Figure 5.5 provides a graph of the density of the standard normal distribution. Notice the following important characteristics of this distribution: it is unimodal, symmetric, and can take on all possible real values both positive and negative. The curve in Figure 5.5 suffices to understand all of the normal distributions due to the following lemma.

Lemma 5.5.2. If $X \sim \mathsf{Norm}(\mu,\sigma)$ then the random variable $Z = (X-\mu)/\sigma$ has the standard normal distribution.

Proof. To see this, we show that $\mathrm{P}(a \le Z \le b)$ is computed by the integral of the standard normal density function.

$$\mathrm{P}(a \le Z \le b) = \mathrm{P}(a \le \frac{X-\mu}{\sigma} \le b) = \mathrm{P}\left(\mu + a\sigma \le X \le \mu + b\sigma\right) = \int_{\mu+a\sigma}^{\mu+b\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \, dx \, .$$

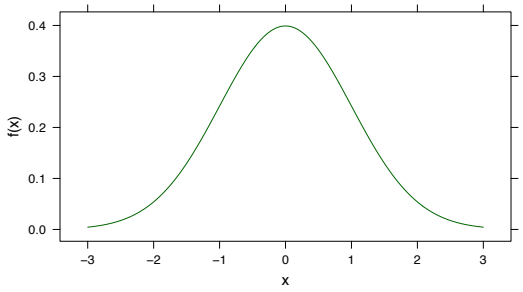


Figure 5.5: The pdf of a standard normal random variable.

Now in the integral, make the substitution $u = (x - \mu)/\sigma$. We have then that

$$\int_{\mu+a\sigma}^{\mu+b\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du\ .$$

But the latter integral is precisely the integral that computes $\mathrm{P}(a \le U \le b)$ if U is a standard normal random variable. □

The normal distribution is used so often that it is helpful to commit to memory certain important probability benchmarks associated with it.

The 68–95–99.7 Rule

If Z has a standard normal distribution, then

1. $\mathrm{P}(-1 \le Z \le 1) \approx 68\%$
2. $\mathrm{P}(-2 \le Z \le 2) \approx 95\%$
3. $\mathrm{P}(-3 \le Z \le 3) \approx 99.7\%$.

If the distribution of X is normal (but not necessarily standard normal), then these approximations have natural interpretations using Lemma 5.5.2. For example, we can say that the probability that X is within one standard deviation of the mean is about 68%.

Example 5.5.1

In 2000, the average height of a 19-year old United States male was 69.6 inches. The standard deviation of the population of males was 5.8 inches. The distribution of heights of this population is well-modeled by a normal distribution. Then the percentage of males within 5.8 inches of 69.6 inches was approximately 68%. In R,

```
> pnorm(69.6+5.8,69.6,5.8)-pnorm(69.6-5.8,69.6,5.8)
[1] 0.6826895
```

It turns out that the normal distribution is a good model for many variables. Whenever a variable has a unimodal, symmetric distribution in some population, we tend to think of the normal distribution as a possible model for that variable. For example, suppose that we take repeated measures of a difficult to measure quantity such as the charge of an electron. It might be reasonable to assume that our measurements center on the true value of the quantity but have some spread around that true value. And it might also be reasonable to assume that the spread is symmetric around the true value with measurements closer to the true value being more likely to occur than measurements that are further away from the true value. Then a normal random variable is a candidate (and often used) model for this situation.

The most important use of the normal distribution stems from the way that it arises in the analysis of repeated trials of a random experiment. This is a result of what might be called the Fundamental Theorem of Statistics — The Central Limit Theorem. Before we state the theorem, we give two examples illustrating the principles.

Example 5.5.2

Suppose that X is the result of tossing a single die and recording the number. Now suppose that we wish to toss the die 100 times and record the results, $x_1, \ldots, x_{100}$ we obtain. These data can be viewed as the result of performing 100 random processes represented by random variables $X_1, \ldots, X_{100}$ which all have the same distribution and are independent one from another. Consider now the sum $y = x_1 + \cdots + x_{100}$ of the 100 tosses. (We'd expect this number to be in the ballpark of 350, wouldn't we?) We can consider this number y to be the result of a random variable, namely $Y = X_1 + \cdots + X_{100}$. Y itself has a distribution and in theory we could write the pmf for Y. (Y is discrete with possible values $100, 101, \ldots, 599, 600$.) A simulation suggests what happens.

```
> trials10000=replicate(10000,sum(sample(c(1:6),100,replace=T)))
> summary(trials10000)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  286.0   339.0   350.0   350.2   362.0   414.0
> histogram(trials10000,xlab="Sum of 100 dice rolls")
```

Note that the histogram in Figure 5.6 suggests that Y has a distribution that is unimodal and symmetric.

Example 5.5.3

The random variable in the previous example was discrete. Suppose instead that X is a continuous random variable. For example, suppose that $X \sim \mathsf{Exp}(1)$. X might be a waiting time random variable that measures the time until the next radioactive event

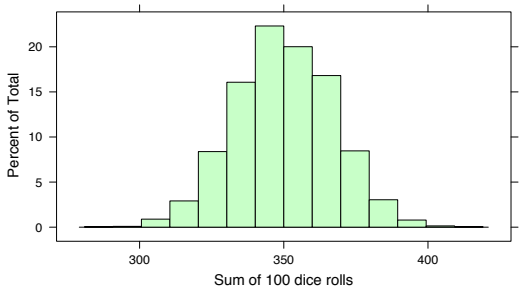


Figure 5.6: 10,000 trials of the sum of 100 dice.

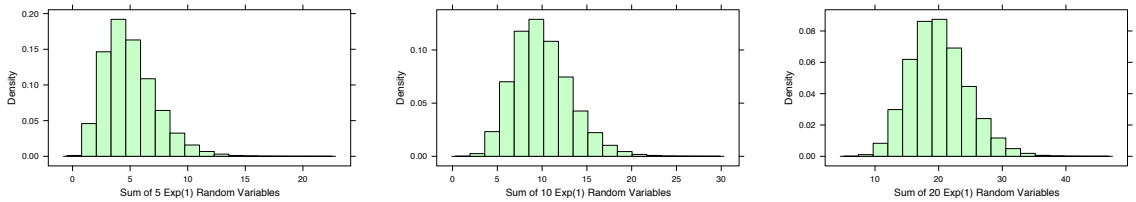


Figure 5.7: Sums of independent exponential random variables.

detected at a Geiger counter. Suppose that $X_1, \ldots, X_n$ are n independent trials of the random process X. This would be a natural model for the experiment in which we wait for not just one radioactive event but for n in succession. In this case $Y = X_1 + \cdots + X_n$ is just the time until n events have happened. The histograms in Figure 5.7 shows what might happen if $n = 5$, $n = 10$, and $n = 20$. One can see that as the number of trials of the experiment increase, the distribution of the sum becomes more symmetric.

To describe this situation in general, we note that the situation we are imagining is that we have n random variables $X_1, \ldots, X_n$ that have the same distribution and that are independent one from another (i.e., the dice don't talk to each other). We will call such random variables **i.i.d.** (for independent and identically distributed). Random variables that arise from repeating a random process and observing the same random variable are the canonical example of i.i.d. random variables. In this situation, we sometimes refer to the original random variable that describes the distribution in question as being the **population** random variable. If $X_1, \ldots, X_n$ are i.i.d. random variables, $X_1, \ldots, X_n$ are usually called a **random sample**. Note that this is the same term that we used to describe a sample from a population. These meanings are related but different. Before we state the Central Limit Theorem, we consider the properties of $Y = X_1 + \cdots + X_n$ in terms of those of X. Specifically we have

Lemma 5.5.3. Suppose that $X_1, \ldots, X_n$ are random variables and that $Y = X_1 + \cdots + X_n$. Then

1. $E(Y) = \sum_{i=1}^{n} E(X_i)$, and
2. if in addition the X_i are independent, then $\text{Var}(Y) = \sum_{i=1}^{n} \text{Var}(X_i)$, and
3. if in addition the X_i have normal distributions (and are still independent), then Y has a normal distribution.

In particular, if the X_i are i.i.d. with mean μ and variance σ^2, then $\mu_Y = n\mu$ and $\text{Var}(Y) = n\sigma^2$.

The lemma says that the sum of random variables that have normal distributions and are independent is normal. The Central Limit Theorem says that even if the X_i are not normal, if n is large the sum of the X_i is approximately normal.

Theorem 5.5.4 (Central Limit Theorem)**.** Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables with common mean μ and variance σ^2. Then as n gets large the random variable

$$Y_n = X_1 + \cdots + X_n$$

has a distribution that approaches the normal distribution $\mathsf{Norm}(n\mu, \sigma\sqrt{n})$.

Given i.i.d. random variables $X_1, \ldots, X_n$, we will often be interested in the mean of the values of the random variables rather than the sum.

Definition 5.5.5 (sample mean)**.** Given i.i.d. random variables $X_1, \ldots, X_n$ (i.e., a random sample), the **sample mean** is the random variable $\overline{X_n}$ defined by

$$\overline{X_n} = (X_1 + \cdots + X_n)/n$$

.

Corollary 5.5.6. Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables with common mean μ and variance σ^2. Then as n gets large the sample mean $\overline{X_n}$ has a distribution that is approximately normal with mean μ and variance σ^2/n.

Returning to Example 5.5.3, we have that if we find the sum Y of 10 different exponential random variables with $\lambda = 1$, the mean and variance of Y are each 10. (Recall that the mean and variance of $X \sim \mathsf{Exp}(\lambda)$ are $1/\lambda$ and $1/\lambda^2$ respectively.) Corollary 5.5.6 is especially important since we are often interested in the mean of data values $x_1, \ldots, x_n$ that can be modeled as resulting from repeating a random process n times.

Example 5.5.4

In Example 4.6.3 we considered simple random samples of size 5 from a population of 134 MIAA basketball players. We observed the points per game of each of the 5 players in our sample. We could consider the sample of size 5 that we generated as the result of 5 random variables $X_1, \ldots, X_5$. Now this sample does not fit exactly the framework of Corollary 5.5.6. Namely, these random variables are not independent. Once we choose the first player at random (X_1, a discrete random variable with 134 possibilities) the distribution of points per game of X_2 changes. This is because we generally sample without replacement. We can rectify this in two ways. First, we may sample with replacement. That guarantees that the five random variables are i.i.d. Else, we can sample with replacement but believe that the random variables X_i are close enough to being independent so as not to affect the result too much. This is especially true if the sample size (5) is much smaller than the population size (134).

5.6 Exercises

5.1 For each of the following random processes, write a complete list of all outcomes in the sample space.

a) A nickel and a dime are tossed and the resulting faces observed.

b) Two different cards are drawn from a hat containing five cards numbered 1–5 are put in a hat. (For some reason, lots of probability problems are about cards in hats.)

c) A voter in the Michigan 2008 Primary elections is chosen at random and asked for whom she voted. (See Problem B.2.)

5.2 Two six-sided dice are tossed.

a) List all the outcomes in the sample space (you should find 36) using some appropriate notation.

b) Let F be the event that the sum of the dice is 7. List the elements of F.

c) Let E be the event that the sum of the dice is odd. List the elements of the event E.

5.3 If a Calvin College student is chosen at random and his/her height is recorded, what is a reasonable listing of the possible outcomes? Explain the choices that you have to make in determining what the outcomes are.

5.4 Weathermen in Grand Rapids are fond of saying things like "The probability of snow tomorrow is 70%." What do you think this statement really means? Can you give a frequentist interpretation of this statement? A subjectivist interpretation?

5.5

a) Suppose that four coins (a penny, nickel, dime and quarter) are tossed and the face-up side of each is observed as heads or tails. How many equally likely outcomes are there? List them.

b) For each $x = 0, 1, 2, 3, 4, 5$, compute the probability that exactly x many heads occurs in the toss of *five* coins.

5.6 Suppose that ten coins are tossed.

a) How many equally likely outcomes are there? Do not list them!

b) In how many of these outcomes is exactly one head showing?

c) What is the probability that exactly one head is showing?

d) What other event would have exactly the same probability?

5.7 Use `R` to simulate the rolling of a fair six-sided die. (e.g., `sample(c(1:6),1)` will do the trick). Roll the die 600 times.

a) How many of each of the numbers 1 through 6 did you "expect" to occur?

b) How many of each of the numbers 1 through 6 actually occurred?

c) Are you surprised by the discrepancy between your answers to (a) and (b)? Why or why not?

5.8 A 20-sided die (with sides numbered 1–20) is used in some games. Obviously the die is constructed in such a way that the sides are intended to be equally likely to occur when the die is rolled. (The die is in fact an icosohedron.) Using `R`, simulate 1,000 rolls of such a die. How many of each number did you expect to see? Include a table of the actual number of times each of the 20 numbers occurred. Is there anything that surprises you in the result?

5.9 Suppose that a small class of 10 students has 4 male students and 6 female students. A random sample of two students is chosen from this class. What is the probability that both of the students are male? (Hint: first find the number of equally likely outcomes.)

5.10 A poker hand consists of 5 cards. When all 5 cards in a hand are from the same suit (♣♠♡◊), the hand is called a *flush*. Explain how one might use simulation to determine the probability of a *flush* being dealt from a well-shuffled deck. (Give the commands you use, along with an explanation of these commands and the result you obtain. Remember that there are 52 cards in a deck, 13 in each suit.)

5.11 Toss a coin 1,000 times (a simulated coin, not a real one!).

a) What is the number of heads in the 1,000 tosses? (You can do this very easily if you code heads as 1 and tails as 0.)

b) Now repeat this procedure 10,000 times (that is toss 1,000 coins 10,000 times). You now have 10,000 different answers to part (a). Don't write them all down but describe the distribution of these 10,000 numbers using the terminology and techniques for describing distributions.

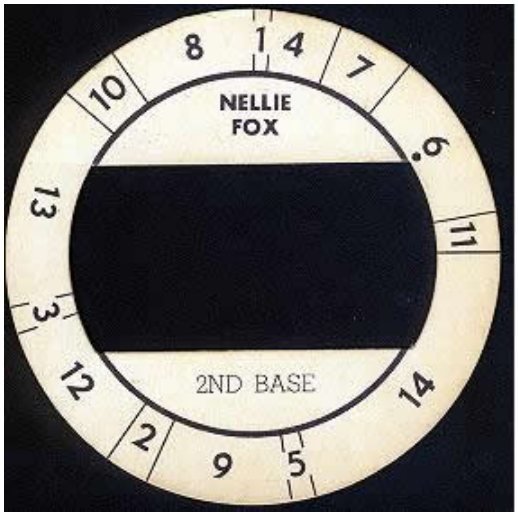
5.12 Let E^C be the event "E doesn't happen." For example, if we toss one die and E is the event that the die comes up 1 or 2, then E^C is the event that the die doesn't come up 1 or

2 (and so E^C is the event that the die comes up 3, 4, 5, or 6). Show from the axioms of probability that $P(E^C) = 1 - P(E)$.

5.13 Suppose that you roll 5 standard dice. Determine the probability that all the dice are the same.

5.14 Suppose that you deal 5 cards from a standard deck of cards. Determine the probability that all the cards are of the same color. (A standard deck of cards has 52 cards in two colors. There are 26 red and 26 black cards.)

5.15 Many games use spinners rather than dice to initiate action. A classic board game published by Cadaco-Ellis is "All-American Baseball." The game contains discs for each of several baseball players. The disk for Nellie Fox (the great Chicago White Sox second baseman) is pictured below.



The disc is placed over a peg with a spinner mounted in the center of the circle. The spinner is spun and comes to rest pointing to the one of the numbered areas. Each number corresponds to the possible result of Nellie Fox batting. (For example, 1 is a homerun and 14 is a flyout.)

a) Why is it unreasonable to believe that all the numbered outcomes are equally likely?

b) Explain how one could use the idea of equal likelihood to predict the probability that the spinner will land on the sector numbered 14 and then make an estimate of this probability.

(Spinners with regions of unequal size are used heavily in the K–8 textbook series *Everyday Mathematics* to introduce probability to younger children.)

5.16 The traditional dartboard is pictured below.

A dart that sticks in the board is scored as follows. There are 20 numbered sectors each of which has a small outer ring, a small inner ring, and two larger areas. A dart landing in the larger areas scores the number of the sector, in the outer ring scores double the number of the sector, and in the inner ring scores triple the number of a sector. The two circles near the center score 25 points (the outer one) and 50 points (the inner one). Unlike the last problem, it does not seem that an equal likelihood model could be used to compute the probability of a "triple 20." Explain why not.

5.17 Acceptance sampling is a procedure that tests some of the items in a lot and decides to accept or reject the entire lot based on the results of testing the sample. Suppose that the test determines whether an item is "acceptable" or "defective". Suppose that in a lot of 100 items, 4 are tested and that the lot is rejected if one or more of those four are found to be defective.

a) If 10% of the lot of 100 are defective, what is the probability that the purchaser will reject the shipment?

b) If 20% of the lot of 100 are defective, what is the probability that the purchaser will reject the shipment?

5.18 Suppose that there are 10,000 voters in a certain community. A random sample of 100 of the voters is chosen and are asked whether they are for or against a new bond proposal. Suppose that, in fact, only 4,500 of the voters are in favor of the bond proposal.

a) What is the probability that fewer than half of the sampled voters (i.e., 49 or fewer) are in favor of the bond proposal?

b) Suppose instead that the sample consists of 2,000 voters. Answer the same question as in the previous part.

5.19 If the population is very large relative to the size of the sample, it seems like sampling with replacement should yield very similar results to that of sampling without replacement. Suppose that an urn contains 10,000 balls, 3,000 of which are white.

a) If 100 of these balls are chosen at random **with** replacement, what is the probability that at most 25 of these are white?

b) If 100 of these balls are chosen at random **without** replacement, what is the probability that at most 25 of these are white?

5.20 In the days before calculators, it was customary for textbooks to include tables of the cdf of the binomial distribution for small values of n. Of course not all values of π could be included—often only the values $\pi = .1, , 2, \ldots, .8, .9$ were included. Let's supposes that one of these tables includes the value of the cdf of the binomial distribution for all $n \leq 25$, all $x \leq n$ and all these values of π.

a) To save space, the values of $\pi = .6, .7, .8, .9$ could be omitted. Give a clear reason why $F(x; n, \pi)$ could be computed for these values of π from the other values in the table.

b) On the other hand, we could instead omit the values of $x \geq n/2$. Show how the value of $F(x; n, \pi)$ could be computed from the other values in the table for such omitted values of x.

(Hint: one person's success is another person's failure.)

5.21 A random variable X has the triangular distribution if it has pdf $f_X(x) = \begin{cases} 2x & x \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}$

a) Show that f_X is indeed a pdf.

b) Compute $P(0 \leq X \leq 1/2)$.

c) Find the number m such that $P(0 \leq X \leq m) = 1/2$. (It is natural to call m the median of the distribution.)

5.22 Let $f(x) = \begin{cases} k(x-2)(x+2) , & -2 \leq x \leq 2 \\ 0 , & \text{otherwise.} \end{cases}$

a) Determine the value of k that makes f a pdf. Let X be the corresponding random variable.

b) Calculate $P(X \geq 0)$.

c) Calculate $P(X \geq 1)$.

d) Calculate $P(-1 \leq X \leq 1)$.

5.23 Describe a random variable that is neither continuous nor discrete. Does your random variable have a pmf? a pdf? a cdf?

5.24 Show that if f and g are pdfs and $\alpha \in [0, 1]$, then $\alpha f + (1 - \alpha)g$ is also a pdf.

5.25 Suppose that a number of measurements that are made to 3 decimal digits accuracy are each rounded to the nearest whole number. A good model for the "rounding error" introduced by this process is that $X \sim \mathsf{Unif}(-.5, .5)$ where X is the difference between the true value of the measurement and the rounded value.

a) Explain why this uniform distribution might be a good model for X.

b) What is the probability that the rounding error has absolute value smaller than .1?

5.26 If $X \sim \mathsf{Exp}(\lambda)$, find the median of X. That is find the number m (in terms of λ) such that $P(X \leq m) = 1/2$.

5.27 A part in the shuttle has a lifetime that can be modeled by the exponential distribution with parameter $\lambda = 0.0001$, where the units are hours. The shuttle mission is scheduled for 200 hours.

a) What is the probability that the part fails on the mission?

b) The event that is described in part (a) is BAD. So the shuttle actually runs three of these systems in parallel. What is the probability that the mission ends without all three failing if they are functioning independently?

c) A different (and perhaps more realistic(?)) alternative (to running three systems in parallel) would be to run one system but carry a replacement for the part (a total of two altogether). We would like to compare (to the answer in part (b)) the probability the mission ends without both parts failing under this alternative. We have not discussed the tools to evaluate this probability exactly, but you should be able to design a simulation in `R` that allows you to find it approximately. Submit the code you use, along with accompanying explanation and resulting (approximate) probability.

5.28 The lifetime of a certain brand of water heaters in years can be modeled by a Weibull distribution with $\alpha = 2$ and $\beta = 25$.

a) What is the probability that the water heater fails within its warranty period of 10 years?

b) What is the probability that the water heater lasts longer than 30 years?

c) Using a simulation, estimate the average life of one of these water heaters.

5.29 Prove Theorem 5.4.5.

5.30 Suppose that you have an urn containing 100 balls, some unknown number of which are red and the rest are black. You choose 10 balls without replacement and find that 4 of them are red.

a) How many red balls do you think are in the urn? Give an argument using the idea of expected value.

b) Suppose that there were only 20 red balls in the urn. How likely is it that a sample of 10 balls would have **at least** 4 red balls.

5.31 The file http://www.calvin.edu/~stob/data/scores.csv contains a dataset that records the time in seconds between scores in a basketball game played between Kalamazoo College and Calvin College on February 7, 2003.

a) This waiting time data might be modeled by an exponential distribution. Make some sort of graphical representation of the data and use it to explain why the exponential distribution might be a good candidate for this data.

b) If we use the exponential distribution to model this data, which λ should we use? (A good choice would be to make the sample mean equal to the expected value of the random variable.)

c) Your model of part (b) makes a prediction about the proportion of times that the next score will be within 10, 20, 30 and 40 seconds of the previous score. Test that prediction against what actually happened in this game.

5.32 Show that it is not necessarily the case that $E(t(X)) = t(E(X))$.

5.33 Let X be the random variable that results from tossing a fair six-sided die and reading the result (1–6). Since $E(X) = 3.5$, the following game seems fair. I will pay you 3.5^2 and then we will roll the die and you will pay me the square of the result. Is the game fair? Why or why not?

5.34 Not every distribution has a mean! Define

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \qquad -\infty < x < \infty\,.$$

a) Show that f is a density function. (The resulting distribution is called the **Cauchy distribution**.)

b) Show that this distribution does not have a mean. (You will need to recall the notion of an improper integral.)

5.35 In this problem we compare sampling with replacement to sampling without replacement. You will recall that the former is modeled by the binomial distribution and the latter by the hypergeometric distribution. Consider the following setting. There are 4,224 students at Calvin and we would like to know what they think about abolishing the interim. We take a random sample of size 100 and ask the 100 students whether or not they favor abolishing the interim. Suppose that 1,000 students favor abolishing the interim and the other 3,224 misguidedly want to keep it.

a) Suppose that we sample these 100 students with replacement. What is the mean and the variance of the random variable that counts the number of students in the sample that favor abolishing the interim?

b) Now suppose that we sample these 100 students without replacement. What is the mean and the variance of the random variable that counts the number of students in the sample that favor abolishing the interim?

c) Comment on the similarities and differences between the two. Give an intuitive reason for any difference.

5.36 Scores on IQ tests are scaled so that they have a normal distribution with mean 100 and standard deviation 15 (at least on the Stanford-Binet IQ Test).

a) MENSA, a society supposedly for persons of high intellect, requires a score of 130 on the Stanford-Binet IQ test for membership. What percentage of the population qualifies for MENSA?

b) One psychology text labels those with IQs of between 80 and 115 as having "normal intelligence." What percentage of the population does this range contain?

c) The top 25% of scores on an IQ test are in what range?

d) If two different individuals are chosen at random, what is the probability that the sum of their IQ scores is greater than 240?

5.37 In this problem we investigate the accuracy of the Central Limit Theorem by simulation. Suppose that X is a random variable that is exponential with parameter $\lambda = 1/2$. Then $\mu_X = 2$ and $\sigma_X^2 = 4$. Suppose that we repeat the random experiment n times to get independent random variables $X_1, \ldots, X_n$ each of which is exponential with $\lambda = 1/2$. (The `R` function `rexp(n,lambda=.5)` will simulate this experiment.)

a) From Lemma 5.5.3 $\overline{X} = (X_1 + \cdots + X_5)/5$ has what mean and variance?

b) If $n = 5$, what does the Central Limit Theorem predict for $P(1.5 < \overline{X} < 2.5)$?

c) Simulate the distribution of $\overline{X}$ by taking many samples of size 5. Compute the proportion of your samples for which $(1.5 < \overline{x} < 2.5)$ and compare to part (b).

d) Repeat parts (b) and (c) for $n = 30$.

5.38 In this problem you are to investigate the accuracy of the approximation of the Central Limit Theorem for the exponential distribution. Suppose that $X \sim \mathsf{Exp}(0.1)$, and that a random sample of size 20 is chosen from this population.

a) What are the mean and variance of X?

b) What are the mean and variance of $\overline{X}$?

c) Using the central Limit Theorem approximation, compute the probability that $\overline{X}$ is within 1, 2, 3, 4, and 5 of μ_X.

d) Now choose 1,000 random samples of size 20 from this distribution. Count the number of samples in which $\overline{x}$ is within 1, 2, 3, 4 and 5 of μ_X, and compare to part (c). Comment on what you find.

5.39 Scores on the SAT Test were redefined (recentered) in 1990 and were set to have mean 500 and standard deviation 110 on each of the Mathematics and Verbal tests. The scores were constructed so that the population had a normal distribution (or at least very close to normal). In a random sample from this population of size 500,

a) What is the probability that the sample mean will be between 490 and 510?

b) What is the probability that the sample mean will exceed 500? 510? 520?

5.40 Continuing Exercise 5.39, the total SAT score for each student is formed by adding his verbal score V and his math score M.

a) If the two scores for an individual are independent of each other, what are the mean and standard deviation of $V + M$?

b) It is not likely that the verbal and mathematics scores of individuals in the population behave like *independent* random variables. Do you expect that the standard deviation of $V + M$ is more or less than you computed in part (a)? Why?