MATH 340 Notes

Mike Hitchman

Fall, 2024

Contents

Chapter 1

Introduction

MATH 340 introduces us to probability and statistics. The prerequisite for the course is MATH 175: Calculus II, a course that gets us through integration techniques and series. We will integrate, and we will evaluate series in this course. Multivariable calculus is not required in MATH 340, though it is required for the second course in this sequence, MATH 440.

Content in these notes is tied to two classic texts, *Introduction to Probability*, by Grinstead and Snell; and *Mathematical Statistics with Applications, 7th ed.*, by Wackerly, Mendenhall, and Scheaffer. This content also happens to coincide with content one finds in the first actuarial exam (https://www.beanactuary.com).

Probability theory is of fundamental importance in the field of statistics. Suppose we roll a die five times and a 4 comes up all 5 times!!!

- A **probability question**: What is the likelihood of rolling five 4s in a row if the die is fair?
- A **statistics question**: Is this die fair?

We have two approaches to answering likelihood questions:

1. simulation (repeat the experiment many, many times and see how often the desired result is obtained)
2. probability theory

We study probability theory in this course to answer likelihood questions without simulation. This study develops intuition and a rigorous foundation for the subject. We also learn simulation techniques - using R - because simulation can produce approximate solutions quite easily when exact solutions are beyond our grasp.

After studying probability up through the Central Limit Theorem, we will practice statistics. We will investigate estimation, hypothesis testing, and, time permitting, an introduction to linear models. The second term of this sequence, MATH 440, continues the study of mathematical statistics.

Chapter 2

Sets

We use sets to build probability models for chance experiments and to communicate features about these models. To help us manage all this effectively, we begin this course with set theory.

Think of a set as a collection of elements.

Example 2.1. We regularly encounter the following sets:

- $\mathbb{N} = \{1, 2, 3, ...\}$, the set of **natural numbers**
- $\mathbb{Z} = \{... , -2, -1, 0, 1, 2, ...\}$, the set of **integers**,
- $\mathbb{R}$, the set of **real numbers**,
- $I = [0, 1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$, the **unit interval**,
- $\emptyset$, the **empty set**, the set with no elements.

To indicate whether item x is an element of set A, we write

- $x \in A$ if x is an element of A, and
- $x \notin A$ if x is not an element of A.

Definition 2.1. Let A and B be sets. We say A is a **subset** of B, denoted $A \subseteq B$, if $x \in A$ implies $x \in B$; and A is a **proper subset** of B, denoted $A \subset B$, if $A \subseteq B$ and there is some element $x \in B$ such that $x \notin A$. We say A and B are **equal**, denoted $A = B$, if $A \subseteq B$ and $B \subseteq A$.

Example 2.2. Here are a few sets (we usually name our sets with capital letters).

- $A = \{2, 4, 6, 8, ...\}$, the set of even natural numbers.
- $B = \{8, 4\}$ is a set with my two favorite natural numbers.
- $C = \{x \in \mathbb{R} \mid |x - 3| > 2\}$. This set consists of all real numbers x whose distance from 3 is greater than 2.
- $S = \{$all spiders on Earth alive today$\}$.

Observe:

- $18 \in A$ and $19 \notin A$.
- $A, B \subset \mathbb{N}$, and $C \subset \mathbb{R}$, and $B \subset A$.
- $6.7 \in C$, and $4.1 \notin C$.
- $\emptyset \subset A$. In fact $\emptyset \subseteq X$ for *any* set X.
- S is a set I'd rather not encounter all at once.

2.1 Algebra of Sets

Definition 2.2. Given sets A and B, the **union** of A and B is the set

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

The **intersection** of A and B is the set

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

The **difference** of A and B is the set

$$A - B = \{x \mid x \in A \text{ and } x \notin B\}.$$

Example 2.3. Let $A = \{2, 4, 6, 8\}$ and $B = \{0, 1, 2, 3, 5, 8\}$.

Then $A \cup B$ gives the set of all elements in A or B (or both):

$$A \cup B = \{0, 1, 2, 3, 4, 5, 6, 8\}.$$

The set $A \cap B$ gives those elements that are in both A and B:

$$A \cap B = \{2, 8\},$$

and $A - B$ gives the set of elements in A that aren't in B:

$$A - B = \{4, 6\}.$$

Note that $(A \cap B) \cup (A - B) = A$, something that will be true for any two sets.

Definition 2.3. If $A \subseteq U$ where U is viewed as a universal set, the **complement of A** in U, denoted $\overline{A}$, consists of those elements in U that are not in A:

$$\overline{A} = \{x \in U \mid x \notin A\}.$$

Example 2.4. Two examples of complements:

a) If $E = \{2, 4, 6, ...\}$ is the set of even natural numbers, then, viewed in the universe of all natural numbers $\mathbb{N}$,

$$\overline{E} = \{1, 3, 5, ...\},$$

the set of odd natural numbers.

b) Suppose our universe is the unit inveral $[0, 1]$. The complement of the open interval $A = (0.3, 0.7)$ in this universe is the union of two closed intervals:

$$\overline{A} = [0, 0.3] \cup [0.7, 1].$$

Definition 2.4. If $A \cap B = \emptyset$, then A and B are called **disjoint** sets. Disjoint sets have no common elements. For $k \geq 2$, the sets $A_1, A_2, \dots, A_k$ are called **pairwise disjoint** if all pairs of sets in this collection are disjoint.

Theorem 2.1. *Let A, B, and C be sets, viewed in a universal set U. Then*

1. $A \cap \overline{A} = \emptyset$ *and* $A \cup \overline{A} = U$.

2. ***Distributive Laws***

a) $A \cap (B \cup C) = (A \cup B) \cap (A \cup C)$.
b) $A \cup (B \cap C) = (A \cap B) \cup (A \cap C)$.

3. ***De Morgan's Laws***

a) $\overline{A \cup B} = \overline{A} \cap \overline{B}$.
b) $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

Proof. We prove 3a), the first De Morgan's Law, by showing that an arbitrary element of either set belongs to the other set (so each set is a subset of the other).

$$\begin{aligned}
x \in \overline{A \cup B} &\iff x \notin A \cup B && \text{by def'n of complement}\\
&\iff x \notin A \text{ and } x \notin B && \text{by def'n of union}\\
&\iff x \in \overline{A} \text{ and } x \in \overline{B} && \text{by def'n of complement}\\
&\iff x \in \overline{A} \cap \overline{B} && \text{by def'n of intersection}
\end{aligned}$$

It follows that $\overline{A \cup B} \subseteq \overline{A} \cap \overline{B}$ and $\overline{A} \cap \overline{B} \subseteq \overline{A \cup B}$, so the two sets are equal. □

2.2 Set sizes

For a finite set A, we let $|A|$ denote the number of elements in A. Note that in Example 2.3, $|A| = 8, |B| = 9, |A \cup B| = 15$, and $|A \cap B| = 2$.

Theorem 2.2. *Let A and B be finite sets. Then*

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

We omit the proof here.

An infinite set is called **countably infinite** if its elements can be counted, i.e., can be put in one-to-one correspondence with the positive integers

$$\mathbb{N} = \{1, 2, 3, 4, ...\}.$$

An infinite set is called **uncountable** if it is not countably infinite.

The set of positive even integers $\{2, 4, 6, ...\}$ is countably infinite, and the unit interval I is uncountable.

We distinguish between these two types of infinite sets in this class because when an infinite set represents the possible outcomes of some random process, the type of probability model we apply to the situation depends on whether this set is countably infinite or uncountable.

We will study two types of probability distributions in this class: discrete distributions, and continuous distributions. We use discrete distributions to model a random process in which the set of outcomes is either finite or countably infinite. We use continuous distributions when the random process of interest has an uncountable set of possible outcomes (which is generally an interval of real numbers in this class).

For instance, if we flip a coin and are interested in how many flips it takes to get our 100th heads, the set of possible outcomes for this experiment is countably infinite (it might take n flips, for any integer $n \geq 100$), and we will use a discrete distribution to model probability in this setting. On the other hand, if we are interested in how far we can throw the coin in frustration after 1000 flips, the set of possible outcomes is better described as an interval in the real line (maybe the interval $(0, \infty)$, units in feet). Since intervals are uncountable sets, we would model probability in this setting with a continuous distribution.

In this class we first study discrete distributions before turning to continuous distributions.

2.3 Sets in R

We generally define a finite set in R as a structure called a data vector. Appendix A.1 dives into data vectors in R, with an eye toward sampling, but we can also use R to perform set operations. We define a data vector via the `c()` command in R. Here are two sets A and B:

```r
A = c(2,4,8)
B = c(2,4,9,12)
```

Basic set operations in R:

- `length(A)` $= 3$ returns $|A|$, the size of A.
- `union(A,B)` $= 2, 4, 8, 9, 12$ gives $A \cup B$
- `intersect(A,B)` $= 2, 4$ gives $A \cap B$
- `setdiff(A,B)` $= 8$ gives $A - B$
- `setequal(A,B)` = FALSE asks whether $A = B$ (returns TRUE or FALSE)
- `is.element(3,A)` = FALSE asks whether $3 \in A$ (returns TRUE OR FALSE)

Chapter 3

Discrete Probability Distributions

In this chapter we develop basic notions of a probability model for a chance experiment.

3.1 Sample Space

A chance experiment is some repeatable process whose outcome on any given trial cannot be known ahead of time. Here are a few examples of chance experiments:

1. Flip a coin.
2. Roll a 6-sided die.
3. Flip a coin three times.
4. Shoot free throws until we've made three.
5. Count "scintilations" in 72 second intervals caused by radioactive decay of a quantity of polonium (Rutherford and Geiger).

Definition 3.1. The **sample space** of a chance experiment is the set of possible basic outcomes of the experiment. The elements of a sample space are called **sample points** or **simple events**, and any subset of a sample space is called an **event**.

We often have some chioce in how to record the possible outcomes of a chance experiment. For instance, we might record the sample spaces for the experiments above as follows:

1. $S = \{H, T\}$ (H for heads, T for tails).
2. $S = \{1, 2, 3, 4, 5, 6\}$ (recording the value that is face up after rolling the die.)
3. $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$ (record the result of each flip in order). Alternatively, we might just record how many heads we flipped, in which case $S = \{0, 1, 2, 3\}$, but we lose some information about the experiment in doing so.
4. $S = \{111, 1101, 1011, 0111, 11001, 10101, 01101, 10011, 01011, 00111, ...\}$, where 0 represents missing a shot, and 1 represents making a shot.
5. $S = \{0, 1, 2, 3, 4, ...\}$.

In the first three examples, S is a finite set, while S appears to be an infinite set in the last two examples. There is, of course, a limit to how many free throws I can attempt in my life (if I shoot one free throw every 15 seconds for 100 years, that's only about 210 million attempts :)), but, in the context of building a probability model to describe the chance experiment of shooting free throws until I've made three, I have no reason to limit how many attempts I need to get that done.

Although infinite, the sample spaces in the last two examples are *countably* infinite. Recall, a set is **countably infinite** if its elements can be counted, i.e., can be put in one-to-one correspondence with the positive integers.

Definition 3.2. The sample space of a chance experiment is called **discrete** if the sample space is finite or countably infinite.

If you asked me to pick a random real number from the unit interval $I = [0, 1]$, this is a chance experiment with an uncountable sample space, and something we are not considering in this chapter. We focus on such games in Chapter 9.

Definition 3.3. Given a chance experiment with discrete sample space S, a **probability distribution function** on the elements of S is a real-valued function m which satisfies these two conditions:

1. $m(s) \geq 0$ for all $s \in S$, and
2. $\displaystyle\sum_{s \in S} m(s) = 1$.

We define the probability of any event E of S to be

$$P(E) = \sum_{s \in E} m(s).$$

Let's consider our first three chance experiments once more.

1. If we flip a fair coin once, then $S = \{H, T\}$, and it is reasonable to assign the probabilities
$$m(H) = \frac{1}{2},\ m(T) = \frac{1}{2}.$$

2. If a 6-sided die is balanced, it is reasonable to assign the probabilities
$$m(i) = \frac{1}{6}$$
for each $i = 1, 2, 3, 4, 5, 6$.

3. If we flip a fair coin 3 times, it seems reasonable that each of the 8 possible sequences of three flips in S is equally likely, so we can assign the probability distribution function $m(s) = 1/8$ for each element $s \in S$.

In the case of a countably infinite sample space (such as shooting free throws until we've made three), defining a valid probability function requires more care: to check that the sum of all $m(s)$ equals 1 requires the evaluation of an infinite series.

3.2 Discrete Random Variables

Definition 3.4. A **discrete random variable** is a real-valued function defined over a discrete sample space. We usually let X or Y denote a random variable. Given random variable X, the **space of** X is the set of possible outcomes for X.

Example 3.1 (Flip a coin 3 times)**.** Consider the experiment of flipping a coin three times. We record as much information as possible about this experiment by providing the sequence of the results of the three flips. Thus, the sample space for this experiment is:

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

We might be interested in knowing how many times we flipped heads, or perhaps we want to know whether we ever flipped heads twice in a row. We can use random variables to keep track of these sorts of things.

Let

$$X = \text{the number of heads in three flips.}$$

Note that the space of X is the set $\{0, 1, 2, 3\}$ (we can get anywhere between 0 and 3 heads in 3 flips).

Or, if we're interested in whether we ever flipped consecutive heads in our 3 flips, we could let

$$Y = \begin{cases} 1 & \text{if we ever flipped consecutive heads} \\ 0 & \text{else.} \end{cases}$$

The space of Y is $\{0, 1\}$.

Again, formally, the random variables X and Y are functions whose inputs are elements in S, and whose outputs are real numbers. We can display these functions in table form when the sample space is small, as in Table 3.1.

Table 3.1: Random variables X and Y associated to the event of flipping a coin 3 times.

S	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
X	3	2	2	2	1	1	1	0
Y	1	1	0	1	0	0	0	0

If X is a random variable associated to an experiment, and we have a probability distribution function assigned to the sample space S, we can naturally ask about the probability that X takes on a particular value x.

Definition 3.5. The probability that a random variable X takes on value x, denoted $P(X = x)$ or $p(x)$, is defined as the sum of the probabilities of all sample points in S that are assigned the value x. The function $p(x)$ is called the **distribution function** of the discrete random variable X, and the **probability distribution** of X refers to the the list of possible values for x along with their associated probabilities $p(x)$ (usually given as a table or function).

Example 3.2 (Flip a coin 3 times (Part II))**.** Consider again the "flip a coin three times" Example 3.1 and the associated random variables X and Y, which counted the number of heads flipped, and whether we flipped consecutive heads, respectively. Table 3.1 provides the values for these random variables.

We assume $m(s) = 1/8$ for each $s \in S$ (all 8 sequences are equally likely), so we have the following probability distributions:

x	0	1	2	3
$p(x)$	1/8	3/8	3/8	1/8

and

y	0	1
$p(y)$	5/8	3/8

Example 3.3 (Rolling Two Dice)**.** The chance experiment of rolling two regular 6-sided dice is a staple of the board game industry. A convenient way to describe the sample space in this setting is to treat the dice as distinct (say, one red die and one blue die), and write down all possible pairs of values (r, b) where r is the red die value, b is the blue die value. The sample space for rolling two 6-sided dice thus has 36 elements, which we can describe via a 6×6 grid.

Table 3.2: The sample space for rolling two dice

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

We may be interested in X, the sum of the two dice. The 6×6 grid is handy for representing this random variable:

Table 3.3: X, the sum of two dice

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Assuming the probability of each element in S is 1/36, the probability distribution for X is

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

More succinctly, we have

$$p(x) = \frac{6 - |x - 7|}{36} \quad \text{for } x = 2, 3, \dots, 12.$$

Maybe we're interested in how far apart the two values are, so we consider the random variable Y equal to the absolute value of the difference of the two dice:

Table 3.4: Y, the absolute value of the difference of two dice

	1	2	3	4	5	6
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

So, the probability distribution for Y is

y	0	1	2	3	4	5
$p(x)$	6/36	10/36	8/36	6/36	4/36	2/36

3.3 Calculating Probabilities

Recall the scene:

1) We conduct a chance experiment, to which we associate the sample space S of possible outcomes.
2) To each sample point s in S we assign a reasonable probability, $m(s)$, that s occurs (being sure that all $m(s)$ are non-negative and that they sum to 1).
3) For any event A associated to this experiment (i.e., A is a subset of S), we define $P(A) = \sum_{s \in A} m(s)$.
4) For a random variable X associated to S, $P(X = x)$ equals the sum of the $m(s)$ for which s is assigned value x.

3.3.1 Sample Point Method

So far we have been finding probability distributions by following what is called the **sample-point method** (list all the sample points, assign probabilities to each, and go!).

Here's one more example of finding probabilities via the sample-point method.

Example 3.4 (Random Phones)**.**

Four phones are found in a classroom after class. The professor returns them at random to the four students the next class. Let X denote the number of students who receive the correct phone. Let's determine the probability distribution for X by the sample-point method.

The chance experiment here is straight-forward: randomly return 4 phones to the 4 students who own them. We list the basic outcomes as follows:

- Name the students “a”, “b”, “c”, and “d”, and name their phones by the same letter (student “a” owns phone “a”, etc).
- Return the phones randomly to the students so that “a” receives the first phone, “b” the second, and so on.
- record the results of the experiment by writing down the phone names in the order in which they were returned.

- For instance, recording “c b a d” would mean student a received phone c, student b received phone b (their own phone!), student c received phone a, and student d received their own phone, d.

In this way, the 24 different permutations of the letters “a b c d” listed in Table 3.5 correspond to the 24 basic outcomes possible in this experiment. For each basic outcome in the table we also record X, the number of students to receive their own phone for that basic outcome.

Table 3.5: Returning 4 phones at random, X counts how many students receive their own phone.

a	**b**	**c**	**d**	**X**		**a**	**b**	**c**	**d**	**X**
a	b	c	d	4		b	a	c	d	2
a	b	d	c	2		b	a	d	c	0
a	c	b	d	2		b	c	a	d	1
a	c	d	b	1		b	c	d	a	0
a	d	b	c	1		b	d	a	c	0
a	d	c	b	2		b	d	c	a	1
c	a	b	d	1		c	b	a	d	2
c	a	d	b	0		c	b	d	a	1
c	d	a	b	0		c	d	b	a	0
d	a	b	c	0		d	b	a	c	1
d	a	c	b	1		d	b	c	a	2
d	c	a	b	0		d	c	b	a	0

If the professor truly returns the phones at random, each of the 24 possible outcomes is equally likely. In other words, for each element s in the sample space S, $m(s) = 1/24$. It follows that the probability distribution for X is

x	0	1	2	3	4
$p(x)$	9/24	8/24	6/24	0	1/24

It looks like the most likely scenario upon returning the phones at random is that no one gets their phone back, and there is about a 4 percent chance that everyone gets their phone back.

This sample-point method for determining probabilities will not be much help if we have a huge sample space, and huge sample spaces arise easily, such as in a friendly game of cards. We examine 5-card poker hands later, beginning with Example 4.7, but mention here that a player can be dealt about 2.6 million possible 5-card hands from a regular 52 card deck. So, in an effort to determine the probability of obtaining a particular type of hand, say a 3 of a kind, I will not be using the sample-point method!

We have two alternatives to the sample-point method:

- simulation (draw 5 cards at random many, many times, and see how often you get a 3 of a kind).
- learn counting techniques in Chapter 4!!

Chapter 4

Counting Techniques

Here we develop a toolbox of counting techniques to help us calculate probabilities.

4.1 Multipiclation Principle

Proposition 4.1 (Addition Principle)**.** *Let A and B be disjoint sets with m and n elements, respectively. Then the total number of elements of $A \cup B$ is $m + n$.*

The addition principle extends to any number of pairwise disjoint sets: The size of the union of pairwise disjoint sets equals the sum of the individual set sizes.

We use the addition principle when we count the size of a set by first breaking the set into disjoint subsets and then counting the size of each subset. The addition principle just says that the size of the original set is found by adding the sizes of these disjoint subsets.

Here's a simple example to illustrate the point:

How many Major League Baseball teams are there?

Ok, I can do this! Major League Baseball (MLB) is the (dijoint) union of two leagues, the American League (AL), which has 15 teams, and the National League (NL), which also has 15 teams. So all of Major League Baseball has $15 + 15 = 30$ teams!

Proposition 4.2 (Multiplication Principle)**.** *Given a set A with m elements, and a set B with n elements, it is possible to form $m \cdot n$ pairs containing one element from each set.*

A simple illustration of the multiplication principle:

In Major League Baseball, the world series is a best-of-7 series between the champion of the AL and the champion of the NL.

How many different world series matchups are possible?

We have 15 possible AL champions and 15 possible NL champions, so we have $15 \cdot 15 = 225$ possible world series matchups (none of which, in all the years past, have included the Seattle Mariners).

The multiplication principle is the hammer of our counting toolbox.

We often use this hammer in the following manner: Suppose a task is completed by completing k subtasks. If the subtasks can be completed in $n_1, n_2, \dots, n_k$ ways, respectively, then the task itself can be completed in $n_1 \cdot n_2 \cdot \dots \cdot n_k$ ways.

The World Series example above fits this mold. We can think of determining the World Series matchup as our task, which we complete by completing two subtasks: (1) Choose the AL champion (15 choices); and (2) Choose the NL champion (15 choices). So we can build ourselves a World Series matchup in 15^2 ways.

Let's look at several more examples.

Example 4.1.

a) A menu at a restaurant has 5 salads, 7 main dishes, and 4 desserts. If a dinner consists of ordering a salad, main dish, and dessert (because you're hungry), how many different dinners are possible?

- Subtask 1: order a salad (5 choices);
- Subtask 2: order a main dish (7 choices);
- Subtask 3: order a dessert (4 choices).

So we have $5 \cdot 7 \cdot 4 = 140$ possible dinners. If we go to this restaurant once a week, it will take about 2.7 years to try every possible dinner.

b) Suppose a license plate consists of six characters, where each character can be a letter (A-Z) or a digit (0-9). How many different license plates are there?

Here's a blank license plate, needing to be created:

$$\underline{\quad}\ \ \underline{\quad}\ \ \underline{\quad}\ \ \underline{\quad}\ \ \underline{\quad}\ \ \underline{\quad}$$

To create the plate, we pick a character for each of the six spots. We have 36 choices at each stage, so the number of distinct plates is

$$\underline{36} \cdot \underline{36} \cdot \underline{36} \cdot \underline{36} \cdot \underline{36} \cdot \underline{36} = 36^6 = 2,176,782,336,$$

just shy of 2.18 billion.

c) How many 7-digit phone numbers are there, assuming the first digit cannot be 0 or 1?

Count our digit choices at each stage in the process of creating a valid number, and multiply our choices:

$$\underline{8} \cdot \underline{10} \cdot \underline{10} \cdot \underline{10} \cdot \underline{10} \cdot \underline{10} \cdot \underline{10} = 8 \cdot 10^6,$$

8 million on the nose. If we have more than 8 million phones in an area, we need more than one area code.

d) A baseball team has 13 batters. How many different batting lineups of 9 players are possible?

We have to create a lineup with 9 players, and the total number of lineups possible will be found by multiplying our choices at each stage. Since we can't pick the same player twice, the number of choices decreases by one at each stage in the selection process:

$$\underline{13} \cdot \underline{12} \cdot \underline{11} \cdot \underline{10} \cdot \underline{9} \cdot \underline{8} \cdot \underline{7} \cdot \underline{6} \cdot \underline{5} = 259,459,200.$$

Over quarter of a billion possible lineups? The season isn't quite long enough to test out every possible lineup.

e) How many 4-digit integers bigger than 5000 have distinct odd digits?

Record our choices as we set about building such a four-digit number:

$$\underline{\quad} \quad \underline{\quad} \quad \underline{\quad} \quad \underline{\quad}$$

Each digit must be odd (1, 3, 5, 7, or 9), and since the number must be bigger than 5000, we only have 3 choices for the "thousands place" (5,7, or 9). Once that has been chosen, we have 4 odd numbers left, so we have 4 choices for the hundreds place. Then 3 choices remain for the tens place, and 2 for the ones place. Multiplying these choices we have

$$\underline{3} \cdot \underline{4} \cdot \underline{3} \cdot \underline{2} = 72$$

4-digit integers bigger than 5000 with distinct odd digits.

In examples (b)-(e) above, we counted the number of **ordered arrangements** - order matters when you're dialing a phone number, or writing down a license plate, or sending players up to bat, or expressing a 4-digit number. In the case of license plates and phone numbers, the same value can be chosen twice. With the lineup no repeat choices are allowed, the lineup must consist of distinct batters. No repeats for those special 4-digit integers either.

To summarize, in the examples so far we have effectively used what we might call the **enumerate subtasks** strategy of counting how many different objects are possible as follows:

- break the task of creating the object into a sequence of subtasks
- count how many choices we have for completing each subtask, and
- multiply all these choice counts.

This process works as long as we would create each object we're trying to count exactly once if we followed every possible combination of step choices.

Example 4.2 (With or Without Replacement)**.** We have reason to consider two variations on the theme of "pick k elements from the set A." We can either pick **with replacement**, meaning each pick is made from the entire set (allowing the same element to be picked multiple times), or we pick **without replacement**, meaning once an element has been picked, it can't be picked again.

How many ways can we pick 3 names from the set $M = \{$Evelyn, Eddie, Gordon, Oriana$\}$:

- with replacement? $\underline{4} \cdot \underline{4} \cdot \underline{4} = 4^3 = 64$.
- without replacement? $\underline{4} \cdot \underline{3} \cdot \underline{2} = 24$.

4.2 Permutations

Definition 4.1. A **permutation** is an ordered arrangement of distinct objects. The number of ways of ordering n distinct objects taken r at a time will be denoted P_r^n.

The symbol P_r^n denotes the number of ways to create an ordered list of length r without repeats from a set of n distinct elements. So, in the baseball lineup example we found $P_9^{13} = 259,459,200$.

In general,

$$P_r^n = n \cdot (n-1) \cdot \dots \cdot (n-r+1),$$

though we can more effectively express P_r^n via factorials.

Recall, $n!$ (read "**n factorial**"), is shorthand for

$$n! = n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1.$$

For instance, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Two notes:

1. We set $0! = 1$
2. Notice that $n! = n \cdot (n-1)!$, so $\dfrac{n!}{(n-1)!} = n$.

So we have the following formula for P_r^n:

$$P_r^n = \frac{n!}{(n-r)!}. \tag{4.1}$$

Referring to the random phones example 3.4, without listing all the possible ordered arrangements we know there will be $4! = 24$ ways to return the phones to the 4 students at random (4 phones to choose from when returning one to the first student, 3 phones for the second student, 2 for the third, and 1 for the fourth).

4.3 Combinations

We let C_r^n equal the number of ways to choose an unordered arrangement of r distinct elements from a set of n distinct objects. We also let $\begin{pmatrix} n \\ k \end{pmatrix}$ denote this number (read as "n choose r").

For instance, $\begin{pmatrix} 4 \\ 2 \end{pmatrix}$ gives how many distinct subsets of size 2 can be formed from a set of size 4, and we see that $\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 6$ because one can form 6 distinct subsets of size 2 from the set $\{A, B, C, D\}$: $\{A, B\}$, $\{A, C\}$,$\{A, D\}$, $\{B, C\}$,$\{B, D\}$, $\{C, D\}$.

We have the following formula for $\begin{pmatrix} n \\ r \end{pmatrix}$:

$$C_r^n = \begin{pmatrix} n \\ r \end{pmatrix} = \frac{n!}{(n-r)! \cdot r!}. \tag{4.2}$$

Proof. We may build an ordered list of r distinct elements from a set of n distinct elements by completing these two tasks:

1. Choose a combination of size r from the set of size n. The number of ways to do this is exactly what we're calling $\binom{n}{r}$.
2. We may choose a particular ordering of the "unordered" r chosen in step 1 in $r!$ ways.

The result is an ordered arrangement of r chosen from a set of n, and the number of such ordered arrangements will be

$$\begin{pmatrix} n \\ r \end{pmatrix} \cdot r!$$

by the multiplication principle.

But we've also denoted the number of ordered arrangments of r from n, as P_r^n, so

$$P_r^n = \begin{pmatrix} n \\ r \end{pmatrix} \cdot r!,$$

and since $P_r^n = n!/(n-r)!$, it follows that

$$C_r^n = \begin{pmatrix} n \\ r \end{pmatrix} = \frac{n!}{(n-r)! \cdot r!}.$$

□

Theorem 4.1. *Facts about* $\begin{pmatrix} n \\ r \end{pmatrix}$ *for* $n \geq 1$ *and* $0 \leq r \leq n$*:*

1. $\begin{pmatrix} n \\ 0 \end{pmatrix} = 1$; $\begin{pmatrix} n \\ 1 \end{pmatrix} = n$; $\begin{pmatrix} n \\ n \end{pmatrix} = 1$.
2. $\begin{pmatrix} n \\ r \end{pmatrix} = \begin{pmatrix} n \\ n-r \end{pmatrix}$.
3. ***Pascal's Formula:*** $\begin{pmatrix} n \\ r \end{pmatrix} = \begin{pmatrix} n-1 \\ r-1 \end{pmatrix} + \begin{pmatrix} n-1 \\ r \end{pmatrix}$.

4. ***Binomial Theorem:*** *For real numbers x, y and $n \in \mathbb{N}$,*

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r.$$

5. $\displaystyle\sum_{r=0}^{n} \binom{n}{r} = 2^n.$

Counting arguments justifying these properties are fun, we'll do these in class, and Pascal's triangle encodes most of them. Because of the binomial theorem, $\binom{n}{k}$ are also called **binomial coefficients**.

With these binomial coefficients in our toolbox, let's return to counting.

Example 4.3.

How many subcommittees of size 3 can be formed from a group of 7 people?

We treat a subcommittee as an unordered subset of the group, so the number of possible subcommittees of size 3 will be $\binom{7}{3} = 35$. I would arrive at the number with paper and pencil by first taking advantage of a lot of cancellations:

$$\begin{aligned}
\binom{7}{3} &= \frac{7!}{4! \cdot 3!} \\
&= \frac{7 \cdot 6 \cdot 5 \cdot 4!}{4! \cdot 3!} \\
&= \frac{7 \cdot 6 \cdot 5}{3!} && \text{canceling the } 4! \text{ terms} \\
&= \frac{7 \cdot 6 \cdot 5}{6} && \text{since } 3! = 6 \\
&= 7 \cdot 5 \\
&= 35.
\end{aligned}$$

Example 4.4.

An ultimate frisbee team is travelling in two vans to a tournament. The purple van seats 8, and the white van seats 12. How many different ways can the 20-player team be split into two groups, the purple group of size 8 and the white group of size 12?

There are $\binom{20}{8}$ ways to choose the purple group, and once they're chosen, the white group has also been formed, so there are $\binom{20}{8} = 125,970$ ways to split the teams into two groups. (Of coures, we could have also answered the question by finding the number of ways to choose the white group, which is $\binom{20}{12}$. This produces the same answer, a fact stated in generality in Part 2 of Theorem 4.1.

Example 4.5.

A market stand has 20 ears of corn left. We plan to purchase 5 ears. How many different combinations of 5 ears can we purchase? If 3 of the 20 ears are actually "bad", how many of these possible purchase combinations would have at least one bad ear?

The first question is answered by finding

$$\binom{20}{5} = \frac{20!}{15! \cdot 5!} = 15,504.$$

In R, binomial coefficients are computed with `choose(n,r)`.

```
choose(20,5)
```

```
## [1] 15504
```

The second question is more interesting.

Let A denote the set of all combinations of size 5 that have at least one bad ear. We want to know the size of A, $|A|$. It's actually easier to find $|\overline{A}|$, the size of the complement of A; that is, it's easier to count how many combinations of size 5 have zero bad ears.

The entire stand has 20 ears, 3 of which are bad, meaning 17 are good. So, the number of combinations with 5 good ears (and hence 0 bad ears) is

$$|\overline{A}| = \binom{17}{5} = 6188.$$

So, of the 15504 different combinations of 5 we could purchase, 6188 of them have zero bad ears, and 9316 have at least one bad ear.

So, if 3 out of 20 are bad, and you pick 5 at random to buy, chances are good you'll end up with at least one bad one: about a 60% chance (9316/15504).

Example 4.6. You inherit a working lottery ping-pong ball machine from a magnificent uncle. The machine has 20 ping-pong balls, numbered 1 through 20. You want to set up a weekly lottery for your favorite charity, and you are considering two options:

Scenario 1: Have the machine pick 4 balls at random, and record the *ordered* arrangement. In this scenario, hopeful lottery participants fill out the lottery card with an ordered list of 4 distinct numbers (from the set 1 to 20).

Scenario 2: Have the machine pick 4 balls at once (it can do this!), record the *unordered* arrangement. Then put all the balls back, and pick one ball as the "wildcat" number. In this scenario, hopeful lottery participants fill out the lottery card with an unordered list of 4 distinct numbers, followed by a choice (from 1 to 20) for the wildcat.

Which lottery game would be more difficult to win?

We count how many distinct tickets are possible in each scenario, using our enumerate subtasks strategy.

In Scenario 1 we see there will be

$$\underline{20} \cdot \underline{19} \cdot \underline{18} \cdot \underline{17} = 116,280$$

distinct tickets.

For Scenario 2 we have $\binom{20}{4}$ ways to choose 4 numbers from the 20, and then 20 choices for the wildcat number, giving

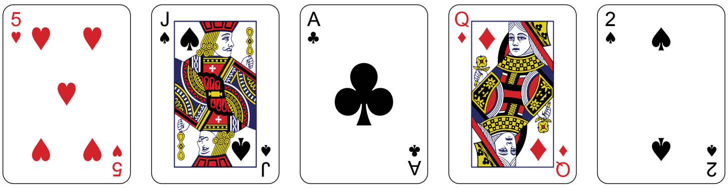
$$\binom{20}{4} \cdot 20 = \frac{20 \cdot 19 \cdot 18 \cdot 17}{4!} \cdot 20 = 96,900.$$

It looks like a lottery following scenario 1 would be a bit more difficult to win than the scenario 2 lottery.

Example 4.7 (Poker)**.** Poker is a family of card games where players bet on who has the best hand according to the rules of the particular game. Most poker games use a standard deck having 52 cards. Each card has two features:

- a **rank**, and there are 13 ranks: 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king), and A (ace).
- a **suit**, and there are 4 suits: “spades”, “hearts”, “diamonds”, and “clubs”.

Here are 5 random cards from a standard deck:



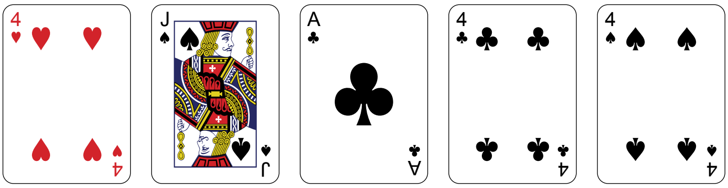
How many different 5 card poker hands are there?

52 distinct cards, choose 5, order doesn’t matter here, so the answer is

$$\binom{52}{5} = \frac{52!}{47! \cdot 5!} = 2,595,960.$$

Now let’s consider a few types of poker hands.

A **Three of a Kind** is a hand of five cards that has 3 of one rank, and the other 2 cards do not have equal rank, such as this hand:



How many different "Three of a Kind" hands are there?

We count the number of ways to build such a hand (we need to fill in each card with a rank and a suit).

1. Pick the rank that appears 3 times - 13 choices.
2. Pick the 3 suits for this rank- we can choose 3 of the 4 possible suits in $\binom{4}{3}$ ways.
3. Pick the remaining (distinct) 2 ranks - $\binom{12}{2}$ ways (since the hand is unordered we do not impose an order on the choice of the remaining 2 ranks).
4. Pick the remaining 2 suits- each of the remaining 2 cards can have any of the 4 suits, we have $4 \cdot 4$ ways to pick these suits.

The total number of Three of a Kind hands is thus

$$13 \cdot \binom{4}{3} \cdot \binom{12}{2} \cdot 4 \cdot 4 = 54912.$$

And we may reasonably define the probability of obtaining a 3-of-a-kind in a deal of 5 cards to be

$$\frac{54912}{\binom{52}{5}} \approx 0.021.$$

How many different full house hands are possible?

A **full house** is a five card hand which has 3 cards of one rank, and 2 cards of another rank.

1. Choose the rank that appears three times. We have $\binom{13}{1} = 13$ ways to do this.
2. Choose three of the four suits for these three cards: $\binom{4}{3}$.
3. Choose the rank that appears twice in the hand. We have 12 ranks left to choose from, so we have $\binom{12}{1} = 12$ ways to do that.
4. Choose two of the four suits for the pair: $\binom{4}{2}$.

Completing these steps builds a full house, so we have

$$\binom{13}{1} \cdot \binom{4}{3} \cdot \binom{12}{1} \cdot \binom{4}{2} = 3744$$

different full houses.

How many different "Two Pair" hands are possible?

A **Two Pair** is a five card hand which has 3 cards of one rank, and 2 cards of another rank.

We present two arguments, but *only one of them is correct.* Which is it?

First argument:

1. Pick distinct ranks for the two pairs. We have $\binom{13}{2}$ ways to do this.
2. Choose two of the four suits for one pair: $\binom{4}{2}$.
3. Choose two of the four suits for the other pair: $\binom{4}{2}$.
4. The fifth card can be any of the remaining 44 cards: $\binom{44}{1}$.

Second Argument:

1. Pick the rank for the first pair. We have $\binom{13}{1}$ ways to do this.
2. Choose two of the four suits for the first pair: $\binom{4}{2}$.
3. Pick the rank for the second pair. We have $\binom{12}{1}$ ways to do this.
4. Choose two of the four suits for the second pair: $\binom{4}{2}$.
5. The fifth card can be any of the remaining 44 cards: $\binom{44}{1}$.

The first argument gives us this number:

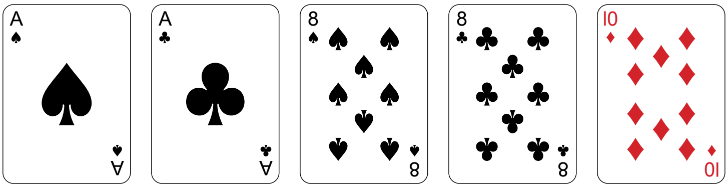
$$\binom{13}{2} \cdot \binom{4}{2} \cdot \binom{4}{2} \cdot \binom{44}{1} = 123552.$$

The second argument gives us a number that is twice as big as the first:

$$\binom{13}{1} \cdot \binom{4}{2} \cdot \binom{12}{1} \cdot \binom{4}{2} \cdot \binom{44}{1} = 247104.$$

Which number is correct?

It turns out the second argument double counts. By picking the two ranks separately in the second argument, we impose an order on the ranks that is not a part of a poker hand. For instance, the second argument allows us to build the following famous Two Pair in two different ways:



The first way to build the hand following the second argument:

1. Choose the rank ace (A).
2. Choose the suits clubs and spades for the aces.
3. Choose the rank 8.
4. Choose the suits clubs and spades for the eights.
5. Choose the last card (10 of diamonds).

The second way to build the hand following the second argument:

1. Choose the rank 8.
2. Choose the suits clubs and spades for the eights.
3. Choose the rank A.
4. Choose the suits clubs and spades for the aces.
5. Choose the last card (10 of diamonds).

In the end the two hands are the same because the order of the cards in your hand doesn't matter, so the second argument double counts. The first argument gives the correct answer.

The number of Two Pair hands is 123552.

4.4 Multinomial Coefficients

Recall, $\binom{n}{r}$ counts the number of ways to choose r from a set of n. Effectively, then, $\binom{n}{r}$ counts the number of ways to assign n elements to two groups of a specific size: r in the "chosen" group, and $n-r$ in the "not chosen" group.

We can generalize this scene to k groups for $k \geq 2$.

Definition 4.2. Let

$$\binom{n}{n_1, n_2, \cdots, n_k}$$

denote the number of ways of partitioning n distinct objects into k groups whose sizes are $n_1, n_2, \ldots, n_k$, where $n_1 + n_2 + \cdots + n_k = n$. These expressions are called **multinomial coefficients**.

Fact:

$$\binom{n}{n_1, n_2, \cdots, n_k} = \frac{n!}{n_1! \cdot n_2! \cdot \cdots \cdot n_k!} \tag{4.3}$$

Note that the binomial coefficent $\begin{pmatrix} n \\ r \end{pmatrix}$ can be written as a multinomial coefficent: $\begin{pmatrix} n \\ r, n-r \end{pmatrix}$.

Example 4.8. A group of 15 volunteers will be split into 5 groups of 3, with each group working on a distinct project (weeding, spreading bark, setting up irrigation, constructing raised beds, and putting a roof on a shed).

> How many distinct ways can the volunteers be split into five groups of three to work on these distinct projects?

The multinomial coefficient

$$\binom{15}{3,3,3,3,3} = \frac{15!}{(3!)^5} = 168,168,000$$

counts the number of ways to do this if we treat the groups as distinct, which we should because the work done by each group is different: Mike in the weeding group is a much different scenario than Mike in the roof-raising group!

Example 4.9.

> How many "words" are spelled from the letters in 'BANANAS.'

If all seven letters in the word were distinct, we could form 7! words, but we have repeat letters: 3 As, 2 Ns, 1 B, and 1 S.

We can arrive at the number of "words" from the 'enumerate subtasks' approach:

We have to build a 7 letter word from the letters in BANANAS.

1. Pick a location for the unique B: 7 choices
2. Pick a location for the unique S: 6 choices left
3. Choose 2 locations from the remaining 5 for the 2 identical Ns: $\binom{5}{2}$
4. Choose 3 locations from the remaining 3 for the 3 identical As: $\binom{3}{3} = 1$

All told, we have $7 \cdot 6 \cdot \binom{5}{2} \cdot 33 = 420$ "words".

Alternatively, the number of words equals the multinomial coefficient

$$\binom{7}{3, 2, 1, 1} = 420$$

because it counts the number of ways to assign 7 distinct elements (the 7 spots for letters in the word) into four groups of size 3, 2, 1, and 1 (3 locations for the As group, 2 for the Ns, 1 for the B and 1 for the S).

4.5 Balls and Bins

Suppose I have 30 Watermelon *Jolly Ranchers* to give away to 8 friends.

How many ways can I pass them out so that each person gets at least 1?

To be precise, if we let n_i denote the number of jolly ranchers that person i receives (i runs from 1 to 8), I want to count how many different vectors $(n_1, n_2, \ldots, n_8)$ are possible with the conditions that each $n_i \geq 1$ and the sum of the n_i equals 30.

To answer this question, we first consider balls and bins.

Theorem 4.2. *The number of ways to distribute n identical balls into r distinct bins is*

$$\binom{n+r-1}{r-1}.$$

Proof. Here's an outline of the proof. Say we have 8 balls and 3 bins. Here's a schematic of one way to distribute them:

$$\underline{|\ \text{oo}\ |}\quad \underline{|\ \text{oooo}\ |}\quad \underline{|\ \text{oo}\ |}$$

Two balls into the first bin, 4 into the second, and 2 into the third. We can also represent the distribution by pushing the bins together so that they share walls:

$$\underline{|\ \text{oo}\ |\ \text{oooo}\ |\ \text{oo}\ |}$$

In fact, we don't really need those two outermost walls to communicate the distribution, or the bin bottoms for that matter:

oo|oooo|oo

So each distribution of the 8 balls into 3 bins corresponds to an ordered arrangement of 8 balls and 2 "inner walls", and we have $\binom{10}{2}$ ways to choose the 2 spots for the inner walls from the 10 spots needed to create the arrangement.

More generally, with n balls and r bins we will have $r-1$ inner walls, which leads us to the formula

$$\binom{n+r-1}{r-1}.$$

□

Let's return to the Watermelon *Jolly Ranchers.*

In this scenario, the friends are the bins, and the candies are the balls, and the number of ways to distribute the candies to the 8 friends is $\binom{30+8-1}{8-1} = \binom{37}{7}$, which we can calculate in R: `choose(37,7)` = 10295472. I've got options.

However, this count doesn't answer the original question because it counts *all* the ways to distribute the candies, including, say, all 30 going to one person. The original question here was to count the number of ways we can distribute the candies so that each friend gets at least 1.

So we do the following: first give each friend one *Jolly Rancher*, then count the number of ways to distribute the remaining 22 using balls (22) and bins (8). It follows that the number of ways to distribute the candies so that each friend gets at least 1 is

$$\binom{22+8-1}{8-1} = \binom{29}{7},$$

which is `choose(29,7)` = 1560780. Still, I've got options.

4.6 Calculating More Probabilities

With our counting tools in hand, we can set about calculating the probability that an event A happens without having to first every possible outcome in the sample space.

Here's the general scene in this section. We have some random experiment with finite sample space S, defined in such a way that the probability of each outcome in S is the same. In this case, the probability of any event A will simply be the size of the set A divided by the size of the set S:

$$P(A) = \frac{|A|}{|S|}.$$

Example 4.10.

> Suppose we want to pick a 4 person subcommitee from a committee of 8 having 5 Republicans, and 3 Democrats. If we pick the subcommittee at random, what is the probability that all three Democrats are on it?

The sample space S here is all possible subcommittees of size 4. Treating each of the 8 members as distinct people we have

$$|S| = \binom{8}{4}.$$

The event A that all Democrats are in the subcommitte can be enumerated as follows:

1. Choose all 3 Democrats for the subcommittee: $\binom{3}{3} = 1$ way to do that!
2. Choose 1 Republican from 5 to fill out the subcommittee: $\binom{5}{1} = 5$.

So $|A| = 5$, and using pencil and paper for this one:

$$\begin{aligned} P(|A|) &= \frac{5}{(8 \cdot 7 \cdot 6 \cdot 5)/4!} \\ &= \frac{4!}{8 \cdot 7 \cdot 6} \\ &= \frac{4 \cdot 3 \cdot 2}{8 \cdot 7 \cdot 6} \\ &= \frac{3}{7 \cdot 6} \\ &= \frac{1}{7 \cdot 2} \\ &= \frac{1}{14}. \end{aligned}$$

Example 4.11.

A wine expert samples 8 wines blindly, two of which are genuinely "high quality," and the others are "budget wines." The expert choses their 3 favorites. If they are really just picking at random, what is the probability that they pick both high quality wines?

The sample space S consists of all possible ways to choose 3 from 8, and we consider each of these combinations of 3 equally likely.

We have $|S| = \binom{8}{3} = 56$.

The event of interest here, A, is choosing a combination that contains both "high quality" wines, and $|A| = \binom{2}{2} \cdot \binom{6}{1} = 6$ (choose both of the high quality wines and choose one of the six budget wines). Thus,

$$P(A) = \frac{6}{56},$$

about a 10% chance of picking both high quality wines if, in fact, the expert is simply picking at random.

This probability is useful to know. If the expert does select both the high quality wines, does this suggest to you they know their business?

Example 4.12 (Good Potatoes Bad Potatoes)**.** A truck has 3000 potatoes (of which 75 are bad). We inspect 50 potatoes at random. We reject the shipment if more than two of the potates in the sample are bad.

What is the probability that we reject the shipment?

We have 3000 potatoes, 2925 good potatoes, and 75 bad potatoes.

The sample space S here is all combinations of size 50, so

$$|S| = \binom{3000}{50}.$$

The event of interest A is chosing a sample with more than 2 bad potatoes.

The number of samples of size 50 with 50 good and 0 bad potatoes is:

$$\binom{2925}{50} \cdot \binom{75}{0}.$$

The number of samples of size 50 with 49 good and 1 bad potato is:

$$\binom{2925}{49} \cdot \binom{75}{1}.$$

The number of samples of size 50 with 48 good and 2 bad potatoes is:

$$\binom{2925}{48} \cdot \binom{75}{2}.$$

The number of samples with *more than 2* bad potatoes is thus

$$|A| = |S| - \left[\binom{2925}{50} \cdot \binom{75}{0} + \binom{2925}{49} \cdot \binom{75}{1} + \binom{2925}{48} \cdot \binom{75}{2}\right],$$

and the probability of rejecting the sample is

$$P(A) = \frac{|A|}{|S|} \approx 0.1279.$$

Example 4.13 (Poker II)**.**

Detemrine the probabilty of drawing each of these standard 5 card poker hands: a Straight, a Flush, and a Straight Flush.

These poker hands just keep on coming!

A **Straight Flush** is a five cards poker hand in which all five suits are the same, and all five ranks form a run (either 2-6, 3-7, 4-8, 5-9, 6-10, 7-J, 8-Q, 9-K, 10-A).

A **Straight** consists of five cards such that all five ranks form a run and the hand is not a Straight Flush.

A **Flush** consists of five cards such that all five suits are the same and the hand is not a Straight Flush.

Here the sample space is $S = \{\text{all possible 5 card hands}\}$, and we know $|S| = \binom{52}{5}$.

How many different Straight Flushes are possible. Let's build one in steps, counting choices:

1. Pick the low card for the run (which then determines all 5 ranks): 9 choices (2, 3, 4, 5, 6, 7, 8, 9, or 10)
2. Pick the one suit for all five cards: 4 choices.

So there are $9 \cdot 4 = 36$ Straight Flushes, and the probability of being dealt a Straight Flush is

$$P(\text{Straight Flush}) = \frac{36}{\binom{52}{5}} \approx 0.00001385.$$

We now turn to the Straight.

As we saw in the case of the Straight Flush, we have 9 choices for the rank of the low card in a straight. Once that has been chosen, all five ranks are determined.

We have 4 choices for the suit of each of the 5 cards (allowing for the possibility that all five cards have the same suit), so $9 \cdot 4^5$ counts the number of Straights + Straight Flushes. The number of Straights is thus $9 \cdot 4^5 - 36 = 9180$. It follows that

$$P(\text{Straight}) = \frac{9180}{\binom{52}{5}} \approx 0.00353.$$

Finally, we count Flushes:

1. Choose the flush suit: $\binom{4}{1}$ ways.
2. Choose the 5 ranks: $\binom{13}{5}$ ways.

Then $\displaystyle\binom{4}{1} \cdot \binom{13}{5} = 5148$ counts the number of Flushes, *almost.* When we chose the 5 ranks, we allowed for the possibility that the ranks also formed a straight, so we must subtract out the Straight Flushes. The total number of Flushes: $5148 - 36 = 5112$.

So

$$P(\text{Flush}) = \frac{5112}{\binom{52}{5}} = .00197.$$

So we have about a 1 in 500 chance of being dealt a Flush, and we note that a Flush is less common than a Straight (which is why a Flush beats a Straight in regular 5-card draw).

Example 4.14. Classic Oregon license plates consist of 3 digits (0-9) followed by three letters (A-Z).

> Find the probability that a randomly selected classic Oregon license plate has two 8s.

The sample space S consists of all possible license plates, and $|S| = 10^3 \cdot 26^3$, since order matters for license plates.

We enumerate the event A of drawing a plate with two 8s as follows:

1. Choose the 2 spots for the 8s: $\binom{3}{2} = 3$ choices
2. Pick the other number: 9 choices
3. Pick the three letters: 26^3 choices.

So

$$P(A) = \frac{3 \cdot 9 \cdot 26^3}{10^3 \cdot 26^3} = \frac{27}{1000} = 0.027.$$

Example 4.15.

Determine the probability that the first time we roll an 8 or higher with a 10-sided die is on the 5th roll.

The sample space S consists of all sequences of 5 rolls of the die. For instance, $(4, 5, 2, 9, 5)$ is one element of S, and $|S| = 10^5$.

The event of interest is

$$A = \{\text{roll 7 or lower in the first four rolls and 8 or higher on the 5th}\}.$$

Then $|A| = 7^4 \cdot 3$ (7 choices for the 1st roll, 7 for the second, 7 for the 3rd, 7 for the 4th, and 3 for the 5th).

So

$$P(A) = \frac{7^4 \cdot 3}{10^5}.$$

Example 4.16.

A class has 12 people: 6 juniors, 4 sophomores, and 2 first-years. The class is randomly divided into 3 subgroups of size 5, 4, and 3. What is the probability that the 2 first-years are in the same subgroup?

The sample space S consists of all possible partitions of the 12 people into the 3 subgroups, and we know

$$\begin{aligned}|S| &= \binom{12}{5,4,3} \\ &= \frac{12!}{5! \cdot 4! \cdot 3!} \\ &= \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{24 \cdot 6} \\ &= 27{,}720.\end{aligned}$$

The event of interest, A, consists of all partitions in which the two first-years are in the same subgroup. To count $|A|$ we consider three cases:

Case 1: The two first-years are in the subgroup of 5. This means, after placing them in there, we have 10 people left to place in groups of size 3, 4, and 3, and we can do this in

$$\binom{10}{3,4,3}$$

ways.

Case 2: The two first-years are in the subgroup of 4, leaving 10 to place in groups of size 5, 2, and 3:

$$\binom{10}{5,2,3}.$$

Case 3: The two first-years are in the subgroup of 3, which can happen in

$$\binom{10}{5,4,1}$$

ways.

So,

$$|A| = \binom{10}{3,4,3} + \binom{10}{5,2,3} + \binom{10}{5,4,1},$$

which evaluates to $|A| = 7980$. So,

$$P(A) = \frac{7980}{27720} \approx 0.288.$$

Chapter 5

Probability Theory

5.1 Conditional Probability and Independence

Suppose we have a probability model associated to a sample space S. If we are told some event B has occurred, how would the probability of other events change? Calculating a new probability for event A, given that B has occurred is called a conditional probability, and will be denoted $P(A|B)$.

For instance, Let X denote the outcome if we roll a fair six-sided die. Let A be the event that we roll a 4, and B the event that we roll an even number. Since the die is fair, we expect that $P(A) = 1/6$. Now suppose that the die is rolled and we are told that the event B has occurred. This leaves only three possible outcomes: 2, 4, and 6. The new, conditional probability of A given B would be $P(A|B) = 1/3$.

Definition 5.1. The **conditional probability** of an event A, given that an event B has occurred, denoted $P(A|B)$, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{5.1}$$

provided that $P(B) > 0$.

Example 5.1. For the 2023 Major League Baseball season, 328 hitters had at least 250 plate appearances. In this group, 71% of them hit at least 10 HR for the season, and 31% of them hit at least 20 HR. If you pick a player from this group at random, and you are told they hit over at least 10 HR, what is the probability that they hit at least 20 HR?

Let A be the event that the player hits at least 20 HR, and B the event that a player hit at least 10. Then we have been asked to find $P(A|B)$. Note that A is a subset of B (if a player hit at least 20, then they also hit at least 10), so $A \cap B$

$= A$, and it follows that

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)}{P(B)} \\ &= \frac{.31}{.71} \\ &\approx .437. \end{aligned}$$

Note that from the conditional probability formula,

$$P(A \cap B) = P(B) \cdot P(A \mid B), \text{ provided } P(B) > 0,$$

and that

$$P(A \cap B) = P(A) \cdot P(B \mid A), \text{ provided } P(A) > 0.$$

Definition 5.2. Two events are called **independent** if any of the following statements holds:

$$\begin{aligned} P(A \mid B) &= P(A) \\ P(B \mid A) &= P(B) \\ P(A \cap B) &= P(A) \cdot P(B) \end{aligned}$$

Example 5.2. The chance experiment "roll a 6-sided die" has sample space $S = \{1, 2, 3, 4, 5, 6\}$. Consider the events

$$\begin{aligned} A &= \{1, 3, 5\} \\ B &= \{1\} \\ C &= \{2\} \\ D &= \{1, 2\} \end{aligned}$$

Then $P(A) = 1/2, P(B) = 1/6, P(C) = 1/6$, and $P(D) = 1/3$.

Claim 1: A and D are independent events.

Reason 1: Well, $A \cap D = \{1\}$, so

$$P(A \mid D) = \frac{P(A \cap D)}{P(D)} = \frac{1/6}{1/3} = 1/2 = P(A).$$

Claim 2: B and C are not independent events.

Reason 2: $B \cap C = \emptyset$, so

$$P(B \mid C) = \frac{P(B \cap C)}{P(C)} = \frac{0}{1/6} = 0 \neq P(B).$$

These simple examples help me remember that for events associated to a sample space,

> independent is different than disjoint!

In this example A and D are independent but not disjoint, while B and C are disjoint but not independent.

Example 5.3. Suppose over the last five years, 10% of all people in a town who have hired a plumber to do some work have been unhappy with the job that was done. One of the plumbers in the town is Frances. Frances has done 40% of the plumbing jobs in town over these five years, and 25% of all plumbing jobs that left the customer unhappy have been done by Frances.

> Find the probability that a customer will be unhappy with the results if they hire Francis?

We begin by translating what we know and what we want into symbols and probability notation.

Let $S = \{\text{all plumbing jobs in the town over the last 5 years}\}$.

Let $A = \{\text{jobs handled by Frances}\}$.

Let $B = \{\text{jobs in which the customer is unhappy}\}$.

What do we want? $P(B \mid A)$.

When do we want it? Now! So let's write down what we know.

What do we know?

- 10% of the jobs have left the customer unhappy $\Rightarrow \quad P(B) = 0.1$.
- Frances has done 40% of all jobs $\Rightarrow \quad P(A) = 0.4$.
- 25% of all complaints dealt with Frances $\Rightarrow \quad P(A \mid B) = 0.25$.

Well,

$$\begin{aligned} P(B \mid A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(A \mid B) \cdot P(B)}{P(A)} \\ &= \frac{(0.25)(0.1)}{(0.4)} \\ &= 0.0625. \end{aligned}$$

Ok, there is about a 6 percent chance that a customer will be unhappy with their plumbing job if they hire Francis.

An irresponsible person who wants to be intentionally misleading could rant in all caps "25 PERCENT OF UNHAPPY CUSTOMERS HIRED FRANCIS!!!" Let's be better than that. Knowing the full context here, which includes what proportion of the town's plumbing jobs have gone to Frances, is necessary to establish how effective Frances has been as a plumber these last five years.

Example 5.4.

> Roll 2 6-sided dice. What is the probability that both values are less than 3?

We assume the two dice are independent. (What appears on one die is independent of what appears on the other.)

Let $A = \{\text{first die is less than 3}\}$ and $B = \{\text{second die is less than 3}\}$.

"Less than 3" means "1 or 2" so $P(A) = P(B) = 2/6 = 1/3$. The question asks for $P(A \cap B)$. Since A and B are independent,

$$P(A \cap B) = P(A) \cdot P(B) = (1/3) \cdot (1/3) = 1/9.$$

We remark that we can also use counting techniques to find this probability directly in Section 4.6.

The sample space S of this chance experiment consists of all possible rolls of the two dice. That is, S consists of all ordered pairs of the form (i, j) with $i, j \in \mathbb{N}$ and $1 \leq i, j \leq 6$, and $|S| = 6^2$. The event of interest, E, consists of all rolls (i, j) in S such that $i, j \leq 2$. So, $|E| = 2^2$, and $P(E) = 4/36 = 1/9$.

So, thinking of the chance experiment as a sequence of independent events we calculate the probability of interest as $\frac{2}{6} \cdot \frac{2}{6},$, and thinking of the probability via the "(outcomes of interest)/(all possible outcomes)" approach we think of the probability as $\frac{2 \cdot 2}{6 \cdot 6}$.

Example 5.5. Roll a regular 6-sided die until a 4 comes up. What is the probability that this occurs on the 8th roll?

The values of the rolls are independent events, the probability of not rolling a four on a given roll is 5/6, and the probability of rolling a 4 is 1/6. It follows that the probability of rolling 7 non-4s followed by 1 4 is

$$P(\text{first 4 on roll 8}) = \left(\frac{5}{6}\right)^7 \cdot \left(\frac{1}{6}\right).$$

More generally, the probability that our first 4 comes up on roll n (for any $n \in \mathbb{N}$) will be

$$P(\text{first 4 on roll } n) = \left(\frac{5}{6}\right)^{n-1} \cdot \left(\frac{1}{6}\right).$$

5.2 Two Laws of Probability

Theorem 5.1. *Suppose A and B are two events.*

1. ***Multiplicative Law of Probability:***

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B \mid A) \\ &= P(B) \cdot (A \mid B) \end{aligned}$$

2. ***Additive Law of Probability:***

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof.

1. This proof follows directly from the definition of conditional probability (5.1).
2. For finite sets A and B, we know

   $$|A \cup B| = |A| + |B| - |A \cap B|,$$

   from which the result follows.

More generally, for any sets A and B, the union $A \cup B$ can be decomposed into disjoint sets: $A \cup B = A \cup (\overline{A} \cap B)$. So,

$$P(A \cup B) = P(A) + P(\overline{A} \cap B).$$

Similarly, we can decompose the set B into disjoint sets as follows: $B = (\overline{A} \cap B) \cup (A \cap B)$. So

$$P(B) = P(\overline{A} \cap B) + P(A \cap B).$$

Combining these two probability equations we see

$$P(A \cup B) = P(A) + [P(B) - P(A \cap B)]\,.$$

□

For three events, A_1, A_2, A_3 it follows that

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P((A_1 \cap A_2) \cap A_3) \\ &= P(A_1 \cap A_2) \cdot P(A_3 \mid A_1 \cap A_2) \\ &= P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2). \end{aligned}$$

This formula may be extended to the intersection of any number of sets.

$$P(A_1 \cap \cdots \cap A_k) = P(A_1) \cdot P(A_2 \mid A_1) \cap \cdots \cap P(A_k \mid (A_1 \cap \cdots \cap A_{k-1})) \tag{5.2}$$

Example 5.6. Flip over 4 cards from a regular 52-card deck. What is the probability they are all hearts?

We flip the cards over one at a time, and define the four events A_i = the event that card i is a hearts, for $i = 1, 2, 3, 4$.

We want the probability that all four events occur. That is, we want $P(A_1 \cap A_2 \cap A_3 \cap A_4)$, and we find this using Equation (5.2).

$$\begin{aligned}P(A_1 \cap A_2 \cap A_3 \cap A_4) &= P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot P(A_4 \mid A_1 \cap A_2 \cap A_3) \\ &= \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \\ &\approx 0.0026.\end{aligned}$$

Example 5.7.

Find the probability that a random 4 digit number has distinct odd digits.

Two notes: We have 5 odd digits, and the leading (thousands) digit of a 4-digit number cannot be 0.

Using the multiplicative law of probability we calculate the probability of a sequence of events and multiply them:

- The probability that the leading digit is odd: 5/9.
- The probability that the 2nd digit is odd, given the first was: 4/10.
- The probability that the 3rd digit is odd, given the first two were: 3/10.
- The probability that the 4th digit is odd, given the first three were: 2/10.

By the multiplicative law of probability, the probability that a random 4-digit number has distinct odd digits is

$$\frac{5}{9} \cdot \frac{4}{10} \cdot \frac{3}{10} \cdot \frac{2}{10} \approx 0.0133.$$

We have three corollaries to Theorem 5.1.

Corollary 5.1. *For any event A,*

$$P(\overline{A}) = 1 - P(A).$$

Proof. By the additive law,

$$P(A \cup \overline{A}) = P(A) + P(\overline{A}) - P(A \cap \overline{A}).$$

Since $A \cup \overline{A} = S$ and $A \cap \overline{A} = \emptyset$, $P(A \cup \overline{A}) = 1$ and $P(A \cap \overline{A}) = 0$, so

$$1 = P(A) + P(\overline{A}),$$

and the result follows. □

Corollary 5.2. *If A and B are disjoint events, then $P(A \cup B) = P(A) + P(B)$.*

Proof. Since $A \cap B = \emptyset$, $P(A \cap B) = 0$ and the result follows from the Additive Law of Probability. □

Corollary 5.3. *For three events A, B, C.*

The probability of their union is

$$\begin{aligned} P(A \cup B \cup C) = P(A) &+ P(B) + P(C) \\ &- [P(A \cap B) + P(A \cap C) + P(B \cap C)] \\ &+ P(A \cap B \cap C). \end{aligned}$$

The probability of their intersection is

$$P(A \cap B \cap C) = P(A) \cdot P(B \mid A) \cdot P(C \mid A \cap B).$$

Proof. First we tackle the union case by appealing to the additive law of probability twice, along a the distributive law of sets.

$$\begin{aligned} P(A \cup B \cup C) &= P(A \cup (B \cup C)) \\ &= P(A) + P(B \cup C) - P(A \cap (B \cup C)) \\ &= P(A) + [P(B) + P(C) - P(B \cap C)] - P((A \cap B) \cup (A \cap C)) \\ &= P(A) + P(B) + P(C) - P(B \cap C) \\ &\quad - [P(A \cap B) + P(A \cap C) - P((A \cap B) \cap (A \cap C))] \\ &= P(A) + P(B) + P(C) \\ &\quad - [P(A \cap B) + P(A \cap C) + P(B \cap C)] \\ &\quad + P((A \cap B) \cap (A \cap C)) \end{aligned}$$

from which the result follows since $A \cap B \cap A \cap C = A \cap B \cap C$.

For the intersection case, by definition of conditional probability the right hand side of the equation is

$$\text{RHS } = P(A) \cdot \frac{P(A \cap B)}{P(A)} \cdot \frac{P(C \cap A \cap B)}{P(A \cap B)},$$

and the result follows by cancellation. □

5.3 Event-Composition Method

We've been using a method called the event-composition method for calculating probabilities associated to an experiment.

Event-Composition Method: Describe the sample space and relevant events; write down what information we're given regarding probabilities etc. via the symbols representing these events; Express what we want to know via these symbols, and use our probability laws to use what we know to determine what we want.

Example 5.8. Two regular 6-sided dice are rolled and we record the sum.

What is the probability that we roll a 4 before we roll a 7?

We define two key events:

- A: we roll a sum of 4; and
- B: we do not roll a 4 or 7.

From our handy 6×6 grid describing all possible sums when rolling 2 dice (Example 3.3), we know

$$P(A) = \frac{3}{36}, P(B) = \frac{27}{36}.$$

In this game we roll the dice until we roll a 4 or a 7, and we **win** if we roll a 4 before rolling a 7.

There is no limit to how many rolls we might need in order to win. We might win on the 1st roll, or the 2nd roll, or the 3rd roll, or the 4th roll, ... etc. In fact, for each $n \in \mathbb{N}$, we might win on roll n.

Put another way, we can partition the event of winning into a countably infinite collection of mutually disjoint events: winning on roll 1, winning on roll 2, winning on roll 3, etc. Then,

$$P(\text{winning}) = \sum_{n=1}^{\infty} P(\text{winning on roll } n).$$

The probability of winning on the 1st roll is $P(A)$.

The probability of winning on the 2nd roll is $P(B) \cdot P(A)$ (no 4 or 7 on 1st roll and yes 4 on the 2nd roll).

The probability of winning on the 3rd roll is $P(B) \cdot P(B) \cdot P(A)$, and, more generally, the probability of winning on roll n is

$$P(B)^{n-1} \cdot P(A).$$

The sum of these probabilities is an honest-to-goodness-living-in-the-wild geometric series:

$$\begin{aligned} P(4 \text{ before } 7) &= \sum_{n=1}^{\infty} P(B)^{n-1} P(A) \\ &= P(A) \sum_{n=0}^{\infty} P(B)^n \\ &= (3/36) \sum_{n=0}^{\infty} (27/36)^n \end{aligned}$$

As a reminder, the sum of a geometric series is given by

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r} \text{ if } |r| < 1 \tag{5.3}$$

Using this formula with $r = 27/36$, we see the probability that we roll a 4 before a 7 is

$$P(4 \text{ before } 7) = \frac{1}{3}.$$

5.4 Bayes' Rule

Definition 5.3. A collection $\{B_1, B_2, \ldots, B_n\}$ of nonempty subsets of S is called a **partition** of S provided that

1. $S = B_1 \cup B_2 \cup \cdots \cup B_n$, and
2. the collection is pairwise disjoint.

Theorem 5.2 (Law of Total Probability)**.** *Assume $\{B_1, B_2, \ldots B_k\}$ is a partition of the sample space S, and $P(B_i) > 0$ for each $i = 1, 2, \ldots k$. Then for any event A,*

$$P(A) = \sum_{i=1}^{k} P(A \mid B_i) \cdot P(B_i).$$

Proof. Notice that as a set,

$$\begin{aligned} A &= A \cap S \\ &= A \cap (B_1 \cup B_2 \cup \cdots \cup B_k) \\ &= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_k). \end{aligned}$$

Furthermore, since the B_i are pairwise disjoint, the $(A \cap B_i)$ are pairwise dijoint as well, and by the additive law of probability

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i).$$

The result then follows since each $P(A \cap B_i) = P(A \mid B_i) \cdot P(B_i)$. □

Example 5.9. An ad agency notices

- 1 in 50 potential buyers of a particular product sees an advertisement for it on television
- 1 in 5 potential buyers sees the ad on YouTube.
- 1 in 100 sees the ad on both TV and YouTube.
- 1 in 3 potential buyers actually purchase the product after seeing an ad, and
- 1 in 10 potential buyers buy it without seeing an ad.

> What is the probability that a radnomly selected potential buyer will purchase the product?

We define relevant events for the sample space $S = \{$ all potential buyers $\}$. In particular, we let

- $A =$ the set of potential buyers who purchase the product,

- $B =$ the set of potential buyers who see the ad on TV,
- $C =$ the set of potential buyers who see the ad on YouTube.

Next we translate what we know and what we want into symbols.

What we want: $P(A)$.

What we know:

- $P(B) = 1/50$,
- $P(C) = 1/5$,
- $P(B \cap C) = 1/100$,
- $P(A \mid B \cup C) = 1/3$, and
- $P(A \mid \overline{B \cup C}) = 1/10$.

We also know by the additive law of probability that

$$P(B \cup C) = \frac{1}{50} + \frac{1}{5} - \frac{1}{100} = \frac{21}{100},$$

so by Corollary 5.1,

$$P(\overline{B \cup C}) = 1 - \frac{21}{100} = \frac{79}{100}.$$

Finally, notice that since $B \cup C$ and $\overline{B \cup C}$ partition the sample space,

$$A = [A \cap (B \cup C)] \cup \left[A \cap (\overline{B \cup C})\right],$$

where these two sets are disjoint.

By the Law of Total Probability,

$$\begin{aligned} P(A) &= P(A \cap (B \cup C)) + P(A \cap (\overline{B \cup C})) \\ &= P(B \cup C) \cdot P(A \mid B \cup C) + P(\overline{B \cup C}) \cdot P(A \mid \overline{B \cup C}) \\ &= \frac{21}{100} \cdot 13 + \frac{79}{100} \cdot 110 \\ &= \frac{7}{100} \cdot \frac{7.9}{100} \\ &= 0.149. \end{aligned}$$

Example 5.10. Suppose you have taken a test for a deadly disease. The doctor tells you that the test is quite accurate in that if you have the disease then the test will correctly tell you that you have the disease 100% of the time. However, if you don't have the disease, the test will very occasionally (1 time in 10) mistakenly tell you that you have it.

The test comes back positive (it says you have the disease)! Are you worried!? In particular, can you estimate the probability that you actually have the disease given that the test came back positive?

What information were we given? What information do we want?

We were given:

- The probability of a positive test given I have the disease is 1.

- The probability of a positive test given I do not have the disease is 0.1

We want:

- The probability I have the disease given that I have a positive test.

Let A denote the event that I have a positive test, B the event that I have the disease.

Then I want $P(B|A)$ and I know $P(A|B) = 1$ and $P(A|\overline{B}) = 0.1$.

It turns out I need more information than I've been given to answer this question, as the following scenarios demonstrate.

Scenario 1: Suppose the population consists of 100 people, and 50 people, in fact, have the disease (blue - healthy, red - sick in the figure below). Then, if we tested each individual we would find 55 positive tests (the circled dots below) (all 50 of the sick people would test positive, and 10% of the 50 healthy people, so 5 healthy people would also test positive.)

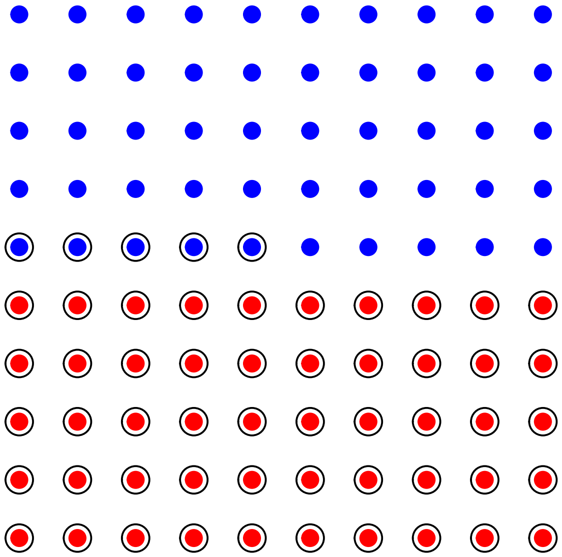


Figure 5.1: 50% of population has the disease

In this scenario, given that I tested positive, there is a 50/55 chance that I have the disease.

Scenario 2: Suppose the population consists of 100 people, and 1 person, in fact, has the disease (blue - healthy, red - sick in the figure below). Then, if we tested each individual we would find about 11 positive tests (the circled dots below) (the one sick people would test positive, and 10% of the 99 healthy people, so about 10 healthy people would also test positive.)

In this scenario, given that I tested positive, there is a 1/11 chance that I have the disease.

So, it seems to answer the question in this example, it is important to know what percentage of the population have the disease. Baye's Theorem below tells us that this is all the additional information we need.

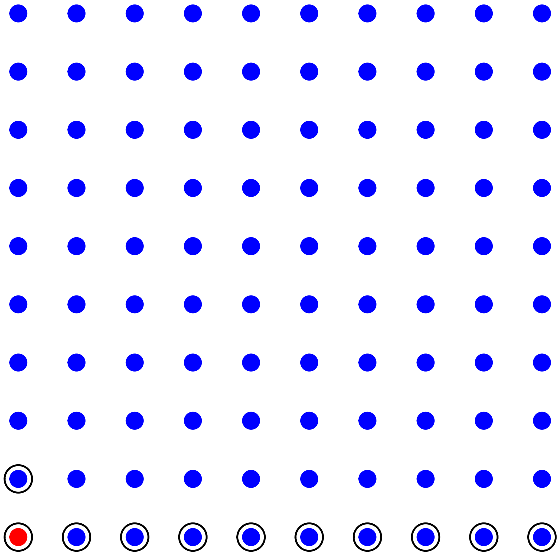


Figure 5.2: 1% of population has the disease

Theorem 5.3 (Bayes' Rule)**.** *Assume $\{B_1, B_2, \ldots B_k\}$ is a partition of the sample space S, and $P(B_i) > 0$ for each $i = 1, 2, \ldots k$. Then for any particular B_j,*

$$P(B_j \mid A) = \frac{P(A \mid B_j) \cdot P(B_j)}{\sum_{i=1}^{k} P(A \mid B_i) \cdot P(B_i)}.$$

Example 5.11. A grocery store has an apple bin. 70% of the apples are Liberty, and 30% are Braeburn. From past experience, we know that 8% of Liberty apples are bad, and 15% of Braeburn apples are bad. Suppose you pick an apple at random and find it is bad. What is the probability that the apple is a Braeburn?

We define our relevant sets.

- S = all apples in the bin
- B_1 = all Liberty apples in S
- B_2 = all Braeburn apples in S (and the B_i partition S!)
- A = bad apples in S.

We know:

- $P(B_1) = .7$, and $P(B_2) = .3$
- $P(A \mid B_1) = .08$, and $P(A \mid B_2) = .15$.

We want:

- $P(B_2 \mid A)$.

This task calls for Bayes' Rule.

$$P(B_2 \mid A) = \frac{P(A \mid B_2) \cdot P(B_1)}{P(A \mid B_1) \cdot P(B_2) + P(A \mid B_2) \cdot P(B_2)}.$$

We know each probability in the right-hand side of the equation:

$$P(B_2 \mid A) = \frac{(.15)(.3)}{(.08)(.7) + (.15)(.3)} \approx .446.$$

About a 44% chance that if we drew a bad apple it's a Braeburn.

We can also use a tree diagram to arrive at this answer:

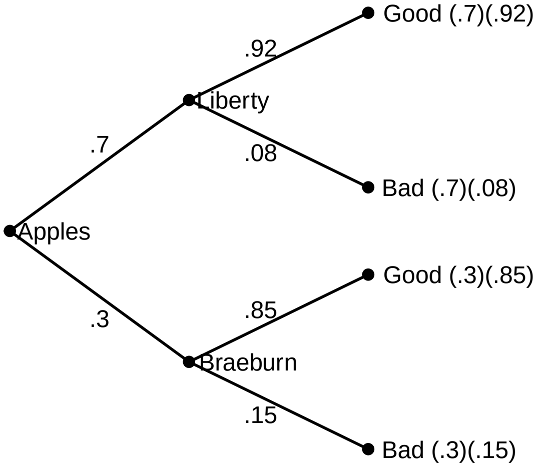


Figure 5.3: Picking an Apple

From the diagram, the probability of picking a bad apple is $(.7)(.08) + (.3)(.15)$ and the probability of picking a bad Braeburn is $(.3)(.15)$, so the probability of having picked a Braeburn, given we picked a bad apple is

$$\frac{(.3)(.15)}{(.7)(.08) + (.3)(.15)}.$$

Example 5.12. Two methods, A and B are available for teaching a certain skill at a factory. The failure rate for A is 20%, and for B is 10%. However, B is more expensive and is used only 30% of the time (A is used the other 70%). A worker was taught the skill by one of the methods but failed to learn it correctly. What is the probability they were taught by Method A?

Let S denote the sample space of all workers who have been trained in this skill.

We have this partition of S:

- $A =$ those taught by method A
- $B =$ those taught by method B.

We also have $F =$ those who fail to learn the skill correctly.

We want: $P(A \mid F)$.

We know: $P(A) = .7$, $P(B) = .3$, $P(F \mid A) = .2$, and $P(F \mid B) = .1$.

Using Bayes' Rule,

$$\begin{aligned}
P(A \mid F) &= \frac{P(A) \cdot P(F \mid A)}{P(A) \cdot P(F \mid A) + P(B) \cdot P(F \mid B)} \\
&= \frac{(.7)(.2)}{(.7)(.2) + (.3)(.1)} \\
&\approx .82.
\end{aligned}$$

Given a worker failed to learn the skill, there is about an 82% chance they had been taught by Method A.

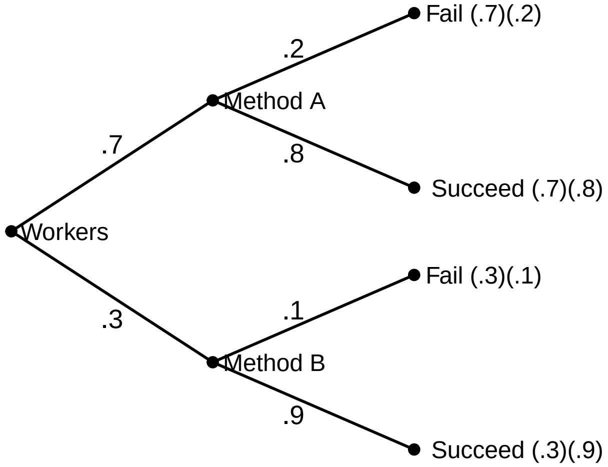


Figure 5.4: Learning by a Method

Chapter 6

Discrete Random Variables

Recall, a random variable X is a real-valued function defined over a sample space associated with a chance experiment.

The **space of** X is the set of possible outcomes for X, and a **probability model** for X is an assignment $p(x)$ to each x in the space of X such that

- each $p(x) \geq 0$, and
- the sum of all the $p(x)$ equals 1.

Let's look at some examples.

Example 6.1. Five balls numbered 1 through 5 are placed in a hat. Two balls are randomly selected without replacement. We consider two random variables associated to this chance experiment:

- X is the largest of the two selected balls, and
- Y is the sum of the two selected balls.

Find the space of X, the space of Y, and reasonable probability models for both random variables.

OK, drawing two balls from five, without replacement, we have $\binom{5}{2} = 10$ possible outcomes. These 10 possible outcomes form the sample space associated with the chance experiment, and we assume each of these 10 outcomes is equally likely.

We can brute-force our answers by following the sample-point method: list all the sample points, and go!

Table 6.1: Two random variables associated with drawing two balls from a hat.

S	{1,2}	{1,3}	{1,4}	{1,5}	{2,3}	{2,4}	{2,5}	{3,4}	{3,5}	{4,5}
X	2	3	4	5	3	4	5	4	5	5
Y	3	4	5	6	5	6	7	7	8	9

Again, we assume each of the 10 elements in S is equally likely, so the probability that $X = 4$, say, will be the $3/10$ since X takes the value 4 for 3 of the 10 elements in S.

The probability model for X in table form:

x	2	3	4	5
$p(x)$	.1	.2	.3	.4

The probability model for Y in table form:

y	3	4	5	6	7	8	9
$p(y)$	.1	.1	.2	.2	.2	.1	.1

Observe that in each case, we have a valid probability model. Each probability is non-negative, and $\sum_x p(x) = 1$ and $\sum_y p(y) = 1$.

Example 6.2. Let X equal the number of rolls of a 6-sided die needed to roll your first 4.

Find the space of X and give a reasonable probability model for X.

The space of X is $\mathbb{N} = \{1, 2, 3, ...\}$.

We assume each of the 6 values is equally likely on any given roll, and that the values are independent from roll to roll. So, the probability of rolling a 4 is $1/6$, and the probability of not rolling a 4 is $5/6$. Then, the probability of rolling our first 4 on roll x, for each $x \geq 1$, is

$$P(X = x) = \left(\frac{5}{6}\right)^{x-1} \cdot \frac{1}{6}.$$

Is this a valid probability model? Certainly each probability is non-negative. Do they all sum to 1? This requires the geometric series formula from Calc II to check:

$$\begin{aligned}\sum_{x=1}^{\infty} \left(\frac{5}{6}\right)^{x-1} \frac{1}{6} &= \frac{1}{6} \sum_{x=0}^{\infty} \left(\frac{5}{6}\right)^{x} \\ &= \frac{1}{6} \cdot \frac{1}{1 - 5/6} \\ &= 1.\end{aligned}$$

Yes!

6.1 Expected Value

Recall, a random variable is a real-valued function defined over a sample space, usually denoted by X or Y, and X is **discrete** if the space of X is finite or countably infinite.

Definition 6.1. If X is a discrete random variable with probability function $p(x)$, then the **expected value of** X, denoted $E(X)$, is

$$E(X) = \sum_{\text{all } x} x \cdot p(x).$$

The expected value $E(X)$ is also called the **mean of** X, and is often denoted as μ_X, or μ if the random variable X is understood.

Example 6.3. In example 6.1 we defined two random variables associated to the experiment of drawing two balls (numbered from 1 to 5) out of a hat.

The expected value of X, the larger value of the two drawn, is

$$E(X) = 2 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.3 + \cdot 0.4 = 4.$$

So, we should expect that after a large number of repetitions of this game the average value of X is about 4.

The expected value of Y, the sum of the two values drawn, is

$$E(Y) = 3 \cdot .1 + 4 \cdot .1 + 5 \cdot .2 + 6 \cdot .2 + 7 \cdot .2 + 8 \cdot .1 + 9 \cdot .1 = 6.$$

. We should expect the average value of Y to be about 6 after a large number of repetitions of this game.

In Example 6.2, the expected number of rolls needed to obtain a 4 is an infinite series:

$$E(X) = \sum_{x=1}^{\infty} x \cdot (5/6)^{x-1} \cdot (1/6).$$

which requires Calc II techniques to evaluate. We do this review in Section 7.2, but will mention here that this infinite sum is 6. That is, the expected value for the number of rolls to get our first 4 turns out to be 6.

Example 6.4 (Chuck-a-luck)**.** The game Chuck-a-luck works like this. Roll 3 dice after choosing a number (1-6). If your chosen number comes up once, you win \$1. If it comes up twice, you win \$2. If it comes up three times, you win \$5. If it doesn't come up at all, you lose \$1.

> Would you expect to win in the long run if you played this game lots of times?

Let's frame Chuck-a-luck as follows:

- We have the chance experiment of rolling 3 dice. We assume the three dice are distinct colors (red, blue, green).
- We have sample space

  $$S = \{(r, b, g) \mid 1 \leq r, b, g \leq 6\}$$

  (r is the value of the red die, b is the value of the blue die, and g is the value of the green die).
- The size of the sample space is $|S| = 6^3 = 216$, and we assume each of these 216 outcomes is equally likely.

- For the sake of argument, let's say that our chosen number is 4.
- We define the random variable X to be the number of 4s we roll.
- The space of X is $\{0, 1, 2, 3\}$.

Now let's find the probability model for X, one value of x at a time.

- $p(0)$ is the probability that all 3 dice are not 4, which is $(5/6)^3 = 125/216$.
- $p(3)$ is the probability that all 3 dice are 4, which is $(1/6)^3 = 1/216$.
- For $p(1)$ we have three cases to consider (based on which die comes up 4):
  - Red die is 4, the others aren't. This probability is $(1/6) \cdot (5/6) \cdot (5/6)$.
  - Blue die is 4, the others aren't: $(5/6) \cdot (1/6) \cdot (5/6)$.
  - Green die is 4, the others aren't: $(5/6) \cdot (5/6) \cdot (1/6)$. So, $p(1) = 3 \cdot (1/6) \cdot (5/6)^2 = 75/216$
- $p(2)$ is found by three cases as well, depending on which die is not 4. We find $p(2) = 3 \cdot (1/6)^2 \cdot (5/6) = 15/216$.

We can check that these four probabilities add to 1. Check! To summarize, X has probability function given here in table form:

x	0	1	2	3
$p(x)$	125/216	75/216	15/216	1/216

With the probability model in hand, we can compute the expected value of X:

$$E(X) = 0 \cdot (125/216) + 1 \cdot (75/216) + 2 \cdot (15/216) + 3 \cdot (1/216) = 1/2.$$

We interpret this result as follows: In a large number of games played, we would expect, on average, 0.5 fours to come up per game played.

This expected value of X doesn't actually answer the original question in this example. Should we expect to win *money* in the long run? Our calculation hasn't taken into account the dollar amounts attached to the various outcomes. These dollar amounts (1 if $X = 1$, 2 if $X = 2$, 5 if $X = 3$ and -1 if $X = 0$), mathematically, describe a function of X (input is a value from the space of X, output is a dollar amount). To decide whether we should expect to win money in the long run, we want to calculate the expected value of *a function* of the random variable X.

We can estimate our average expected winnings by playing the game repeatedly, we could play 100 times, or a 1000 times, and see how we do on average (hello R!). Or we can turn to the computation of the theoretical expected winnings per turn via the following theorem.

Theorem 6.1. *Let X be a discrete random variable with probability function $p(x)$, and suppose $g(X)$ is a real-valued function of X. Then the expected value of $g(X)$ is*

$$E(g(X)) = \sum_{all\ x} g(x) \cdot p(x).$$

Example 6.5 (Chuck-a-luck for a living?)**.** Now we focus on our winnings, W. W is a function of X, and we summarize this function by adding the winnings to the probability model table:

x	0	1	2	3
$p(x)$	125/216	75/216	15/216	1/216
w	-1	1	2	5

Then, Theorem 6.1 says

$$E(W) = -1 \cdot (125/216) + 1 \cdot (75/216) + 2 \cdot (15/216) + 5 \cdot (1/216)$$

which evaluates to $-15/216 \approx -.07$.

In the long run we should expect, on average, to lose 7 cents per game. So, yes, we should definitely play Chuck-a-luck, it's cheap entertainment! If you figure a game pace of 1 roll per minute, it will cost you about \$4.20 per hour to play!!

As an aside, here's code to simulate Chuck-a-luck in R a bunch of times (betting on 4), storing the results of each game, and then printing the table of the results followed by the average winnings of all the trials.

```r
chosen_number = 4
X = c(0,1,2,3)#space of X
W = c(-1,1,2,5)#winnings based on X
trials = 2160
results = c() #stores winnings each trial
for (i in 1:trials){
  rolls = sample(1:6,3,replace=TRUE)
  x = sum(rolls == chosen_number)
  w = W[which(X == x)]
  results[i] = w
}
print(table(results))
```

```
## results
##   -1    1    2    5
## 1239  750  161   10
```

```r
print(mean(results))
```

```
## [1] -0.05416667
```

Example 6.6. Suppose X has probability model

x	0	10	20	30	40	50
$p(x)$	.1	.2	.1	.1	.4	.1

Perhaps X models my scores per roll in skee ball? In any event, let's compute $E(3X^2 + 1)$:

$$\begin{aligned} E(3X^2+1) &= \sum_{\text{all } x} (3x^2+1) \cdot p(x) \\ &= 0 \cdot .1 + 301 \cdot .2 + 1201 \cdot .1 + 2701 \cdot .1 + 4801 \cdot .4 + 7501 \cdot .1 \\ &= 3120 \end{aligned}$$

6.2 Variance

Definition 6.2. If X is a random variable with expected value $E(X) = \mu$, the **variance of** X, denoted $V(X)$, is

$$V(X) = E((X - \mu)^2).$$

The variance of X is often denoted σ_X^2, or σ^2 if the random variable is understood. Also, $\sqrt{V(X)}$, denoted σ_X or σ, is called the **standard deviation of** X.

Example 6.7. Suppose X and Y have the following random probability models.

x	0	1	2	3	4
$p(x)$	.2	.3	.3	.1	.1

y	0	1	2	3	4
$p(y)$	.6	0	0	0	.4

Compute the expected value and variance for each random variable.

The expected value of X is

$$E(X) = 0 + (1)(.3) + (2)(.3) + (3)(.1) + (4)(.1) = 1.6,$$

and the expected value of Y is

$$E(Y) = 0 + (4)(.4) = 1.6,$$

so the two random variables have the same mean: $E(X) = E(Y)$, or, using the alternate notation, $\mu_X = \mu_Y$.

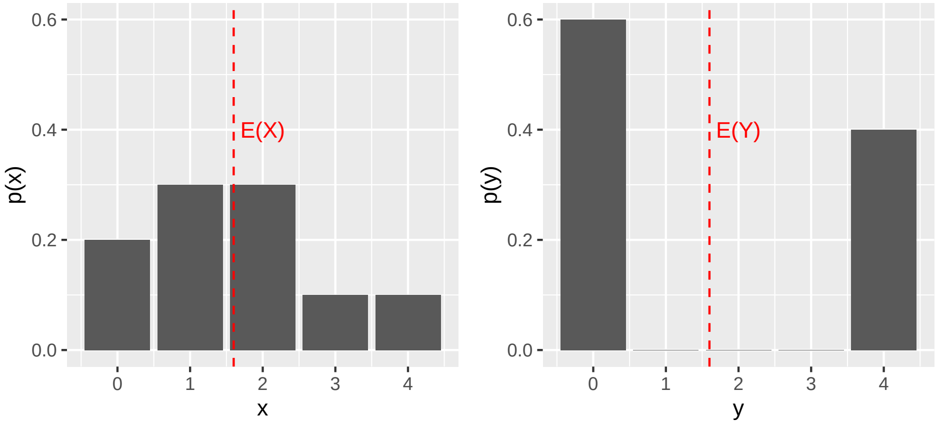
The variance of X:

$$V(X) = E((X - \mu_X)^2) = \sum_{x=0}^{4}[(x - 1.6)^2 \cdot p(x)] = 1.44.$$

The variance of Y is larger:

$$V(Y) = E((Y - \mu_Y)^2) = \sum_{y=0}^{4}[(y - 1.6)^2 \cdot p(y)] = 3.84.$$

The variance of a random variable increases as more of the distribution lies further from μ. In this example, more of the probability distribution for Y lies farther away from 1.6, than the distribution of X does away from its mean (also 1.6).



Example 6.8 (Infinite Variance?)**.** There exist discrete random variables with finite mean and infinite variance. Here's one: Recall that the p-series

$$\sum_{n=1}^{\infty} \frac{1}{n^p}$$

converges for $p > 1$, and diverges when $p = 1$. Let's suppose the series

$$\sum_{n=1}^{\infty} \frac{1}{n^3} = c.$$

(In fact, c equals a known constant close to 1.2, called Apery's constant after the mathematician who proved this constant is irrational.) Consider the discrete random variable X whose distribution function is given by

$$p(x) = \frac{1}{cx^3} \quad \text{for} \quad x = 1, 2, 3, \ldots .$$

Then,

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} x \cdot \frac{1}{cx^3} \\ &= \frac{1}{c} \sum x = 1^{\infty} \frac{1}{x^2} \\ &= \frac{\pi^2}{6c}, \end{aligned}$$

since the p-series $\sum_{n=1}^{\infty}(1/n^2) = \pi^2/6$. So, $E(X)$ exists as a finite number.

However,

$$\begin{aligned} E(X^2) &= \sum_{x=1}^{\infty} x^2 \cdot \frac{1}{cx^3} \\ &= \frac{1}{c} \sum x = 1^{\infty} \frac{1}{x}, \end{aligned}$$

which diverges as a multiple of the harmonic series. So, $V(X)$ does not exist as a finite number.

6.3 Properties of Expected Value

Theorem 6.2. *Suppose X is a discrete random variable, $c \in \mathbb{R}$ is a constant, and g, g_1, and g_2 are functions of X.*

1. $E(c) = c$.
2. $E(c \cdot g(X)) = cE(g(X))$.
3. $E(g_1(X) \pm g_2(X)) = E(g_1(X) \pm g_2(X))$.

These properties can help us evaluate expected values without having to sum over all x.

For instance, suppose we know X is a discrete random variable with expected value $E(X) = 1.6$.

Then

$$\begin{aligned} E(4+3X) &= E(4) + E(3X) && \text{by property 3} \\ &= 4 + 3E(X) && \text{by properties 1 and 2} \\ &= 4 + 3 \cdot 1.6 && \text{since } E(X) = 1.6 \\ &= 8.8. && \text{Nice.} \end{aligned}$$

Let's take the time to prove these properties. Each of them essentially follows by properties of summations.

Proof.

1. Given a constant c, we can view this constant as a function of X, say $f(x) = c$. Then

$$\begin{aligned} E(c) &= \sum_{\text{all } x} c \cdot p(x) \\ &= c \sum_{\text{all } x} p(x) \end{aligned}$$

Since the sum over all x of $p(x)$ is 1 for any probability model, the result follows.

2. Here appeal to Theorem 6.1} and arithmetic:

$$\begin{aligned} E(c \cdot g(X)) &= \sum_{\text{all } x} c \cdot g(x) \cdot p(x) \\ &= c \sum_{\text{all } x} g(x)p(x) && \text{by arithmetic} \\ &= cE(g(X)) \end{aligned}$$

3. Here we also appeal to Theorem 6.1} and arithmetic:

$$\begin{aligned} E(g_1(x) \pm g_2(x)) &= \sum_{\text{all } x} (g_1(x) \pm g_2(x)) \cdot p(x) \\ &= \sum_{\text{all } x} (g_1(x)p(x) \pm g_2(x)p(x)) && \text{by arithmetic} \\ &= \sum_{\text{all } x} g_1(x)p(x) \pm \sum_{\text{all } x} g_2(x)p(x) && \text{by arithmetic} \\ &= E(g_1(X)) \pm E(g_2(X)) \end{aligned}$$

□

Example 6.9. The number N of residential homes that a fire company can serve depends on the distance r (in city blocks) that a fire engine can cover in a fixed period of time. If we assume that N is proportional to the area of a circle R blocks from the fire house, then

$$N = k\pi R^2,$$

where k is a constant, and R is a random variable. For a particular fire company, $k = 8$, and the probability function for R is

r	21	22	23	24	25	26
$p(r)$	.05	.20	.30	.25	.15	.05

Find $E(N)$, the expected number of homes that the fire department can serve.

Well,

$$E(N) = E(8\pi R^2) = 8\pi E(R^2),$$

so

$$\begin{aligned} E(N) &= 8\pi\left(21^2 \cdot .05 + 22^2 \cdot .20 + 23^2 \cdot .30 + 24^2 \cdot .25 + 25^2 \cdot .15 + 26^2 \cdot .05\right) \\ &= 8\pi(549.1) \\ &\approx 13,800 \text{ homes} \end{aligned}$$

Theorem 6.3 (Useful Variance Formula)**.** *Let X be a discrete random variable with probability function $p(x)$ and expected value $E(X) = \mu$. Then*

$$V(X) = E(X^2) - \mu^2.$$

Proof. By definition,

$$\begin{aligned} V(X) &= E((X-\mu)^2) \\ &= E(X^2 - 2\mu X + \mu^2) && \text{by expanding} \\ &= E(X^2) - E(2\mu X) + E(\mu^2) && \text{by E() Property 3} \\ &= E(X^2) - 2\mu E(X) + \mu^2 && \text{by E() Properties 2 and 1} \\ &= E(X^2) - 2\mu^2 + \mu^2 && \text{since } E(X) = \mu \\ V(X) &= E(X^2) - \mu^2 \end{aligned}$$

□

This alternate formula for variance also provides us with a way to compute $E(X^2)$ from the expected value and variance of a random variable:

$$E(X^2) = V(X) + \mu^2,$$

or using the alternate variance notation:

$$E(X^2) = \sigma^2 + \mu^2.$$

Example 6.10. Suppose X is a discrete random variable with expected value $\mu = 5$ and variance $\sigma^2 = 6$. Find $E(3 + 2X + 4X^2)$.

Well,

$$\begin{aligned} E(3 + 2X + 4X^2) &= E(3) + 2E(X) + 4E(X^2) \\ &= 3 + 2 \cdot \mu + 4[\sigma^2 + \mu^2] \\ &= 3 + 2 \cdot 5 + 4[6 + 5^2] \\ &= 3 + 10 + 124 \\ &= 137. \end{aligned}$$

6.4 Tchebysheff's Theorem

Theorem 6.4 (Tchebysheff's Inequality)**.** *Let X be a random variable with mean $E(X) = \mu$ and finite variance $V(X) = \sigma^2 > 0$. Then for any constant $k > 0$,*

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

Equivalently,

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

For instance, if $k = 2$, Tchebbysheff's inequality says that for any random variable X, the probability that X takes a value that is within 2 standard deviations of the mean is at least .75. For many distributions, this proability is closer to .95, but .75 holds for *all* distributions. Of course, letting $k = 1$ in Tchebbysheff's inequality gives us the trivially true statement that the probability that X takes a value within 1 standard deviation of the mean is at least 0.

Proof. We prove Tchebbysheff's inequality in the case for a discrete random variable, and we come back to this theorem after defining continuous random variables.

Let $k > 0$ be given.

Then

$$V(X) = \sum_{\text{all } x} (x - \mu)^2 p(x),$$

by the definition of variance. We can partition the space of X into three disjoint sets, depending on the location of x relative to $\mu \pm k\sigma$:

$$V(X) = \sum_{\text{all } x \leq \mu - k\sigma} (x - \mu)^2 p(x) + \sum_{\text{all } x \text{ s.t. } |x - \mu| < k\sigma} (x - \mu)^2 p(x) + \sum_{\text{all } x \geq \mu + k\sigma} (x - \mu)^2 p(x)$$

Each of these three sums is non-negative, and for the first and third sums we can also say that $(x - \mu)^2 \geq k^2\sigma^2$ for all x in the given range, so it follows that

$$V(x) \geq \sum_{\text{all } x \leq \mu - k\sigma} k^2\sigma^2 p(x) + 0 + \sum_{\text{all } x \geq \mu + k\sigma} k^2\sigma^2 p(x).$$

So,

$$\begin{aligned}
\sigma^2 &\geq \sum_{\text{all } x \leq \mu - k\sigma} k^2\sigma^2 p(x) + 0 + \sum_{\text{all } x \geq \mu + k\sigma} k^2\sigma^2 p(x) \\
&= k^2\sigma^2 \left(\sum_{\text{all } x \leq \mu - k\sigma} p(x) + \sum_{\text{all } x \geq \mu + k\sigma} p(x) \right) \\
&= k^2\sigma^2 \left(P(X \leq \mu - k\sigma) + P(X \geq \mu + k\sigma) \right) \\
&= k^2\sigma^2 P(|X - \mu| \geq k\sigma)
\end{aligned}$$

Dividing both sides of the inequality by the positive value $k^2\sigma^2$ gives us the result:

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

□

Example 6.11. Suppose X is a random variable with $E(X) = 70$ and $V(X) = 25$, so $\mu = 70$ and $\sigma = \sqrt{25} = 5$. According to Tchebbysheff's inequality with $k = 2$, the probability that X takes a value between 60 and 80 is at least 3/4. Setting $k = 3$, we find the probability that X takes a value between 55 and 85 is at least 8/9.

Again, for many distributions, the probability of being within 2 standard deviations of the mean is much higher than .75 (often about .95, in fact), and the probability of being within 3 standard deviations of the mean is much higher than 8/9 (often about .99).

Here's a distribution, however, that shows the bound in Tchebbysheff's inequality cannot be improved.

Example 6.12. We show that there exists a probability distribution for which $P(|X - \mu| < 2\sigma) = .75$.

Consider the discrete random variable X whose probability distribution function is

x	-1	0	1
$p(x)$	.125	.75	.125

Then

$$\mu = E(X) = (-1)(.125) + (0)(.75) + (1)(.125) = 0,$$

and

$$\begin{aligned}
\sigma^2 &= E(X^2) - \mu^2 \\
&= (-1)^2(.125) + 0^2(.75) + 1^2(.125) \\
&= .25,
\end{aligned}$$

So, $\sigma = 0.5$, and

$$\begin{aligned}
P(|X - \mu| < 2\sigma) &= P(|X| < 1) \\
&= P(-1 < X < 1) \\
&= P(X = 0) \qquad \text{since the space of } X \text{ is } \{-1, 0, 1\} \\
&= .75.
\end{aligned}$$

Thus, there exists a discrete random variable for which $P(|X-\mu|<2\sigma)=.75$. In fact, for any $k>0$ the probability distribution given by

x	-1	0	1
$p(x)$	$\frac{1}{2k^2}$	$1-\frac{1}{k^2}$	$\frac{1}{2k^2}$

will satisfy $P(|X-\mu|<2\sigma)=1-\frac{1}{k^2}$, demonstrating that the bound in Tchebbyshef's inequality can not be increased.

Chapter 7

Important Discrete Random Variables

In this chapter we introduce the following well-known discrete random variables: binomial, geometric, Poisson, negative binomial, and hypergeometric. In examples we work through, it will from time to time be convenient to compute probabilities in R. Appendix C contains details about the commands in R useful for doing so.

7.1 Binomial Distributions

It all begins with a Bernoulli trial.

Definition 7.1. A **Bernoulli trial** is a chance experiment with two distinct possible outcomes, "success" and "failure". Typically, we let p denote the probability of success, and q denote the probability of failure (where $q = 1 - p$).

Some examples:

1. Roll a 6-sided die and define success to be "roll a 4", and failure to be "don't roll a 4". Here $p = 1/6$ and $q = 5/6$.
2. Pick a name out of a hat with n names. Success: pick Oriana's name; Failure: do not pick Oriana's name. Here $p = 1/n$ and $q = (n - 1)/n$.
3. Test a person for a particular disease. In medical tests such as these, "success" is often used to describe a positive test (meaning the person tests positive for the disease), and "failure" means the person tests negative for the disease.

Definition 7.2 (Binomial Distribution)**.** Define the random variable X to equal the number of successes in n independent, identical Bernoulli trials with probability of success on any given trial equal to p. Then X is called a **binomial distribution with parameters** n **and** p, and X is denoted `binom`(n, p).

The space of X equals $\{0, 1, \ldots, n\}$, and for $x = 0, 1, \ldots, n$,

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Does this probability distribution make sense? To find the probability of exactly x successes in n independent Bernoulli trials, we first choose which x of the n trials will be successes (and we have $\binom{n}{x}$ choices here). Then, we know the probability of success on each of these x trials is p, and the probability of failure on each of the other $n-x$ trials is $q = 1-p$. Since the trials are independent, the probability of getting exactly x successes and $n-x$ failures is the product $\binom{n}{x}p^x(1-p)^{n-x}$.

Furthermore, by the Binomial Theorem (4.1),

$$\sum_{x=0}^{n} p(x) = (p + (1-p))^n = 1.$$

Theorem 7.1. *If X is* `binom`(n, p),

$$E(X) = np \quad \textit{and} \quad V(X) = np(1-p).$$

Proof. From the definition of expected value,

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x \cdot p(x) \\
&= \sum_{x=1}^{n} x \cdot p(x) && \text{since the } x=0 \text{ term is } 0 \\
&= \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(n-x)!(x-1)!} p^x (1-p)^{n-x} && \text{cancelling an } x \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(n-x)!(x-1)!} p^{x-1} (1-p)^{n-x} && \text{pull out } n \text{ from } n! \text{ and one } p \\
&= np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)}
\end{aligned}$$

Hey! The summation term equals 1 since it is the sum of all the probabilities in a `binom`$(n-1, p)$ distrtibution!

Thus $E(X) = np$.

To prove the $V(X)$ formula, it is helpful to first observe the following

$$E(X(X-1)) = E(X^2 - X) = E(X^2) - E(X),$$

so

$$E(X^2) = E(X(X-1)) + E(X).$$

We computed $E(X)$ above, and $E(X(X-1))$ can be determined similarly:

$$\begin{aligned}
E(X(X-1)) &= \sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum_{x=2}^{n} x(x-1)\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum_{x=2}^{n} x(x-1)\frac{n!}{(n-x)!x!}p^x(1-p)^{n-x} \\
&= \sum_{x=2}^{n} \frac{n!}{(n-x)!(x-2)!}p^x(1-p)^{n-x} \\
&= n(n-1)p^2\sum_{x=2}^{n} \frac{(n-2)!}{(n-x)!(x-2)!}p^{x-2}(1-p)^{n-x} \\
&= n(n-1)p^2\sum_{x=2}^{n} \binom{n-2}{x-2}p^{x-2}(1-p)^{(n-2)-(x-2)}
\end{aligned}$$

This summation also equals one, since it is the sum of all the probabilities in a `binom`$(n-2,p)$ distribution, so $E(X(X-1)) = n(n-1)p^2$, and it follows that

$$E(X^2) = E(X(X-1)) + E(X) = n(n-1)p^2 + np.$$

Finally,

$$\begin{aligned}
V(X) &= E(X^2) - E(X)^2 \\
&= n(n-1)p^2 + np - (np)^2 \\
&= n^2p^2 - np^2 + np - n^2p^2 \\
&= np(1-p)
\end{aligned}$$

□

Example 7.1 (Guessing on a multiple choice test)**.** A multiple choice exam has 15 questions. Each question has 4 possible answers. If a student answers each question with a random guess, what is the probability they will score 10 or higher?

Let X = the score on the test. Then, X is `binom`$(n = 15, p = 1/4)$.

Now,

$$P(X \geq 10) = \sum_{x=10}^{15}\binom{15}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{15-x} \approx .0008.$$

Calculating this sum in R

R has a nice command for calculating cumulative probabilities of the form $P(X \leq x)$.If X is `binom`(n,p), then $P(X \leq x)$ is calculated in R by `pbinom(x,n,p)`. So, in the case of the multiple choice test, $P(X \geq 10) = 1 - P(X \leq 9) =$ `1-pbinom(9,15,1/4)`. See Appendix C) for more information on using R to calculate probabilities for the important discrete distributions we encounter in this class.

Back to the test, they have less than a 1 in a thousand chance of scoring a 10 or better if they are truly guessing. Note also that the mean for X is $E(X) = 15 \cdot \frac{1}{4} = 3.75$.

What if they can eliminate one of the choices on each problem, and randomly guess between the remaining three choices on each problem? Are they likely to do better? If we let Y denote the score on the test following this approach, then Y is $\texttt{binom}(n = 15, p = 1/3)$, and the probability of scoring 10 or greater ends up about 10 times better than it was before, but still miniscule:

$$P(Y \geq 10) = \sum_{y=10}^{15} \binom{15}{y} \left(\frac{1}{3}\right)^{y} \left(\frac{2}{3}\right)^{15-y} \approx .0085,$$

The mean score is now $E(Y) = 15 \cdot \frac{1}{3} = 5$. Ok, not great, but it is better, I guess.

Example 7.2 (Pay the meter?)**.** This example is adapted from an exercise in the Grinstead-Snell text. Flint never puts money in a 25-cent parking meter downtown. He assumes that there is a probability of .03 that he will be caught. The first offense costs nothing, the second costs 10 dollars, and subsequent offenses cost 25 dollars each.

How does the expected cost of parking 100 times without paying the meter compare with the cost of paying the meter each time?

Assume parking events are independent, identical Bernoulli trials with probability $p = .03$ of getting a ticket. Then the random variable X counting the number of tickets in 100 trials is $\texttt{binom}(100, .03)$, and we note that $E(X) = 100 \cdot (.03) = 3$.

In deciding whether to pay the meter, one idea is to consider the cost associated to the expected number of tickets, which would be \$35. This amount is higher than the \$25 Flint would pay by chucking in a quarter each time. But this approach doesn't give the full picture.

Instead, let's determine the expected cost associated to parking 100 times without paying the meter. If Flint never pays the meter, the parking cost C of these 100 trials is the following function of X:

$$C(x) = \begin{cases} 0 & \text{if } x = 0, 1 \\ 10 & \text{if } x = 2 \\ 10 + 25(x-2) & \text{if } x \geq 3 \end{cases}$$

Then the expected cost associated with these 100 trials, $E(C)$, is

$$\begin{aligned}
E(C) &= \sum_{x=0}^{100} C(x) \cdot p(x) \\
&= C(2) \cdot p(2) + \sum_{x=3}^{100} (10 + 25 \cdot (x-2)) \cdot p(x) \\
&= 10 \cdot p(2) + \sum_{x=3}^{100} (25x - 40) \cdot p(x) \\
&= 10 \cdot p(2) + \left(E(25X - 40) - \sum_{x=0}^{2} (25x - 40) \cdot p(x) \right) \\
&= 10 \cdot p(2) + [25 \cdot E(X) - 40] - (-40 \cdot p(0) - 15 \cdot p(1) + 10 * p(2)) \\
&= [25E(X) - 40] + 40 \cdot p(0) + 15 \cdot p(1) \\
&= [(25 \cdot 3) - 40] + 40 \cdot (.97)^{100} + 15 \cdot 100(.03)(.97)^{99} \\
&\approx 39.10.
\end{aligned}$$

Yes, Flint is better off putting a quarter in the meter each time for a cost of $25 in parking. But never tell Flint the odds.

Example 7.3 (Drilling for Oil)**.** An oil exploration firm in the 1970s wants to finance 10 drilling explorations. They figure each exploration has a probability of success (finding oil) equal to 0.1, and that the 10 operations are independent (success in one is independent of success in any other). The company has $50,000 in fixed costs prior to doing its first exploration, and anticipates a cost of $150,000 for each unsuccessful exploration, and a cost of $300,000 for each successful exploration. Find the expected total cost to the firm for its 10 explorations.

Let X = number of successful explorations. Then X is $\texttt{binom}(10, .1)$, and $E(X) = 10 \cdot .1 = 1$. The cost C (in thousands of dollars) can be expressed as a linear function of X:

$$\begin{aligned}
C &= 50 + 150(10 - X) + 300X \\
&= 1550 + 150X.
\end{aligned}$$

It follows from properties of expected value that

$$\begin{aligned}
E(C) &= E(1550 + 150X) \\
&= E(1550) + 150E(X) \\
&= 1550 + 150 \cdot 1 \\
&= 1700.
\end{aligned}$$

So the expected cost is $1.7 million dollars.

Example 7.4 (AI Generated Example)**.** In a galaxy not so far away, there is a soccer ball factory run by enthusiastic, if somewhat clumsy, aliens. The factory manager, Zorg, claims that only 5% of the soccer balls they produce end up being “special” (read: defective). The company’s quality control inspector, an alien named Blurp, is highly skeptical and decides to randomly select 20 soccer balls from a day's production to test this claim.

Questions:

1. If Zorg's claim is correct, what is the probability that exactly 2 of the 20 selected soccer balls are "special"?
2. If Zorg's claim is correct, what is the probability that at most 1 of the 20 selected soccer balls is "special"?
3. If Zorg's claim is correct, what is the probability that more than 3 of the 20 selected soccer balls are "special"?

Solution:

Let X denote the number of "special" soccer balls in the randomly selected set of 20 balls. Then, X is $\texttt{binom}(n = 20, p = .05)$

The first question asks for $P(X = 2)$ which is

$$P(X = 2) = \binom{20}{2} p^2 \cdot (1 - p)^{18} \approx .189.$$

The second question asks for

$$P(X \leq 1) = \sum_{x=0}^{1} \binom{20}{x} p^x (1 - p)^{20-x},$$

and using R, we see that $P(X \leq 1) =$ `pbinom(1,20,.05)` ≈ 0.736.

The third question asks for

$$P(X > 3) = 1 - P(X \leq 3),$$

which we can evaluate in R with `1-pbinom(3,20,.05)` ≈ 0.016.

Example 7.5. Suppose X is $\texttt{binom}(60, 1/4)$. Then $\mu = 60 \cdot \frac{1}{4} = 15$ and $\sigma^2 = 60 \cdot \frac{1}{4} \cdot \frac{3}{4} = 11.25$. Tchebbysheff says at least 75% of the distribution is within 2 standard deviations of the mean, so in this case, at least 75% of the distribution is between $15 - 2 \cdot \sqrt{11.25}$ and $15 + 2 \cdot \sqrt{11.25}$, or between 8.3 and 21.7. The actual percentage is closer to 95%, and it can be found by summing all $p(x)$ for x between 8.29 and 21.7. This sum is calculated in R by

```
pbinom(21.7,60,1/4)-pbinom(8.29,60,1/4)
```

```
## [1] 0.9489842
```

7.2 Geometric Distributions

Suppose we have a sequence of Bernoulli trials, independent, with probability of success p on each trial. Let X equal the number of trials up to and including the trial of the first success. Then X is called a **geometric distribution with parameter** p, denoted $\texttt{geom}(p)$.

The probability function for X is given by

$$p(x) = q^{x-1} \cdot p,$$

for $x = 1, 2, 3, \ldots$, where, again, $q = 1 - p$.

Note that

$$\sum_{x=0}^{\infty} p(x) = p + pq + pq^2 + \cdots$$

is a geometric series with $|q| < 1$ so it converges. Moreover, by the geometric series formula,

$$\sum_{n=0}^{\infty} pq^n = \frac{p}{1-q}.$$

Thus, all the $p(x)$ sum to 1, and we have a valid probability function.

More generally, for any non-negative integer k,

$$\begin{aligned}
P(X > k) &= p(X = k+1) + p(X = k+2) + p(X = k+3) + \cdots \\
&= pq^k + pq^{k+1} + pq^{k+2} + \cdots \\
&= pq^k(1 + q + q^2 + \cdots) \\
&= pq^k \cdot \frac{1}{1-q} \\
&= q^k.
\end{aligned}$$

The geometric distribution can be useful when modeling the behavior of lines. For instance, suppose a line of cars waits to pay their parking fee as they exit the airport. It can be reasonable to assume that over a short interval of time (say 10 seconds), the probability that a car arrives is p, and the probability that a car does not arrive is $q = 1 - p$. Then the time T until the next arrival has a geometric distribution, and by the remark above, the probability that no car arrives in the next k time units is $P(T > k) = q^k$.

Theorem 7.2. *If X is* `geom`*(p) for* $0 < p \leq 1$, *then*

$$E(X) = \frac{1}{p} \quad \textit{and} \quad V(X) = \frac{1-p}{p^2}.$$

Before proving this theorem we consider the geometric series one more time.

Geometric Series Intermission

From Calc II we know that the geometric series $\sum q^x$ converges if and only if $-1 < q < 1$, and that

$$\sum_{x=0}^{\infty} q^x = \frac{1}{1-q} \qquad \text{(provided |q|<1)}$$

Thinking of q as a variable, we can differentiate each side with respect to q, and the resulting infinite series will still converge for $-1 < q < 1$.

$$\begin{aligned}
\frac{d}{dq}\left[\sum_{x=0}^{\infty} q^x\right] &= \frac{d}{dq}\left[\frac{1}{1-q}\right] \\
\frac{d}{dq}\left[1+q+q^2+q^3+\cdots\right] &= \frac{d}{dq}\left[\frac{1}{1-q}\right] \\
\left[0+1+2q+3q^2+\cdots\right] &= \frac{1}{(1-q)^2} \\
\sum_{x=1}^{\infty} xq^{x-1} &= \frac{1}{(1-q)^2}.
\end{aligned}$$

This ends the geometric series intermission.

Proof. By definition of expected value,

$$\begin{aligned}
E(X) &= \sum_{x=1}^{\infty} x \cdot q^{x-1} \cdot p \\
&= p \sum_{x=1}^{\infty} x \cdot q^{x-1}.
\end{aligned}$$

But this series is exactly the one for which we found a formula in the intermission above, so

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

We leave the proof that $V(X) = \frac{1-p}{p^2}$ as an exercise. □

Example 7.6 (Message Received)**.**

Assume that, during each second, my junk box receives one email with probability .01 and no email with probability .99. Determine the probability that I will not receive a junk email in the next minute.

We let X count how many seconds (from now) it takes to be sent an email to my junk box, and we assume X is `geom`$(p = .01)$.

Then the probability of not receiving a junk email in the next ten minutes is

$$P(X > 600) = q^{600} \approx .0024.$$

So you're saying there's a chance!

We also note that $E(X) = 1/p = 100$, the average time to get our next junk mail is 100 seconds.

7.3 Negative Binomial Distribution

Suppose we have a sequence of independent Bernoulli trials, each having probability of success p, and we want to know how many trials it takes to get our rth success, where $r \geq 1$.

For instance, if we set $r = 3$, then $X = 7$ for the following sequence of Bernoulli trials (S stands for success, F for failure)

$$FFFSFSSFSFFS\,...$$

since the third S occurs on the 7th trial.

Let X equal the number of trials up to and including the trial of the rth success. Then X is called a **negative binomial distribution with parameters r and p**, denoted $\texttt{nbinom}(r, p)$.

The space of X is $\{r, r+1, r+2, ...\}$ and for x in the space of X the probability function for X is given by

$$p(x) = \binom{r-1}{x-1} p^r (1-p)^{x-r}.$$

Note, that if $r = 1$ we just have the friendly geometric distribution.

Theorem 7.3. *If X is $\texttt{nbinom}(r, p)$ where $0 < p \leq 1$ then*

$$E(X) = \frac{r}{p} \quad \text{and} \quad V(X) = \frac{r(1-p)}{p^2}.$$

Example 7.7.

> If we roll 2 dice and record the sum, how many rolls, on average, will it take to get our 4th 8?

Well, the probability of rolling a sum of 8 is $p = 5/36$ by our 6x6 dice grid in Example 3.3, and the random variable X counting the number of rolls until we get our 4th 8 is $\texttt{nbinom}(r = 4, p = 5/36)$, so

$$E(X) = \frac{4}{5/36} = 28.8.$$

Would you want to make a bet that it will take me more than 25 rolls to get my 4th 8?

We note that

$$V(X) = \frac{4 \cdot (31/36)}{(5/36)^2} \approx 178.6,$$

so the standard deviation is $\sigma = \sqrt{V(X)} \approx 13.36$.

7.4 Hypergeometric Distribution

Here's the scene: We have a finite population with N total elements, and this population can be partitioned into two distinct groups, where

- group 1 has m elements, and
- group 2 has n elements, (so $m + n = N$).

Think: a box of N marbles, m of them are orange and n of them are green.

Suppose we draw a random sample without replacement of size k from the population.

Let X equal the number of elements in the sample of size k that belong to group 1.

Then X is called a **hypergeometric distribution with parameters** m, n, **and** k, denoted $\texttt{hyper}(m, n, k)$. The space of X is either $x = 0, 1, \dots, k$ if $m \geq k$, or $0, 1, \dots, m$ if $m < k$.

For each x in the space of X,

$$p(x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{N}{k}},$$

where $N = m + n$.

The classic "good potatoes/bad potatoes" Example 4.12 has a hypergeometric distribution.

Theorem 7.4. *If X is $\texttt{hyper}(m, n, k)$ then*

$$E(X) = k \cdot \frac{m}{N} \qquad \text{and} \qquad V(X) = k\left(\frac{m}{N}\right)\left(\frac{n}{N}\right)\left(\frac{N-k}{N-1}\right),$$

where $N = m + n$.

Example 7.8.

> Let's say a bag of 120 skittles has 30 orange ones. If we pick 10 at random, what is the probability that we get more than 5 orange ones?

Let X denote the number of orange skittles in our sample. Then X is $\texttt{hyper}(30, 90, 10)$, and

$$P(X > 5) = \sum_{x=6}^{10} \frac{\binom{30}{x}\binom{90}{10-x}}{\binom{120}{10}} \approx .0153.$$

We note that in this example $E(X) = 10 \cdot \frac{30}{120} = 2.5$.

7.5 Poisson Distribution

The Scene

The Poisson probability distribution can provide a good model for the number of occurrences X of a rare event in time, space, or some other unit of measure. A Poisson random variable X has one parameter, λ, which is the average number of occurrences of the rare event in the indicated time (or space, etc.)

Some examples that might be well-modeled by a Poisson distribution:

- the number of customers going through a check-out lane in a grocery store per hour;
- the number of typos per page in a book;
- the number of goals scored in a World Cup soccer game;
- the number of chocolate chips per cookie in a big batch;
- the number of pitches per baseball in a baseball game;

Definition 7.3 (Poisson Distribution)**.** A random variable X has a **Poisson probability distribution** with parameter $\lambda > 0$, denoted `Poisson`(λ), if and only if

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{for } x = 0, 1, 2, \ldots .$$

We take the time to explain how this probability density function does actually model such things, but first let's check that it is a valid probability density function, and then find $E(X)$ and $V(X)$.

First, since $\lambda > 0$, each $p(x)$ is non-negative. Also, recall the Calc II power series formula for e^λ for any real number λ is

$$e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!},$$

so we can see that all the probabilities in this distribution sum to 1:

$$\begin{aligned} \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \cdot e^{\lambda} \\ &= 1. \end{aligned}$$

Theorem 7.5. *If X is* `Poisson`(λ), $E(X) = \lambda$ *and* $V(X) = \lambda$.

Proof. We tackle the mean first.

$$\begin{aligned}
E(X) &= \sum_{x=0}^{\infty} x \cdot p(x) \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} && \text{since } x = 1 \text{ term is 0} \\
&= e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} && \text{since } \frac{x}{x!} = \frac{1}{(x-1)!} \\
&= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} && \text{pulling out one } \lambda \\
&= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{(k)!} && \text{letting } k = x-1 \\
&= \lambda e^{-\lambda} \cdot e^{\lambda} && \text{power series for } e^{\lambda} \\
&= \lambda.
\end{aligned}$$

Thus, $\mu = \lambda$.

For $V(X)$, we first find $E(X(X-1))$ in much the same way as we found $E(X)$:

$$\begin{aligned}
E(X(X-1)) &= \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\
&= e^{-\lambda} \sum_{x=2}^{\infty} \frac{x(x-1)}{x!} \cdot \lambda^x && \text{since } x = 0, 1 \text{ terms are 0.} \\
&= e^{-\lambda} \sum_{x=2}^{\infty} \frac{1}{(x-2)!} \cdot \lambda^x \\
&= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{1}{(x-2)!} \cdot \lambda^{x-2} && \text{pulling out } \lambda^2 \\
&= \lambda^2 e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{(k)!} \cdot \lambda^k && \text{letting } k = x-2 \\
&= \lambda^2 e^{-\lambda} \cdot e^{\lambda} && \text{power series for } e^{\lambda} \\
&= \lambda^2.
\end{aligned}$$

Finally, we find $V(X)$ using our expectation shortcuts:

$$\begin{aligned}
V(X) &= E(X^2) - \mu^2 \\
&= [E(X(X-1)) + E(X)] - \mu^2 \\
&= [\lambda^2 + \lambda] - \lambda^2 \\
&= \lambda.
\end{aligned}$$

□

Example 7.9. Suppose X gives the number of typos per page in a large printed manuscript, and X is Poisson with $\lambda = 2$. Find the probability that a randomly chosen page has (a) fewer than 2 typos, and (b) more than 5 typos.

Part (a) asks for

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= \frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} \\ &= e^{-2} + 2e^{-2} \\ &= 3e^{-2} \\ &\approx 0.406. \end{aligned}$$

Part (b) asks for

$$\begin{aligned} P(X > 5) &= 1 - P(X \leq 5) \\ &= 1 - \left[\frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} + \cdots + \frac{2^5 e^{-2}}{5!}\right] \end{aligned}$$

Rather than calculate this by hand, we turn to R, and the command `ppois(k,lambda)`, which returns $P(X \leq k)$ when X is $\texttt{Poisson}(\lambda)$.

So $P(X > 5) = 1 - P(X \leq 5) =$ 1 - `ppois(5,2)` ≈ 0.017.

Example 7.10. The number of website views at a seldom visited website is Poisson with an average number of 8 visits per day.

> What is the probability that the site gets 20 or more visits in a day?

We want $P(X \geq 20)$, an infinite sum, so we use the strategy of finding the complement, with the help of the function `ppois()` (Appendix C) in R for the calculation:

$$\begin{aligned} P(X \geq 20) &= 1 - P(X \leq 19) \\ &= 1 - \texttt{ppois}(19, 8) \\ &\approx .00025. \end{aligned}$$

7.5.1 Poisson Process

If we're interested in modelling the number of instances of some rare event over a time interval, we can imagine subdividing the interval into n small pieces, small enough that at most 1 instance can occur in each subinterval. In fact, we can imagine each subinterval constitutes a Bernoulli trial of sorts:

- the probability of 1 instance in a subinterval ("success") equals p;

- the probability of 0 instances ("failure") equals $1 - p$.

Then, if X equals the number of instances in the original interval, then X is binomial on n trials with probability of success p on each trial, assuming identical, independent Bernouilli trials. As we increase n, breaking the original time period into smaller and smaller subintervals, the corresponding probability p of seeing 1 instance of the event in a subinterval will decrease, but what if $n \cdot p$ remains constant? In this case, let $\lambda = np$ and consider what happens to the probability density function for the $texttbinom(n,p)$ distribution:

$$\begin{aligned}
\lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} &= \lim_{n\to\infty} \frac{n\cdot(n-1)\cdot\dots\cdot(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{\lambda^x}{x!}\cdot\frac{n\cdot(n-1)\cdot\dots\cdot(n-x+1)}{n^x}\cdot\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\cdot\frac{n}{n}\cdot\frac{n-1}{n}\cdot\frac{n-2}{n}\cdot\dots\cdot\frac{n-x+1}{n}.
\end{aligned}$$

Now, we have a limit of a product of many terms. If the limit of each term exists, then the overall limit will be the product of each of the individual limits.

First, observe that $(1-\lambda/n)^n \to e^{-\lambda}$ as $n \to \infty$ by one of the greatest limits in mathematics:

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$$

for any real number x.

Second, observe that $(1-\lambda/n)^{-x} \to 1^{-x} = 1$ since $\lambda/n \to 0$ as $n \to \infty$.

Finally, for any $k \geq 0$, $\lim_{n\to\infty} \frac{n-k}{n} = 1$ so each of the ratios from the third term on converges to 1.

So, if np remains constant as $n \to \infty$, then as $n \to \infty$ the binomial distribution $texttbinom(n,p)$ approaches the `Poisson`(λ) distribution, where $\lambda = np$. So, the Poisson distribution can be a good model for counting instances of some rare event.

The process described above, where subdivision of the interval leads to Bernoulli trials in such a way that np remains constant, is called a Poisson process.

Definition 7.4. The process by which an event happens is called a **Poisson process** if the following holds:

1. The dimension over which X is measured can be subdivided into n small pieces, within which the event can occur *at most once*.
2. In each small piece, the probability of seeing one occurrence is the same, say p, and p is proportional to the length of the sub-interval (as n grows, p shrinks, but np remains constant).
3. Occurrences in all the small pieces are independent from one another.

In the limit derivation above we demonstrated the following:

For large n, if we let $\lambda = np$, `Poisson`$(\lambda) \sim$ `binom`(n,p).

Table 7.1: Approximating a Poisson distn with Binomial distns

x	binom(10,4)	binom(40,.1)	binom(400,.01)	Poisson(4)
0	0.0060	0.0148	0.0180	0.0183
1	0.0403	0.0657	0.0725	0.0733
2	0.1209	0.1423	0.1462	0.1465
3	0.2150	0.2003	0.1959	0.1954
4	0.2508	0.2059	0.1964	0.1954
5	0.2007	0.1647	0.1571	0.1563
6	0.1115	0.1068	0.1045	0.1042
7	0.0425	0.0576	0.0594	0.0595

For instance, in the table below we compare the probabilities for `binom`$(10, .4)$, `binom`$(40, .1)$, and `binom`$(400, .01)$ with those of a `Poisson`(4) distribution:

Example 7.11. Industrial accidents occur according to a Poisson process with an average of 3 accidents per month. During the last 2 months 10 accidents occurred. Does this number seem highly improbable if the mean number of accidents per month is still equal to 3? Does it indicate a genuine increase in the mean number of accidents per month?

If X equals the number of accidents in *two* months, then X is Poisson with mean $\lambda = 6$.

Then $P(X \geq 10) = 1 - P(X \leq 9) =$ `1-ppois(9,6)` $= 0.084$.

Let's consider this result. If the mean number of accidents per month is still 3, we would expect to observe 10 or more accidents over a two month period about 8 times out of 100, which is unlikely, but not extremely unlikely. Better safe than sorry, I would send out a safety memo and closely monitor what unfolds over the next month!

Now, if we had 5 more accidents the next month, giving us 15 over a 3 month window, chances of 15 or more over 3 months is $P(X \geq 15)$, where X is `Poisson`(9). Using R, this probability is `1-ppois(14,9)` $= 0.041$.

So we have about a 4% chance of seeing 15 or more over a 3 month window if the mean number of accidents per month is 3. I would investigate whether some practice has changed to make accidents more likely than they used to be.

Example 7.12. For a certain type of soil the number of wireworms per cubic foot has a Poisson distribution with mean of 100.

Give an interval that captures at least 5/9ths of the distribution.

OK, this feels like a job for Tchebbysheff.

Here X is Poisson with parameter $\lambda = 100$, so $\mu = 100$ and $\sigma = \sqrt{100} = 10$.

Thinking of Tchebbysheff's inequality, let's find k so that $1 - 1/k^2 = 5/9$. Then

$$P(|X - \mu| < k\sigma) \geq 5/9,$$

meaning at least 5/9ths of the distribution is in the interval $(\mu - k\sigma, \mu + k\sigma)$. Solving $1 - \frac{1}{k^2} = \frac{5}{9}$ for $k > 0$ yields $k = \frac{3}{2}$, so the interval is 85 to 115 wireworms.

Chapter 8

Moments and Moment-Generating Functions

For random variable X we have seen that $E(X)$ and $E(X^2)$ provide useful information:

- $\mu = E(X)$ gives the mean of the distribution
- $\sigma^2 = E(X^2) - E(X)^2$ gives the variance of the distribution.

Definition 8.1. Let X be a random variable, and $k \geq 1$. The **kth moment of X about the origin** is $E(X^k)$. More generally, for any constant $c \in \mathbb{R}$, $E((X-c)^k)$ is called the **kth moment of X about $x = c$**.

Often times we can encode all the moments of a random variable in an object called a moment-generating function.

Definition 8.2. Let X be a discrete random variable with density function $p(x)$. If there is a positive real number h such that for all $t \in (-h, h)$,

$$E(e^{tx})$$

exists and is finite, then the function of t defined by

$$m(t) = E(e^{tx})$$

is called the moment-generating function of X.

Example 8.1. Suppose X has the density function

x	0	1	2	3
$p(x)$	.1	.2	.3	.4

Then, for any real number t,

$$\begin{aligned}
m(t) &= E(e^{tx}) \\
&= \sum_{x=0}^{3} e^{tx} \cdot p(x) \\
&= e^0 \cdot (.1) + e^t \cdot (.2) + e^{2t} \cdot (.3) + e^{3t} \cdot (.4) \\
&= .1 + .2e^t + .3e^{2t} + .4e^{3t},
\end{aligned}$$

and this sum exists as a finite number for any $-\infty < t < \infty$, so the mgf for X exists.

How does $m(t)$ **encode the moments** $E(X), E(X^2), E(X^3), \ldots$?

Theorem 8.1 (Extracting Moments from the Moment-generating Function)**.** *Suppose X is a random variable with moment-generating function $m(t)$ which exists for t in some open interval containing 0. Then the kth moment of X equals the kth derivative of $m(t)$ evaluated at $t = 0$:*

$$E(X^k) = m^{(k)}(0).$$

Proof. Let's say X is discrete and

$$m(t) = \sum_{\text{all } x} e^{tx} \cdot p(x).$$

Then the derivative of $m(t)$ with respect to the variable t is Then

$$m'(t) = \sum_{\text{all } x} x \cdot e^{tx} \cdot p(x),$$

and letting $t = 0$ we have

$$m'(0) = \sum_{\text{all } x} x \cdot e^0 \cdot p(x),$$

which equals $E(X)$ since $e^0 = 1$.

The second derivative of $m(t)$ is

$$\begin{aligned}
m''(t) &= \frac{d}{dt}\left[m'(t)\right] \\
&= \sum_{\text{all } x} x^2 \cdot e^{tx} \cdot p(x)
\end{aligned}$$

Evaluating this at $t = 0$ gives

$$m''(t) = \sum_{\text{all } x} x^2 \cdot 1 \cdot p(x) = E(X^2).$$

Continuing in this manner, for any $k \geq 1$, the kth derivative of $m(t)$ is

$$m^{(k)}(t) = \sum_{\text{all } x} x^k \cdot e^{tx} \cdot p(x),$$

which evaluates to the defintion of $E(X^k)$ when $t = 0$. □

Example 8.2 (The mgf for a geometric distribution)**.** If X is geometric with parameter p, then

$$p(x) = (1-p)^{x-1} \cdot p,$$

for $x = 1, 2, 3, \ldots$, and

$$\begin{aligned}
m(t) &= E(e^{tx}) \\
&= \sum_{x=1}^{\infty} e^{tx}(1-p)^{x-1} \cdot p \\
&= pe^t \sum_{x=1}^{\infty} e^{t(x-1)}(1-p)^{x-1} \quad \text{since } e^t \cdot e^{t(x-1)} = e^{tx} \\
&= pe^t \sum_{x=1}^{\infty} [e^t(1-p)]^{x-1} \qquad = pe^t \sum_{k=0}^{\infty} [e^t(1-p)]^k \text{where } k = x-1 \text{ is a change of index} \\
&= pe^t \frac{1}{1-e^t(1-p)}
\end{aligned}$$

The last step is true by the geometric series formula, provided $|e^t(1-p)| < 1$. Since $0 \le |e^t(1-p)| = e^t(1-p)$, the series converges by the geometric series formula if and only if $e^t(1-p) < 1$. Well,

$$\begin{aligned}
e^t(1-p) < 1 &\iff e^t < \frac{1}{1-p} \\
&\iff t < \ln\left(\frac{1}{1-p}\right).
\end{aligned}$$

In other words, yes, there exists an interval containing 0 for which $m(t)$ exists for all t in the interval.

Example 8.3 (The mgf for a Poisson distribution)**.** Find the mgf of a Poisson random variable X with parameter λ. Since we're considering a Poisson distribution, our strategy for finding the mgf will be to work our expectation to look like a power series for e^{junk}.

Strategy: Work our series to include

$$\sum_{x=0}^{\infty} \frac{(\text{junk})^x}{x!}$$

since this converges to e^{junk}.

$$\begin{aligned}
m(t) &= E(e^{tx}) \\
&= \sum_{x=0}^{\infty} e^{tx}\frac{\lambda^x e^{-\lambda}}{x!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!} && \text{here it is!} \\
&= e^{-\lambda}e^{[\lambda e^t]} && \text{for all } -\infty < t < \infty \\
&= e^{\lambda(e^t-1)}.
\end{aligned}$$

Let's derive our μ and σ formulas for a Poisson random variable using the mgf.

The first derivative is

$$m'(t) = e^{\lambda(e^t-1)} \cdot \lambda e^t,$$

and $m'(0) = e^{\lambda(1-1)} \cdot \lambda e^0 = \lambda$.

The second derivative is

$$m''(t) = (e^{\lambda(e^t-1)} \cdot \lambda e^t) \cdot \lambda e^t + e^{\lambda(e^t-1)} \cdot \lambda e^t,$$

so

$$m''(0) = \lambda^2 + \lambda.$$

Now

$$\mu = m'(0) = \lambda,$$

check! And,

$$\sigma^2 = m''(0) - [m'(0)]^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda,$$

check again!

Chapter 9

Continuous Random Variables

We now turn our attention to continuous random variables.

9.1 Distribution Functions

Definition 9.1 (Distribution Function of a Random Variable)**.** Let X be a random variable. **The distribution function** of X, denoted $F(x)$, is the function defined on all real numbers x such that

$$F(x) = P(X \leq x).$$

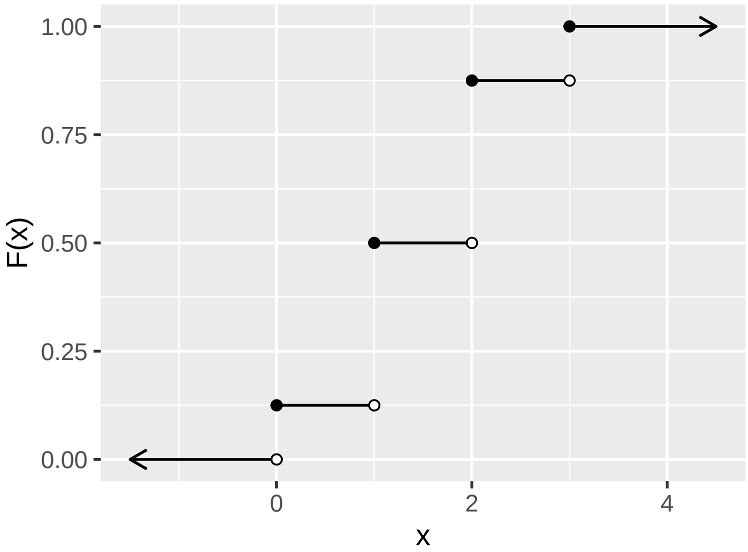
Example 9.1. Suppose X is the discrete random variable given by density function

x	0	1	2	3
$p(x)$	1/8	3/8	3/8	1/8

Note $F(-2.7) = P(X \leq -2.7) = 0$, and $F(1.3) = P(X \leq 1.3) = 1/8 + 3/8 = 4/8$ (since the only X values less than or equal to 1.3 with positive probability are $X = 0$ or $X = 1$).

The distribution function for X is the following step function:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/8 & \text{if } 0 \leq x < 1 \\ 4/8 & \text{if } 1 \leq x < 2 \\ 7/8 & \text{if } 2 \leq x < 3 \\ 1 & \text{if } x \geq 3 \end{cases}$$



Observe that this function $F(x)$ is defined for all $-\infty < x < \infty$ (check out those arrows :)). The jumps in the graph indicate that the function F is not continuous, and the points of discontinuity occur exactly at the values of X in the probability table for X.

Theorem 9.1 (Properties of any distribution function)**.** *If $F(x)$ is a distribution function, then*

a) $\lim_{x\to-\infty} F(x) = 0$*;*
b) *F is non-decreasing. That is, if $x_1 \leq x_2$ then $F(x_1) \leq F(x_2)$; and*
c) $\lim_{x\to\infty} F(x) = 1$.

Definition 9.2 (Continuous Random Variable)**.** A random variable is called a **continuous random variable** if its distribution function F is continuous for all x.

So the distribution function for any continuous random variable has the following sort of look, descriptively (as in Figure 9.1):

- it is continuous,
- its domain is $(-\infty, \infty)$
- as x progresses away from $-\infty$ toward ∞, the values of $F(x)$ rise from 0 to 1, never decreasing along the way.

Definition 9.3. Let F be the distribution for a continuous random variable X. Then the derivative of F, wherever it exists is called the **probability density function** for X. When continuous X has a probability density function, we usually denote it as $f(x)$.

The density function $f(x)$ is a theoretical curve for the frequency distribution of a population of measurements. We'll look at examples shortly.

Theorem 9.2 (Properties of a density function)**.** *If $f(x)$ is a density function for a continuous random variable X, then*

a) $f(x) \geq 0$ *for all x, and*

b) $\displaystyle\int_{-\infty}^{\infty} f(x)\ dx = 1.$

Sketch of Proof:

For a) Recall $f(x) = F'(x)$. One feature of any distribution function is that it is never decreasing, so its slope (derivative) is never negative. Since $f(x)$ gives the slope of F, $f(x) \geq 0$.

For b) F is the antiderivative of f, which is useful to know when we integrate f. Also, $\displaystyle\int_{-\infty}^{\infty} f(x)\ dx$ is an improper integral, which we can tackle by splitting it into two integrals, assuming each of these new integrals converges:

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\ dx &= \int_{-\infty}^{0} f(x)\ dx + \int_{0}^{\infty} f(x)\ dx \\
&= \lim_{a\to-\infty} \int_{a}^{0} f(x)\ dx + \lim_{b\to\infty} \int_{0}^{b} f(x)\ dx \\
&= \lim_{a\to-\infty} [F(0) - F(a)] + \lim_{b\to\infty} [F(b) - F(0)] \\
&= (F(0) - 0) + (1 - F(0)) && \text{by limit properties of } F \\
&= 1.
\end{aligned}$$

Example 9.2. Consider distribution function F pictured below, where $c > 0$ is a fixed constant.

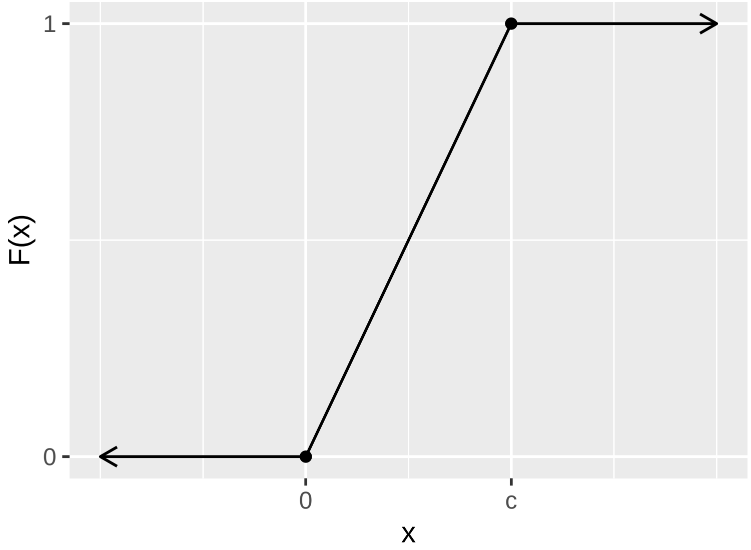


Figure 9.1: Piece-wise linear distribution function

This function is piece-wise linear, continuous, and it is differentiable everywhere except the sharp corners at $x = 0$ and $x = c$. At any other point, $f(x) = F'(x)$ equals the slope of the segment running through the point $(x, F(X))$.

So the probability density function for this random variable is

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/c & \text{if } 0 < x < c \\ 0 & \text{if } x > c, \end{cases}$$

and the graph of f looks like this:

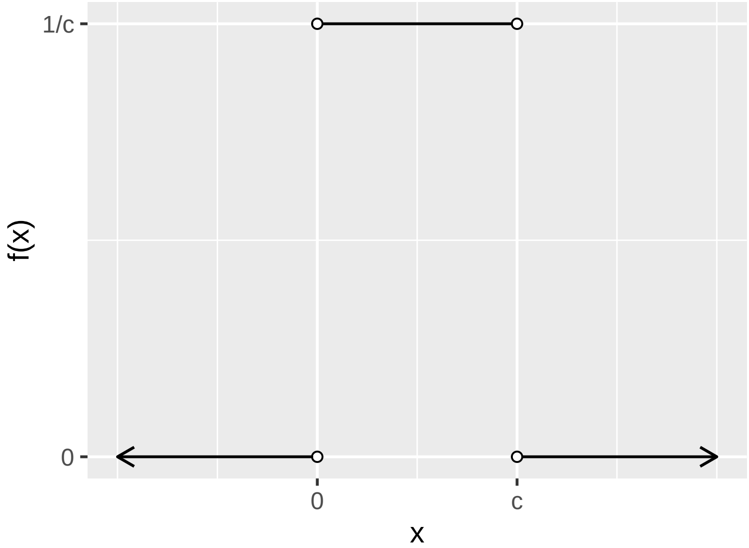


Figure 9.2: probability density function for X

Note that $f(x) \geq 0$ for all x. Also,

$$\int_{-\infty}^{\infty} f(x)\ dx = \int_{0}^{c} f(x)\ dx$$

(we only have to integrate over intervals in which $f(x) > 0$), and this later integral is the area of a rectangle of width c and height $1/c$, so it has area 1. Thus, we have a valid pdf!

Example 9.3. Find the value of k that makes the following function a valid pdf.

$$f(x) = \begin{cases} kx^8 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{else.} \end{cases}$$

We need $k \geq 0$ os that $f(x) \geq 0$ for all x. We also need

$$1 = \int_{-\infty}^{\infty} f(x)\ dx = \int_{0}^{1} kx^8\ dx = \frac{k}{9}x^9 \Big|_0^1.$$

It follows that $k = 9$.

Definition 9.4 (Quantiles)**.** Let X denote a random variable. If $0 < p < 1$, the pth quantile of X, denoted ϕ_p, is the smallest value such that $F(\phi_p) \geq p$. If X is continuous, ϕ_p is the smallest value such that $F(\phi_p) = p$.

Some special quantiles:

- $\phi_{.}25$, denoted $Q1$, is called the first quartile,
- $\phi_{.}5$, denoted M, is called the median of the random variable,
- $\phi_{.}75$, denoted $Q3$, is called the third quartile

Theorem 9.3. *If X is a continuous random variable with density function f, then for any real numbers $a < b$,*

$$P(a \leq X \leq b) = \int_a^b f(x)\ dx.$$

Proof Idea: The distribution function F is an antiderivative of the density function f, so using the Fundamental Theorem of Calculus,

$$\begin{aligned}
\int_a^b f(x)\ dx &= F(b) - F(a) \\
&= P(X \leq b) - P(X \leq a) \\
&= P(a < X \leq b) && \text{since } a < b \\
&= P(a \leq X \leq b) && \text{since } P(X = a) = 0
\end{aligned}$$

Note: For any continuous random variable X, and $a < b$,

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq B)$$

since $P(X = c) = 0$ for any real number c.

Example 9.4 (A quadratic density function)**.** Suppose X has density function

$$f(x) = \begin{cases} \frac{3}{8}x^2 & \text{if } 0 \leq x \leq 2 \\ 0 & \text{else.} \end{cases}$$

Wait! Is this *actually* a valid density function?

1. Ok, yes, $f(x) \geq 0$ for all x.
2. And…

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\ dx &= \int_0^2 \frac{3}{8}x^2\ dx \\
&= \frac{1}{8}x^3\Big|_0^2 \\
&= 1.
\end{aligned}$$

Ok, now to the question: Find $P(1 \leq X \leq 2)$:

$$
\begin{aligned}
P(1 \leq X \leq 2) &= \int_1^2 \frac{3}{8}x^2 \, dx \\
&= \frac{1}{8}x^3 \Big|_1^2 \\
&= 1 - \frac{1}{8} \\
&= \frac{7}{8}.
\end{aligned}
$$

Even though X can take any value between 0 and 2, the probability is 7/8 that X will be between 1 and 2. Most of the area under the density curve is at the high end of the X range:

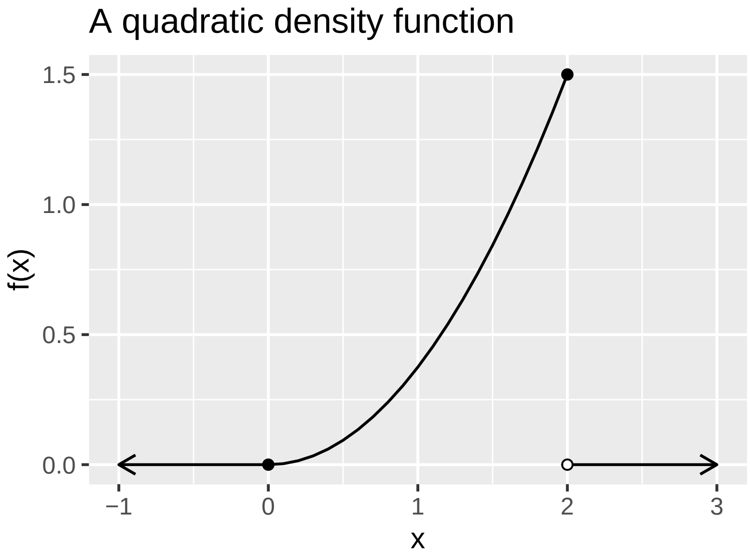


Figure 9.3: A quadratic density function

Example 9.5. Suppose X has density function

$$
f(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{x^2} & \text{if } x \geq 1. \end{cases}
$$

a) Check that this gives a valid density function:

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\ dx &= \int_{1}^{\infty} x^{-2}\ dx\\ &= \lim_{b\to\infty}\left[\int_{1}^{b} x^{-2}\ dx\right]\\ &= \lim_{b\to\infty}\left[-\frac{1}{x}\bigg|_{1}^{b}\right]\\ &= \lim_{b\to\infty}\left[-\frac{1}{b}+1\right]\\ &= 1.\end{aligned}$$

The limit equals 1 in the end since $1/b \to 0$ as $b \to \infty$.

b) Find $F(x)$, the cumulative probability distribution function.

By definition, for any real number x,

$$F(x) = \int_{-\infty}^{x} f(t)\ dt,$$

which, of course, gives the area under f over the interval $(-\infty, x]$. Since f is piece-wise defined, the integrand used in the integral to evaluate F depends on the bounds of the integral.

$$F(x) = \begin{cases} \displaystyle\int_{-\infty}^{x} 0\ dt & \text{if } x < 1\\ \displaystyle\int_{-\infty}^{1} 0\ dt + \int_{1}^{x} \frac{1}{t^2}\ dt & \text{if } x \geq 1.\end{cases}$$

We leave it to the reader to integrate these expressions to obtain

$$F(x) = \begin{cases} 0 & \text{if } x < 1\\ 1 - \dfrac{1}{x} & \text{if } x \geq 1.\end{cases}$$

c) Find $P(1 < X < 3)$.

Well,

$$P(1 < X < 3) = \int_{1}^{3} f(x)\ dx = F(3) - F(1),$$

by the Fundamental Theorem of Calculus (FTC), so

$$P(1 < X < 3) = (1 - 1/3) - (1 - 1/1) = 2/3.$$

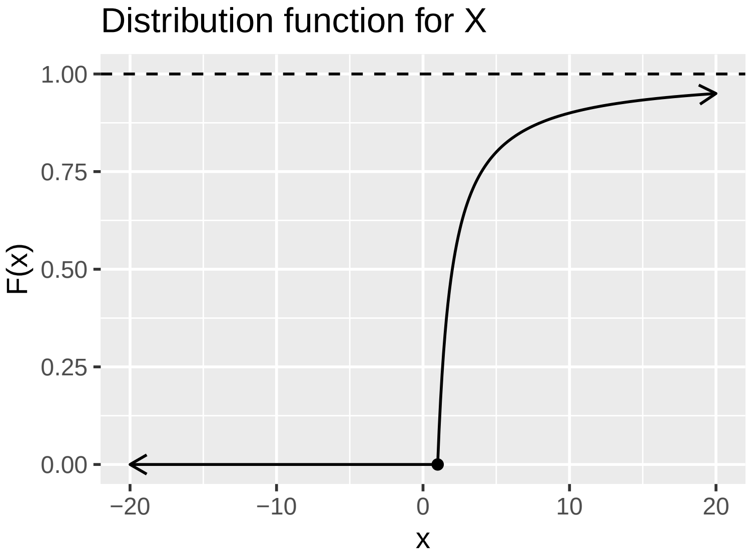


Figure 9.4: Distribution function for X

9.2 Expected Value for Continuous Random Variables

Definition 9.5. If X is a continuous random variable with probability density function $f(x)$, then the **expected value of** X, denoted $E(X)$, is

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

provided this integral exists. The expected value $E(X)$ is also called the **mean of** X, and is often denoted as μ_X, or μ if the random variable X is understood.

The **expected value of the function** $g(X)$ **of** X is

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f(x)\ dx,$$

provided this integral exists.

The **variance of** X is

$$V(X) = E((X - \mu_X)^2),$$

provided this integral exists.

As in the discrete case, one can show $V(X) = E(X^2) - E(X)^2$, a working formula for variance which is sometimes easier to use to calculate variance.

Example 9.6.

Find $E(X)$ and $V(X)$ where X is the continuous random variable from Example 9.4.

Recall X has density function $f(x) = 3x^2/8$ for $0 \le x \le 2$.

Expected Value:

$$\begin{aligned} E(X) &= \int_0^2 x \cdot 3x^2/8 \; dx \\ &= \frac{3}{8}\int_0^2 x^3 \; dx \\ &= \frac{3}{8}\frac{1}{4}x^4 \; \Big|_0^2 \\ &= \frac{3}{2}. \end{aligned}$$

Variance: We first find $E(X^2)$:

$$\begin{aligned} E(X^2) &= \int_0^2 x^2 \cdot 3x^2/8 \; dx \\ &= \frac{3}{8}\int_0^2 x^4 \; dx \\ &= \frac{3}{8}\frac{1}{5}x^5 \; \Big|_0^2 \\ &= \frac{12}{5}. \end{aligned}$$

Then,

$$\begin{aligned} V(X) &= E(X^2) - E(X)^2 \\ &= (12/5) - (3/2)^2 \\ &= 0.15. \end{aligned}$$

The properties of expected value that held for discrete random variables also hold for continuous random variables.

Theorem 9.4. *Suppose X is a continuous random variable, $c \in \mathbb{R}$ is a constant, and g, g_1, and g_2 are functions of X.*

1. *$E(c) = c$.*
2. *$E(c \cdot g(X)) = cE(g(X))$.*
3. *$E(g_1(X) \pm g_2(X)) = E(g_1(X) \pm g_2(X))$.*

These results follow immediately from properties of integration. For instance, to prove property 1 we observe that for constant c,

$$E(c) = \int_{-\infty}^{\infty} c \cdot f(x) \; dx = c\int_{-\infty}^{\infty} f(x) \; dx,$$

and the integral in the last expression equals 1 by definition of a valid probability density function.

Theorem 9.5. *Let X be a random variable (discrete or continuous) with $E(X) = \mu$ and $V(X) = \sigma^2$, and let a, b be constants. Then*

a) $E(aX + b) = aE(X) + b = a\mu + b$.
b) $V(aX + b) = a^2V(X) = a^2\sigma^2$.

Proof.

a) This result follows immediately from properties of expected value (Theorems 9.4 and 9.4).

b) Let $Y = aX + b$. Then (a) says that $E(Y) = a\mu + b$, so

$$\begin{aligned} V(Y) &= E((Y - (a\mu + b))^2) \\ &= E\left(((aX + b) - (a\mu + b))^2\right) \\ &= E\left((aX - a\mu)^2\right) \\ &= a^2E\left((X - \mu)^2\right) \end{aligned}$$

But $E\left((X - \mu)^2\right) = V(X)$ by the definition of variance, so the result follows.

□

Example 9.7 (Ore Sample Impurities)**.** For certain ore samples, the proportion X of impurities per sample is a random variable with density function

$$f(x) = \begin{cases} 1.5x^2 + x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{else.} \end{cases}$$

The dollar value of each sample is $W = 5 - 0.5X$.

Find the mean, variance, and standard deviation of W.

First, let's consider the variable X itself.

$$\begin{aligned} E(X) &= \int_0^1 x \cdot (1.5x^2 + x)\ dx \\ &= \int_0^1 1.5x^3 + x^2\ dx \\ &= \frac{1.5}{4}x^4 + \frac{1}{3}x^3 \Bigg|_0^1 \\ &= \frac{17}{24}. \end{aligned}$$

Also,

$$\begin{aligned} E(X^2) &= \int_0^1 x^2 \cdot (1.5x^2 + x)\ dx \\ &= \int_0^1 1.5x^4 + x^3\ dx \\ &= \frac{1.5}{5}x^5 + \frac{1}{4}x^4 \Big|_0^1 \\ &= \frac{11}{20}. \end{aligned}$$

Thus, $V(X) = (11/20) - (17/24)^2 \approx 0.0483$.

Then, by Theorem 9.5,

$$E(W) = E(5 - 0.5X) = 5 - 0.5E(X) = 5 - 0.5 \cdot (17/24) = 4.65,$$

and

$$V(W) = V(5 - 0.5X) = 0.25V(X) \approx 0.012,$$

so that the standard deviation is $\sigma = \sqrt{V(W)} \approx 0.11$ (about 11 cents).

Chapter 10

Important Continuous Random Variables

In this chapter we introduce the following well-known continuous random variables: uniform, normal, exponential, gamma, chi-square, and beta. In examples we work through, it will from time to time be convenient to compute probabilities in R. Appendix D contains details about the commands in R useful for doing so.

10.1 Uniform Distribution

Definition 10.1. Let $\theta_1 < \theta_2$ be distinct real numbers. A random variable X has **uniform distribution on the interval** $[\theta_1, \theta_2]$ if it has probability density function

$$f(x) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \text{if } \theta_1 \leq x \leq \theta_2 \\ 0 & \text{else.} \end{cases}$$

we may write $X \sim U(\theta_1, \theta_2)$ to mean X is uniform on $[\theta_1, \theta_2]$.

A uniform random variable is a good model for picking a random real number between θ_1 and θ_2.

In R we access the uniform distribution with `unif`. For instance, we can generate a random sample of n numbers in the interval $[a, b]$ with the `runif()` command:

```
n=6; a = 0; b = 10;
runif(n,a,b)
```

```
## [1] 8.8125533 3.9767428 0.8136621 7.6948940 3.5300695 3.6943150
```

Example 10.1 (Average value of a function).

Use R to estimate the average value of $f(x) = x^2$ over the interval [0,2].

Our strategy: Select a large random sample of points in the interval [0,2] and then compute the average of their squares.

```
x = runif(1000,0,2) #picking 1000 points in [0,2]
mean(x^2)
```

```
## [1] 1.334432
```

Note: From Calc I, we know the average value of a function f over the interval $[a, b]$ is

$$\frac{1}{b-a}\int_a^b f(x)\ dx,$$

so here it's

$$\frac{1}{2}\int_0^2 x^2\ dx = \frac{1}{6}x^3\ \Big|_0^2 = 4/3 \approx 1.333.$$

Theorem 10.1. *If X is $U(\theta_1, \theta_2)$, then*

$$E(X) = \frac{\theta_1+\theta_2}{2}, \qquad \textit{and} \qquad V(X) = \frac{(\theta_2-\theta_1)^2}{12}.$$

Proof. Recall, X has pdf

$$f(x) = \begin{cases} \frac{1}{\theta_2-\theta_1} & \text{if } \theta_1 \leq x \leq \theta_2 \\ 0 & \text{else.} \end{cases}$$

So

$$\begin{aligned} E(X) &= \int_{\theta_1}^{\theta_2} x \cdot \frac{1}{\theta_2-\theta_1}\ dx \\ &= \frac{1}{\theta_2-\theta_1} \cdot \frac{1}{2}x^2\ \Big|_{\theta_1}^{\theta_2} \\ &= \frac{1}{\theta_2-\theta_1} \cdot \frac{1}{2}(\theta_2^2-\theta_1^2) \\ &= \frac{1}{\theta_2-\theta_1} \cdot \frac{1}{2}(\theta_2-\theta_1)(\theta_2+\theta_1) \\ &= \frac{\theta_1+\theta_2}{2}. \end{aligned}$$

One can show similarly, that

$$E(X^2) = \int_{\theta_1}^{\theta_2} x^2 \cdot \frac{1}{\theta_2-\theta_1}\ dx = \cdots = \frac{\theta_2^2+\theta_1\theta_2+\theta_1^2}{3},$$

so that

$$V(X) = E(X^2) - E(X)^2 = \frac{(\theta_2-\theta_1)^2}{12}.$$

The fun algebra details are left to the reader. □

10.2 Exponential Distribution

The exponential distribution is often used to model experiments that aim to investigate: How long until something happens?

Definition 10.2. A random variable X has an **exponential probability distribution with parameter** β, denoted $\texttt{Exp}(\beta)$, if it has probability density function

$$f(x) = \frac{1}{\beta} e^{-(x/\beta)} \quad \text{for } x \geq 0 \quad (\text{and } f(x) = 0 \text{ else.})$$

First, let's check that the total area under $f(x)$ is 1.

$$\begin{aligned}
\int_0^\infty \frac{1}{\beta} e^{-x/\beta}\, dx &= \lim_{b\to\infty} \left[\int_0^b \frac{1}{\beta} e^{-x/\beta}\, dx \right] \\
&= \lim_{b\to\infty} \left[-e^{-x/\beta} \Big|_0^b \right] && \text{(try u-sub.} u = -x/\beta \\
&= \lim_{b\to\infty} \left[1 - \frac{1}{e^{b/\beta}} \right] \\
&= 1.
\end{aligned}$$

Having done the above integral, we can write down a formula for the cumulative distribution function for an exponential distribution:

If X is Exp(β) then for $x \geq 0$,

$$\begin{aligned}
F(x) &= \int_0^x \frac{1}{\beta} e^{-t/beta}\, dt \\
&= 1 - e^{-x/\beta}.
\end{aligned}$$

Example 10.2. Suppose X is $\texttt{Exp}(4)$. Find $P(X < 8)$.

Well,

$$\begin{aligned}
P(X < 8) &= F(8) \\
&= 1 - e^{-8/4} \\
&= 1 - e^{-2} \\
&\approx .865.
\end{aligned}$$

Theorem 10.2. *If X is $\texttt{Exp}(\beta)$), then*

$$E(X) = \beta, \quad \textit{and} \quad V(X) = \beta^2.$$

Proof.

$$\begin{aligned}
E(X) &= \int_0^\infty x \cdot \frac{1}{\beta} e^{-x/\beta}\, dx \\
&= \lim_{b\to\infty} \left[\int_0^b x \cdot \frac{1}{\beta} e^{-x/\beta}\, dx \right]
\end{aligned}$$

To evaluate this integral, try integration by parts with $u = x$ and $dv = e^{-x/\beta}\,dx$. Doing so, we obtain

$$\begin{aligned} E(X) &= \lim_{b\to\infty}\left[\int_0^b x\cdot\frac{1}{\beta}e^{-x/\beta}\,dx\right] \\ &= \lim_{b\to\infty}\left[-xe^{-x/\beta}-\beta e^{-x/\beta}\Big|_0^b\right] \\ &= \lim_{b\to\infty}\left[\left(\frac{-b}{e^{b/\beta}}-\frac{\beta}{e^{b/\beta}}\right)-(0-\beta)\right]. \end{aligned}$$

Since $\frac{b}{e^{b/\beta}} \to 0$ and $\frac{\beta}{e^{b/\beta}} \to 0$ as $b \to \infty$, we have proved that $E(X) = \beta$.

To prove that $V(X) = \beta^2$, first find $E(X^2) = \int_0^\infty x^2 f(x)\,dx$ by integration by parts, and then use the fact that $V(X) = E(X^2) - E(X)^2$. We leave details to those nostalgic for Calc II. :) □

10.3 Normal Distribution

Definition 10.3. A random variable X has a **normal probability distribution with parameters μ and $\sigma > 0$**, denoted $N(\mu, \sigma)$, if it has probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } -\infty < x < \infty.$$

The graph of a normal density curve is bell-shaped, with peak at $x = \mu$, and inflection points at $x = \mu \pm \sigma$, facts we can readily demonstrate by analyzing the first and second derivative of f.

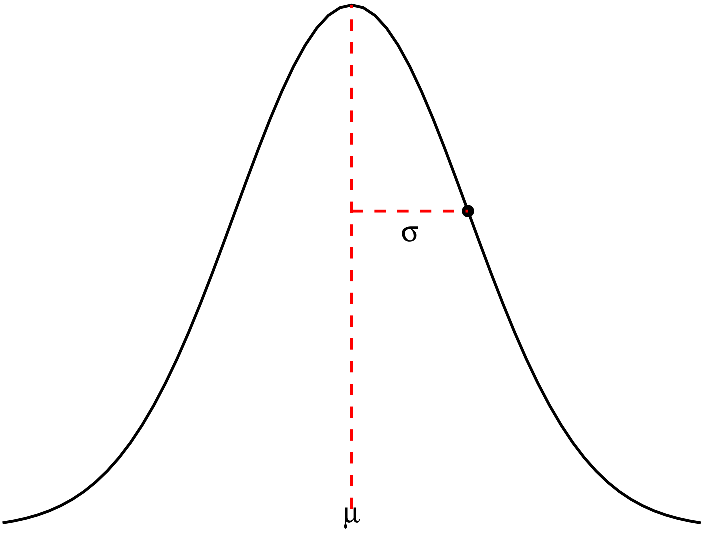


Figure 10.1: A Normal density curve

Theorem 10.3. *If X is $N(\mu, \sigma)$, then*

$$E(X) = \mu, \quad \textit{and} \quad V(X) = \sigma^2.$$

Definition 10.4. The **standard normal probability distribution** is $N(0,1)$. If Z is $N(0,1)$, its pdf is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \text{ for all real numbers } z.$$

As we shall see, the family of normal distributions $N(\mu, \sigma)$ has a special place of importance in statistics; many distributions have a bell-shape (physical measurements, for instance, such as heights of adult males, weights of newborns, wingspans of adult female bald eagles, …). But its special place of importance in statistics comes from the fact that the distribution of sample means from repeated sampling, as we shall see, are well-modeled by normal distributions.

Theorem 10.4. *If X is $N(\mu, \sigma)$ then $Z = (X - \mu)/\sigma$ is $N(0,1)$.*

We prove this theorem later.

In practice, shifting from X to

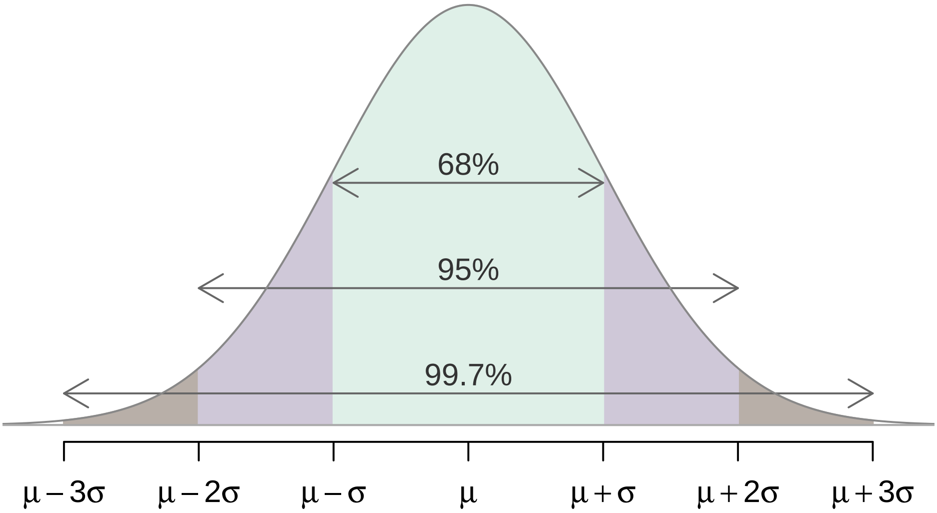
$$Z = \frac{X - \mu}{\sigma}$$

gives us a way to consider unitless, standardized “Z-scores” associated to values in X.

A Z-score for X gives the number of standard deviations above or below the mean X is in its distribution.

Example 10.3 (The 68-95-99.7 Rule)**.** In any normal distribution $N(\mu, \sigma)$:

- About 68% of the distribution is within 1 standard deviation of the mean.
- About 95% of the distribution is within 2 standard deviations of the mean.
- About 99.7% of the distribution is within 2 standard deviations of the mean.



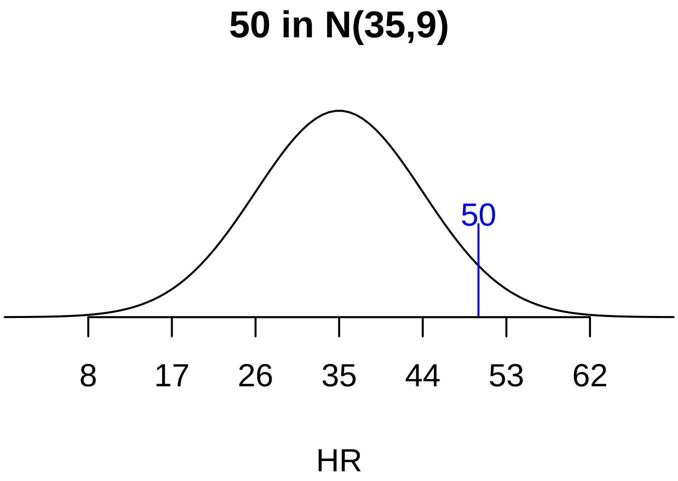
For instance, in $N(10,3)$,

- Roughly, 68% of the distribution is between 7 and 13, and
- 95% of the distribution is between 4 and 16, and
- 99.7% of the distribution is between 1 and 19.

Example 10.4. Which is more "impressive": hitting 50 home runs in a season when the league home run distribution is $N(35, 9)$, or hitting 35 home runs in a season when the league distribution is $N(24, 5)$?

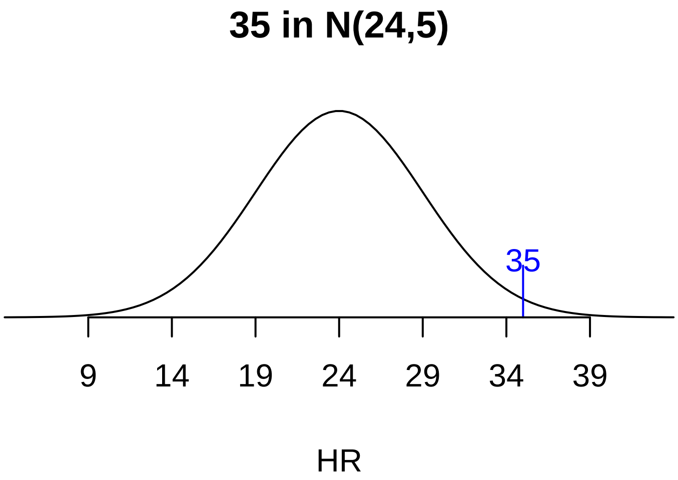
For 50 in $N(35, 9)$,

$$Z = \frac{50 - 35}{9} = \frac{5}{3} \approx 1.67.$$



For 35 in $N(24, 5)$,

$$Z = \frac{35 - 24}{5} = \frac{11}{5} = 2.2.$$



A person hitting 35 HR in a league with distribution $N(24, 5)$ is more extreme (at the high end), and so more impressive in that sense.

Now we focus on some fine print, proving that the density function for a normal distribution is, indeed, a valid density function.

Lemma 10.1.

$$\int_{-\infty}^{\infty} e^{-x^2/2}\ dx = \sqrt{2\pi}.$$

Proof. First, we remark that the integral converges by comparison with $\displaystyle\int_{-\infty}^{\infty} e^{-x/2}\ dx.$

Suppose the value of the integral we want to calculate is A. We use some integration techniques from vector calculus to first find the value of A^2. If you haven't seen vector calculus, don't sweat the details, but demand your vector calculus prof prove this lemma when you take the class :). Ok, let's look at A^2.

$$\begin{aligned}
A^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\ dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\ dy\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2/2}e^{-y^2/2}dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-(x^2+y^2)}{2}}\ dxdy \\
&= \int_{0}^{2\pi}\int_{0}^{\infty} e^{-r^2/2}\ r\ dr\ d\theta && \text{change to polar coordinates} \\
&= 2\pi\int_{0}^{\infty} e^{-r^2/2}\ r\ dr \\
&= -\pi\int_{0}^{\infty} e^{-u}\ du && \text{let } u = r^2/2 \\
&= 2\pi\left[-e^{-u}\Big|_0^{\infty}\right] && = 2\pi[-0+1] \\
&= 2\pi
\end{aligned}$$

Since $A^2 = 2\pi$, $A = \sqrt{2\pi}$. □

We have the following corollaries to this lemma.

Corollary 10.1. *The function* $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, *for* $-\infty < x < \infty$, *is a valid probability density function.*

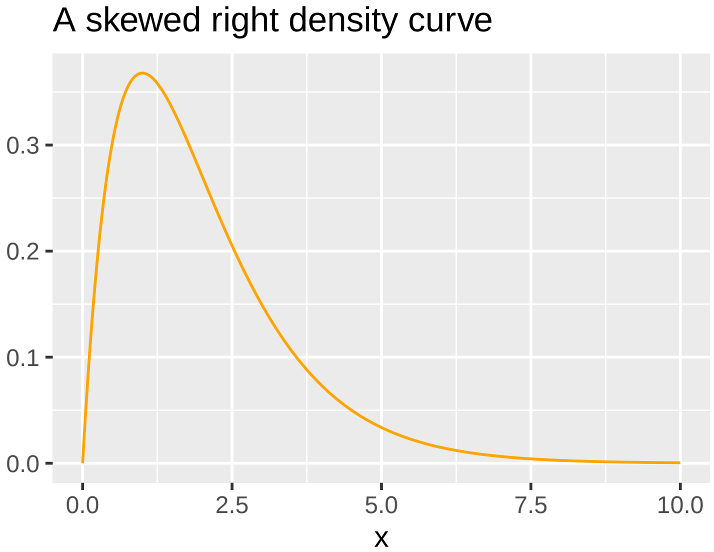
Proof. Clearly, $f(x) \geq 0$ for all x, and the previous lemma ensures that $\int_{-\infty}^{\infty} f(x)\ dx = 1.$. □

Corollary 10.2. *The function* $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$, *for* $-\infty < x < \infty$, *is a valid probability density function.*

Proof. Clearly, $f(x) \geq 0$ for all x, and after u-substitution of $u = (x-\mu)/\sigma$, the previous lemma ensures that $\int_{-\infty}^{\infty} f(x)\ dx = 1$. □

10.4 Gamma Distribution

Some random variables are always nonnegative and yield distributions of data that are skewed right, as pictured below.



Some typically skewed right distributions include household incomes in a city, the length of time between malfunctions of some machine, and major league baseball salaries. The gamma probability distribution, which has two parameters α and β, can model such skewed right distributions. The parameter α is sometimes called the **shape** parameter, β is called the **scale** parameter, and its reciprocal $1/\beta$ is called the **rate**.

The density function for a gamma distribution looks formidable, so we'll take time to go through it carefully.

Definition 10.5. A random variable X has a **gamma probability distribution with parameters** $\alpha > 0$ **and** $\beta > 0$, denoted `gamma`(α, β), if and only if it has probability density function

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-(x/\beta)} \quad \text{for } x \geq 0 \quad (\text{and } f(x) = 0 \text{ else})$$

where

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$$

is called the **gamma function**.

Here are plots of three different gamma distributions.

The quantity $\Gamma(\alpha)$ is called the **gamma function**, which has some nice features.

Lemma 10.2. *If* $n > 0$ *then* $\Gamma(n+1) = n \cdot \Gamma(n)$.

Proof. This follows by integration by parts! First note,

$$\Gamma(n+1) = \int_0^\infty t^{(n+1)-1} e^{-t}\, dt = \int_0^\infty t^n e^{-t}\, dt.$$

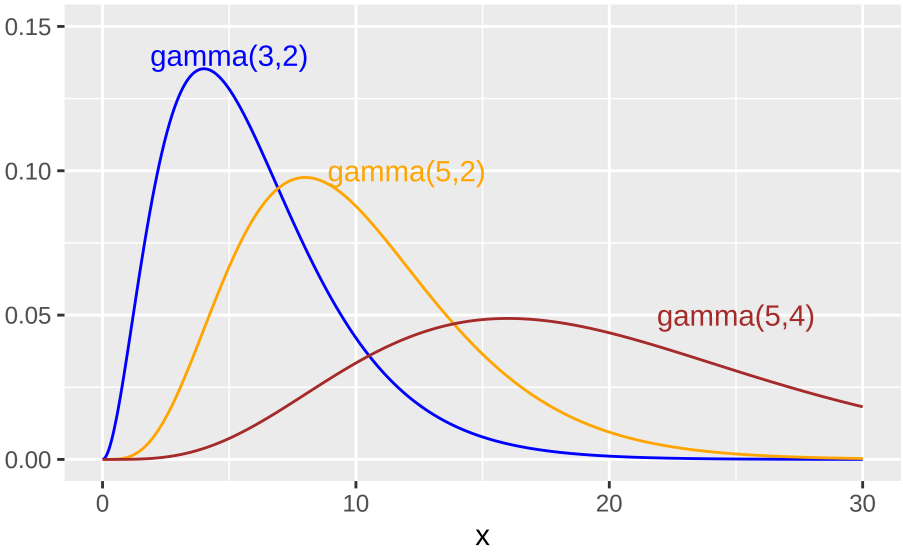


Figure 10.2: Some gamma density functions

Let $u = t^n$, and $dv = e^{-t}\ dt$. Then $du = nt^{n-1}\ dt$ and $v = -e^{-t}$, and

$$\begin{aligned}\int_0^\infty t^n e^{-t}\ dt &= -t^n e^{-t}\Big|_0^\infty - \int_0^\infty nt^{n-1}(-e^{-t})\ dt \\ &= \lim_{b\to\infty}\left[-t^n e^{-t}\Big|_0^b\right] + n\int_0^\infty t^{n-1}e^{-t}\ dt\end{aligned}$$

Apply l'hopital's rule to see that the limit term above evaluates to 0, and note the integral term above is precisely the definition of $\Gamma(n)$. Thus, we have

$$\Gamma(n+1) = n\cdot\Gamma(n).$$

□

This lemma provides us with the following

Fun Fact: $\Gamma(n) = (n-1)!$ for any positive integer n.

To see why this is the case, we first show $\Gamma(1) = 1$:

$$\begin{aligned}
\Gamma(1) &= \int_0^\infty t^0 e^{-t}\ dt \\
&= \lim_{b\to\infty}\left[\int_0^b e^{-t}\ dt\right] \\
&= \overset{\infty}{\lim_{b\to\infty}}\left[-e^{-t}\Big|_0^b\right] \\
&= \overset{\infty}{\lim_{b\to\infty}}\left[-e^{-b}+1\right] \\
&= 1.
\end{aligned}$$

Next, the lemma gives us a recursive way to find $\Gamma(2), \Gamma(3), \Gamma(4)$ and so on. Or, using mathematical induction, $\Gamma(1) = 1$ is our basis step, and the inductive step is proved as follows: Suppose $\Gamma(k) = (k-1)!$ for some $k \geq 1$. Then

$$\begin{aligned}
\Gamma(k+1) &= k \cdot \Gamma(k) && \text{by the lemma} \\
&= k \cdot (k-1)! && \text{by substitution} \\
&= k!
\end{aligned}$$

It follows that $\Gamma(n) = (n-1)!$ for all positive integers n.

The family of gamma distributions contain two special sub-families, one of which we've already seen!

Theorem 10.5. *If X is $\mathtt{gamma}(\alpha, \beta))$, then*

$$E(X) = \alpha\beta, \qquad \textit{and} \qquad V(X) = \alpha\beta^2.$$

We prove this later in Chapter 11.

Example 10.5. Where does the peak of the $\mathtt{gamma}(\alpha, \beta)$ pdf occur?

This looks like a question for calculus. We can find f', set it to 0, and consider critical points.

We leave the details to the reader for now, but find the following results:

- if $\alpha \leq 1$, $f'(x) < 0$ for all $x > 0$, so f is always decreasing and the peak occurs when $x = 0$.
- If $\alpha > 1$, the pdf for $X \sim \mathtt{gamma}(\alpha, \beta)$ has its peak at $x = (\alpha - 1)\beta$.

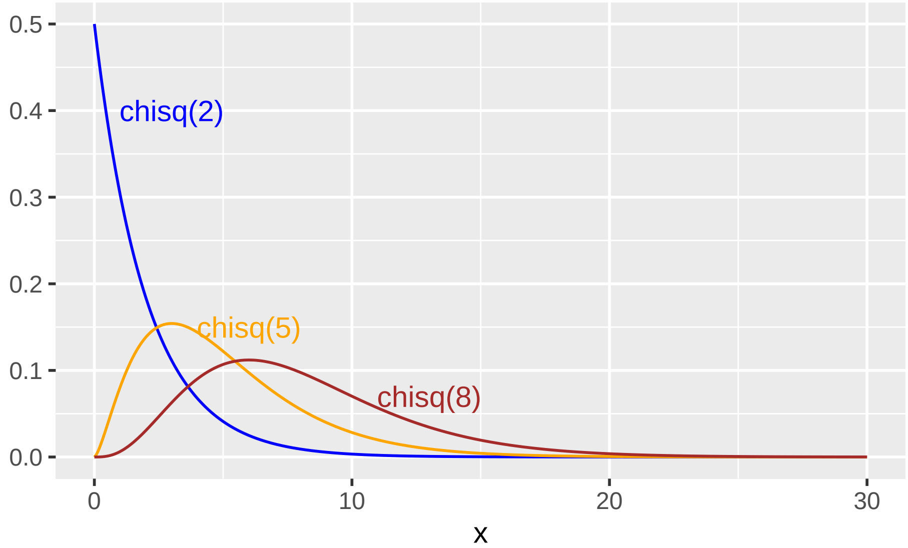
10.4.1 Exponential Distribution

Set $\alpha = 1$ and you will find $\mathtt{gamma}(1, \beta) = \mathtt{Exp}(\beta)$, because their density functions are identical. So exponential distributions are special gamma distributions.

10.4.2 Chi-square distribution

Definition 10.6. Let ν be a positive integer. X has a **chi-square distribution with ν degrees of freedom**, denoted X is $\chi^2(\nu)$, if X is $\texttt{gamma}(\alpha = \nu/2, \beta = 2)$.

Here are plots of three different chi-square distributions.



10.5 Beta Distribution

The beta probability distribution provides a way to model random variables whose possible outcomes are all real numbers between 0 and 1. Such distributions are useful for modeling proportions. As with the gamma and normal distributions, this is a 2-parameter family of distributions. Altering the parameters α and β gives us, well, different shapes for the density curves.

Definition 10.7. A random variable X has a **beta probability distribution with parameters $\alpha > 0$ and $\beta > 0$** if and only if it has probability density function

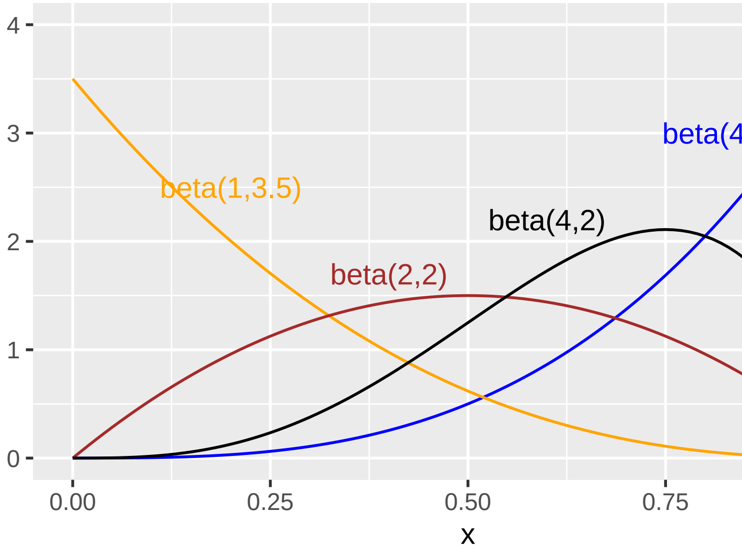
$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0 \le x \le 1 \quad (\text{and } f(x) = 0 \text{ else})$$

If X has such a pdf we say that X is $\texttt{beta}(\alpha, \beta)$.

The gamma function (10.5) appears in this pdf three times. Recall that for positive integers n, $\Gamma(n) = (n-1)!$ so for integer values of α and β, the beta density function is fairly nice. Indeed,

- $X \sim \texttt{beta}(1,1) \Rightarrow f(x) = \dfrac{\Gamma(2)}{\Gamma(1)\Gamma(1)} x^0(1-x)^0 = 1$. Whoa! beta$(1,1)$ is the uniform distribution $U(0,1)$.
- $X \sim \texttt{beta}(1,2) \Rightarrow f(x) = 2(1-x)$.
- $X \sim \texttt{beta}(2,1) \Rightarrow f(x) = 2x$.
- $X \sim \texttt{beta}(2,2) \Rightarrow f(x) = 6x(1-x)$.
- $X \sim \texttt{beta}(n,1) \Rightarrow f(x) = nx^{n-1}$.

- $X \sim \texttt{beta}(1, n) \Rightarrow f(x) = n(1-x)^{n-1}$.



Here are a few beta distributions:

Theorem 10.6. *If X is* $\texttt{beta}(\alpha, \beta))$, *then*

$$E(X) = \frac{\alpha}{\alpha+\beta}, \qquad \textit{and} \qquad V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Chapter 11

Moment Generating Functions

Recall by Definition 8.2, the moment-generating function (mgf) associated with a discrete random variable X, should it exist, is given by

$$m_X(t) = E(e^{tX})$$

where the function is defined on some open interval of t values containing 0. The same definition applies to continuous random variables. We have seen that this mgf encodes information about X: the kth derivative of m evaluated at $t = 0$ gives us the kth moment. That is, for $k = 1, 2, 3, \ldots$,

$$m_X^{(k)}(0) = E(X^k).$$

In fact, it turns out that the mgf gives us *all* the information about a random variable X, per the following theorem, whose proof is beyond the scope of this course.

Theorem 11.1. *Let $m_X(t)$ and $m_Y(t)$ denote the mgfs of random variables X and Y, respectively. If both mgfs exist and $m_X(t) = m_Y(t)$ for all values of t then X and Y have the same probability distribution.*

Example 11.1.

Find the mgf for the standard normal random variable $Z \sim N(0, 1)$.

$$\begin{aligned}
m_Z(t) &= E(e^{tZ}) \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \cdot e^{tz}\, dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz-z^2/2}\, dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t)^2+\frac{1}{2}t^2}\, dz && \text{complete the square} \\
&= e^{\frac{1}{2}t^2} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t)^2}\, dz \right]
\end{aligned}$$

The bracketed portion of this last expression equals 1, for all t, since it is the integral of the density function of a $N(t,1)$ distribution, so

$$m_Z(t) = e^{\frac{1}{2}t^2},$$

for all $-\infty < t < \infty$.

More generally, for $X \sim N(\mu, \sigma)$, one can show its mgf is

$$m(t) = e^{\left(\mu t + \frac{\sigma^2}{2} t^2\right)} \tag{11.1}$$

We now return to the proof of Theorem 10.4, which we restate as the following lemma.

Lemma 11.1. *If X is $N(\mu,\sigma)$ and $Z = \frac{X-\mu}{\sigma}$, then Z is $N(0,1)$.*

Proof. Let X be $N(\mu,\sigma)$, and $Z = \frac{X-\mu}{\sigma}$. Then the mgf for Z is

$$\begin{aligned}
m_Z(t) &= E\left[e^{tZ}\right] \\
&= E\left[e^{t\left(\frac{X-\mu}{\sigma}\right)}\right] \\
&= E\left[e^{\frac{Xt}{\sigma}-\frac{\mu t}{\sigma}}\right] \\
&= E\left[e^{Xt/\sigma} \cdot e^{-\mu t/\sigma}\right] \\
&= e^{-\mu t/\sigma} \cdot E\left[e^{Xt/\sigma}\right] \\
&= e^{-\mu t/\sigma} \cdot m_X(t/\sigma)
\end{aligned}$$

This last step follows because $E\left[e^{Xt/\sigma}\right]$ is the mgf of X evaluated at t/σ. Then,

$$\begin{aligned}
m_Z(t) &= e^{-\mu t/\sigma} \cdot e^{\left(\mu(t/\sigma) + \frac{\sigma^2}{2}(t/\sigma)^2\right)} \\
&= e^{t^2/2}
\end{aligned}$$

But hey! This mgf is the mgf for $N(0,1)$, so by Theorem 11.1, since $Z = (X-\mu)/\sigma$ and $N(0,1)$ have the same mgf, they have the same probability distribution. □

Lemma 11.2. *If Z is $N(0,1)$ then Z^2 is $\chi^2(1)$.*

The proof of this lemma is left for now.

Theorem 11.2. *Let $X_1, X_2, \dots, X_n$ be independent random variables with mgfs $m_1(t), m_2(t), \dots m_n(t)$, respectively. If $U = X_1 + X_2 + \cdots + X_n$ then*

$$m_U(t) = m_1(t) \cdot m_2(t) \cdot \ \cdots \ \cdot m_n(t).$$

Sketch of Proof:

$$\begin{aligned} m_U(t) &= E\left[e^{tU}\right] \\ &= E\left[e^{t(X_1+X_2+\cdots X_n)}\right] \\ &= E\left[e^{tX_1} \cdot \ e^{tX_2} \cdot \ \cdots \ \cdot e^{tX_n}\right] \\ &= E\left[e^{tX_1}\right] \cdot E\left[e^{tX_2}\right] \cdot \ \cdots \ \cdot E\left[e^{tX_n}\right] \\ &= m_1(t) \cdot m_2(t) \cdot \ \cdots \ \cdot m_n(t) \end{aligned}$$

That the $E[\ ]$ distributes through the product in line 4 above follows since the X_i are assumed to be independent. The rpoof of this fact would be given in MATH 440.

Theorem 11.3. *Let $X_1, X_2, \dots, X_n$ be independent normal random variables with $X_i \sim N(\mu_i, \sigma_i)$, and let $a_1, a_2, \dots, a_n$ be constants. If*

$$U = \sum_{i=1}^{n} a_i X_i,$$

then U is normally distribution with

$$\mu = \sum_{i=1}^{n} a_i \mu_i \qquad \text{and} \qquad \sigma^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

Proof. Since X_i is $N(\mu_i, \sigma_i)$, X_i has mgf

$$m_{X_i}(t) = e^{(\mu_i t + \sigma_i^2 t^2/2)},$$

and for constant a_i, the random variable $a_i X_i$ has mgf

$$m_{a_i X_i}(t) = E(e^{a_i X_i t}) = m_{X_i}(a_i t) = e^{(\mu_i a_i t + a_i^2 \sigma_i^2 t^2/2)}.$$

Then by Theorem 11.2 and properties of exponents, for $U = \sum a_i X_i$,

$$\begin{aligned} m_U(t) &= \prod_{i=1}^{n} m_{a_i X_i}(t) \\ &= \prod_{i=1}^{n} e^{(\mu_i a_i t + a_i^2 \sigma_i^2 t^2/2)} \\ &= e^{\left(t \sum a_i \mu_i + \frac{t^2}{2} \sum a_i^2 \mu_i^2\right)} \end{aligned}$$

But hey! This is the mgf for a normal distribution with mean $\sum a_i \mu$ and variance $\sum a_i^2 \sigma_i^2$, so we have proved the result. □

Theorem 11.4. *Let $X_1, X_2, \ldots, X_n$ be independent normal random variables with $X_i \sim N(\mu_i, \sigma_i)$, and $Z_i = \dfrac{X_i - \mu_i}{\sigma_i}$ for $i = 1, \ldots, n$.*

Example 11.2. Suppose the number of customers arriving at a particular checkout counter in an hour follows a Poisson distribution. Let X_1 record the time until the first arrival, X_2, the time between the 1st and 2nd arrival, and so on, up to X_n, the time between the $(n-1)$st and nth arrival. Then it turns out the X_i are independent, and each is an exponential random variable with density

$$f_{X_i}(x_i) = \frac{1}{\theta} e^{-x_i/\theta},$$

for $x_i > 0$ (and 0 else). Find the density function for the waiting time U until the nth customer arrives.

Well $U = X_1 + X_2 + \cdots + X_n$, so by Theorem 11.2,

$$m_U(t) = m_1(t) \cdot \ \cdots \ \cdot m_n(t) = (1 - \theta t)^{-n}.$$

But, hey! This is the mgf for a gamma$(\alpha = n, \beta = \theta)$ random variable so by Theorem 11.1, U *is* gamma(n, θ). So

$$f_U(u) = \frac{1}{(n-1)!\theta^n} u^{n-1} e^{-u/\theta},$$

for $u > 0$ (and 0 else).

Example 11.3. If Y_1 is $N(10, .5)$ and Y_2 is $N(4, .2)$ and $U = 100 + 7Y_1 + 3Y_2$, how is U distributed, and what value marks the 90th percentile for U?

Theorem 11.3 says that U is normal with

$$E(U) = 100 + 7 \cdot 10 + 3 \cdot 4 = 182,$$

and

$$V(U) = 0 + 7^2 \cdot (.5)^2 + 3^2 \cdot (.2)^2 = 12.61,$$

so $\sigma_U = \sqrt{12.61} = 3.55$.

The 90th percentile can be found in R with the `qnorm()` function:

```r
qnorm(.9,mean=182,sd=3.55)
```

```
## [1] 186.5495
```

Example 11.4.

> Find the moment-generating function for $X \sim U(\theta_1, \theta_2)$.

$$\begin{aligned}
m_X(t) &= E(e^{tX}) \\
&= \int_{\theta_1}^{\theta_2} e^{tx} \frac{1}{\theta_2 - \theta_1}\, dx \\
&= \frac{1}{\theta_2 - \theta_1} \frac{1}{t} e^{tx} \Big|_{\theta_1}^{\theta_2} \\
&= \frac{e^{t(\theta_2 - \theta_1)}}{t(\theta_2 - \theta_1)}.
\end{aligned}$$

Example 11.5.

Find the moment-generating function for $X \sim \text{gamma}(\alpha, \beta)$ and compute $E(X)$ and $V(X)$.

$$\begin{aligned}
m_X(t) &= E(e^{tX}) \\
&= \int_0^\infty e^{tx} \cdot \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-(x/\beta)}\, dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x(1/\beta - t)}\, dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \cdot \left(\frac{1}{1/\beta - t}\right)^\alpha \Gamma(\alpha) \int_0^\infty \frac{x^{\alpha-1} e^{-x(1/\beta - t)}}{\left(\frac{1}{1/\beta - t}\right)^\alpha \Gamma(\alpha)} \cdot\, dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \cdot \left(\frac{1}{1/\beta - t}\right)^\alpha \Gamma(\alpha)
\end{aligned}$$

The last integral above evaluates to 1 because it is the pdf for a gamma(α, β) distribution! After simplifying we obtain

$$m_X(t) = (1 - \beta t)^{-\alpha}.$$

With the mgf for a gamma random variable in hand, we can know derive its mean and variance, thus proving Theorem 10.5.

$$\begin{aligned}
m_X'(t) &= -\alpha(1 - \beta t)^{-\alpha - 1} \cdot (-\beta) \\
&= \alpha\beta(1 - \beta t)^{-\alpha - 1},
\end{aligned}$$

so

$$E(X) = m_X'(0) = \alpha\beta.$$

Turning to the second derivative,

$$\begin{aligned}
m_X''(t) &= (-\alpha - 1)\alpha\beta(1 - \beta t)^{-\alpha - 2} \cdot (\beta) \\
&= \alpha(\alpha + 1)\beta^2(1 - \beta t)^{-\alpha - 2},
\end{aligned}$$

so

$$E(X^2) = m_X''(0) = \alpha(\alpha+1)\beta^2.$$

Thus,

$$V(X) = E(X^2) - E(X)^2 = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2.$$

Example 11.6. The average velocity of nails shot from a nail gun is 2000 ft/s. Suppose the velocity varies according to a gamma(4,500) distribution, so the probability density function is

$$f(v) = \frac{v^3 e^{-v/500}}{6 \cdot 500^4},$$

for $v > 0$.

We note that this nail gun has the following (alarming?) velocity distribution:

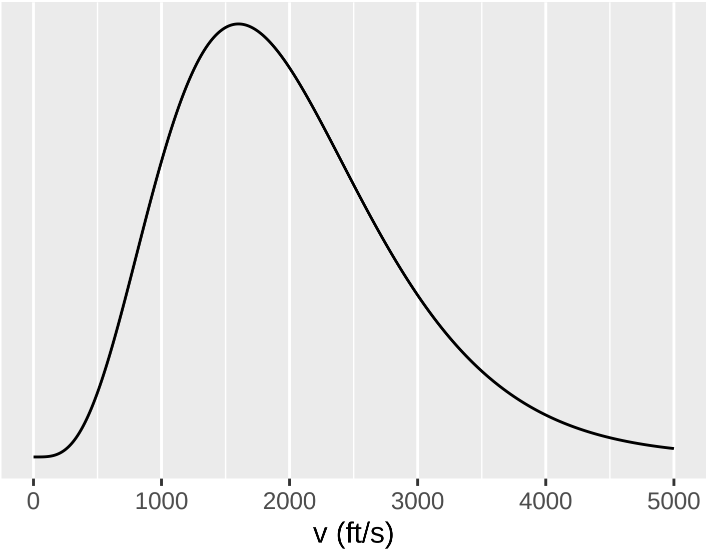


Figure 11.1: Nail gun velocity distribution

The kinetic energy K associated with a nail having mass m moving at velocity V is $K = \frac{1}{2}mV^2$. What is $E(K)$?

$$\begin{aligned} E(K) &= E(\frac{1}{2}mV^2) \\ &= \frac{1}{2}mE(V^2) \\ &= \frac{1}{2}m(\sigma_V^2 + \mu_V^2) \end{aligned}$$

Since V is gamme(4,500), $\mu_V = 4 \cdot 500 = 2000$ (as we were told) and $\sigma_V = 4 \cdot 500^2$, so

$$E(K) = 2500000m \text{ units.}$$

Chapter 12

Central Limit Theorem

12.1 Sums of Random Variables

Suppose $X_1, X_2, \dots, X_n$ are random variables defined via a random sample of size n taken from a distribution that is $N(\mu, \sigma)$.

After the sample is chosen, each $X_i = x_i$ takes on a value (lower case corresponds to data, upper case corresponds to random variable). We may then compute the sample mean

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Prior to picking our actual sample we can consider the function of the random variables

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Theorem 12.1. *If $X_1, X_2, \dots, X_n$ represents a random sample taken from a $N(\mu, \sigma)$ distribution, then*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \textit{is} \quad N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Proof. This theorem is an immediate consequence of Theorem 11.3 where each $a_i = 1/n$. □

Example 12.1. Let X equal the duration of a randomly selected song (in seconds) for a house finch. Suppose X is normal with unknown mean μ (we're trying to get a handle on this) and standard deviation $\sigma = 30$ seconds (somehow we know this). A random sample of 25 song durations is observed. Find the probability that the sample mean will be within 5 seconds of the population mean μ.

If $X_1, X_2, \dots, X_{25}$ denote the 25 song lengths to be observed, each $X_i \sim N(\mu, 30)$, so

$$\overline{X} \sim N\left(\mu, \frac{30}{\sqrt{25}}\right) = N(\mu, 6).$$

We want to know

$$P(|\overline{X}-\mu|<5).$$

$$\begin{aligned} P(|\overline{X}-\mu|<5) &= P(-5<\overline{X}-\mu<5) \\ &= P\left(\frac{-5}{6}<\frac{\overline{X}-\mu}{6}<\frac{5}{6}\right) \\ &= P(-5/6<Z<5/6). \end{aligned}$$

Using R, $P(-5/6<Z<5/6)=$ `pnorm(5/6)-pnorm(-5/6)` $=0.595$.

A secondary question: How big a sample we we need so that the likelihood of the sample mean being within 5 seconds of μ is up to .95?

In this case, we want n so that

$$P\left(\frac{-5}{30/\sqrt{n}}<Z<\frac{5}{30/\sqrt{n}}\right)=.95.$$

Equivalently, we want to find n so that

$$P\left(Z<\frac{-5}{30/\sqrt{n}}\right)=.025.$$

In $N(0,1)$, `qnorm(.025)` = -1.96, which means $P(Z<-1.96)=.025$. So we want

$$\frac{-5}{30/\sqrt{n}}=-1.96,$$

and solving for n and rounding up yields $n=139$.

Theorem 12.2. *Let $X_1, X_2, \dots, X_n$ represent a random sample from a $N(\mu,\sigma)$ distribution, and*

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \quad \textit{and} \quad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

Then

$$\frac{n-1}{\sigma^2}S^2\sim\chi^2(n-1),$$

and $\overline{X}$ and S^2 are independent random variables.

We refer to $\overline{X}$ and S^2 as the sample mean and sample variance associated with the random sample.

Suppose we draw a sample of size $n=25$ from a $N(10,2)$ distribution. In this case the preceding two theorems tell us that

- $\overline{X}\sim N(10,2/\sqrt{25})=N(10,0.4)$
- $6S^2\sim\chi^2(24)$ (since $\frac{n-1}{\sigma^2}=6$ in this case)
- $\overline{X}$ and S^2 are independent random variables.

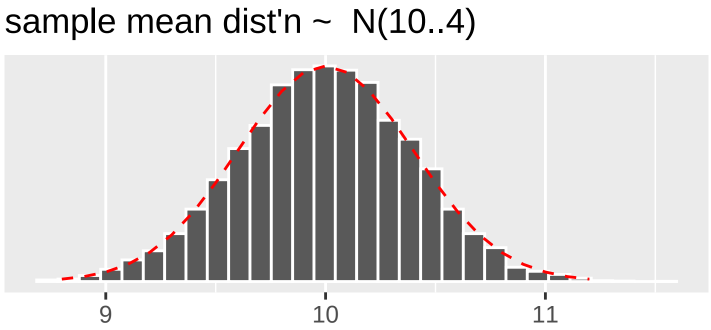
Let's look at a simulation in R to investigate these statements. The simulation works like this:

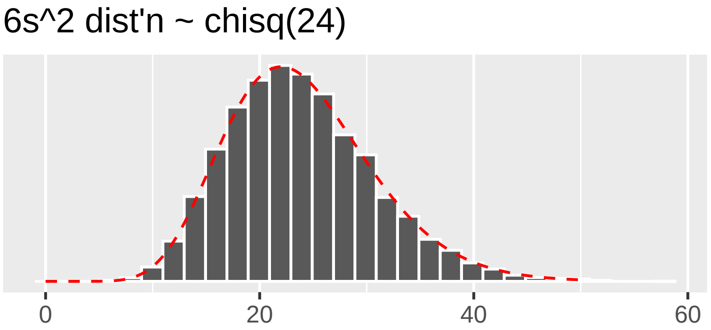
1. Draw a random sample of size 25 from $N(10,2)$
2. Calculate $\overline{x}$ and s^2 from this sample.

3. Repeat steps 1 and 2 for many trials, and then consider - a frequency plot for $\overline{x}$ (does it look $N(10, 0.4)$) - a frequency plot for $\frac{n-1}{\sigma^2}s^2$ (does it look $\chi^2(24)$?) - a scatter plot of $\overline{x}$ against s^2 (do they look independent?)

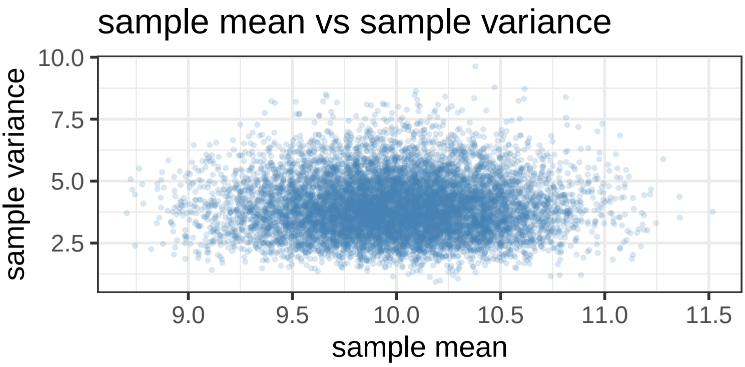
```
trials = 10000
n = 25; mu = 10; sigma = 2 #define sample size and parameters
sample_means = c() #stores mean of each sample
sample_var = c() #stores variance of each sample
for (i in 1:trials){
  x = rnorm(n,mu,sigma) #draw sample
  sample_means[i] = mean(x) #record sample mean
  sample_var[i] = var(x) #record sample variance
}
```

Plots:





The scatter plot below suggests no real association between $\overline{x}$ and s^2.



12.2 T distribution

Definition 12.1. Let $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$. If Z and W are independent then

$$\frac{Z}{\sqrt{W/\nu}}$$

is said to have a **t distribution with ν degrees of freedom.**

Here's our motivation for looking at such a thing. Look again at the house finch example (Example 12.1). We took a sample of 25 song lengths to estimate μ, or rather the likelihood that $\overline{x}$ is within 5 seconds of μ, the population mean. In our solution we assumed we know σ. It is perhaps not reasonable to assume we know σ when we're trying to estimate μ!

So, if we don't know σ, can we estimate it from the sample? Sure! How about estimating σ with s, the sample standard deviation?

Now, recall in our solution there came a point when we considered a Z-score:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}.$$

If we don't know σ can we replace it with the estimate s? Good question! Check this out:

From Theorem 10.4 $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is $N(0,1)$

From Theorem 12.2, $\dfrac{(n-1)S^2}{\sigma^2}$ is $\chi^2(n-1)$,

So the ratio $\dfrac{Z}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}$ has a t distribution with $(n$—$1)$ degrees of freedom!

Finally, observe

$$\begin{aligned}\frac{Z}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} &= \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{s/\sigma} \\ &= \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \cdot (\sigma/s) \\ &= \frac{\overline{X} - \mu}{s/\sqrt{n}}.\end{aligned}$$

The point of this story is this:

If $X_1, X_2, \dots, X_n$ represents a random sample drawn from $N(\mu, \sigma)$ then

- $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$
- so $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is $N(0,1)$
- while $T = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$ is a t distribution with $n-1$ degrees of freedom.

We denote a t distributioin with k degrees of freedom by $t(k)$. The density function for a $t(k)$ distribution, defined for all $-\infty < t < \infty$, is

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(k/2)} \left(1 + \frac{t^2}{2}\right)^{-(\frac{k+1}{2})}$$

Suppose $T \sim t(k)$.

Facts about T:

1. $E(T) = 0$
2. The distribution has mode at 0
3. The distribution is symmetric about the y-axis
4. it has fatter tails than $N(0,1)$, i.e.,for $a > 0$, $P(t > a) > P(Z > a)$.
5. As $k \to \infty$, $t(k) \to N(0,1)$.

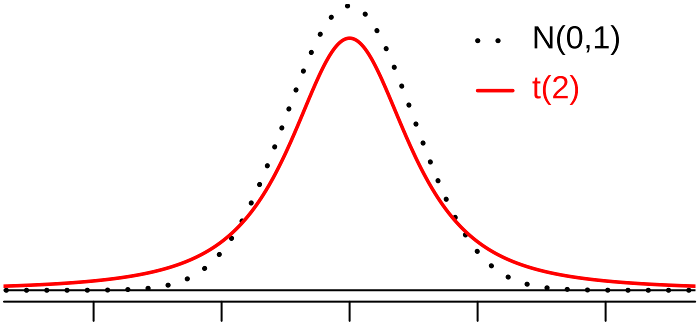


Figure 12.1: A t distribution and N(0,1)

Example 12.2. A forester studying the effects of fertilization on certain pine forests is interested in estimating the average basal area (in ft^2) of pine trees. Let $X_1, X_2, \dots, X_9$ denote a random sample of size 9, and suppose $X_i \sim N(\mu, \sigma)$ with μ, σ unknown.

Find two statistics (i.e., functions of the data) g_1 and g_2 such that

$$P(g_1 \leq \overline{X} - \mu \leq g_2) = .9.$$

(The statistics g_1 and g_2 thus give us a range of values we believe with probability .9 captures μ.)

Well, the statistic

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S}$$

lives in a $t(8)$ distribution.

Now $t(8)$ is plotted in figure 12.2, and we can find constants c and $-c$ such that the shaded area between them is 0.9.

Using R, in which the t distribution is aptly named as `t`, c and $-c$ are readily found with `qt()`:

```
qt(.95,8) #gives c
```

```
## [1] 1.859548
```

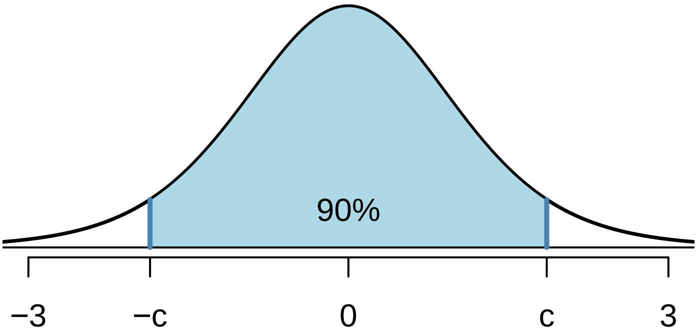


Figure 12.2: Finding the middle 90% of a t(8) distribution

So

$$P(-1.86 < T < 1.86) = .9,$$

where $T = 3(\overline{X} - \mu)/S$, and this allows us to solve the problem:

$$\begin{aligned} .9 &= P(-1.86 < T < 1.86) \\ &= P(-1.86 < 3(\overline{X} - \mu)/S < 1.86) < 1.86) \\ &= P(\frac{-1.86}{3} S < \overline{X} - \mu < \frac{1.86}{3} S) \\ &= P(-.62S < \overline{X} - \mu < .62S) \end{aligned}$$

So $g_1 = -.62S$ and $g_2 = .62S$ work!

In practice, this means that, once we have gathered our data of size $n = 9$, it is "likely" that μ is captured by the interval

$$(\overline{X} - .62S, \overline{X} + .62S).$$

For instance, suppose our data is (units are ft^2)

```
data = c(85.5,71.4,60.4,70.9,78.3,67.9,65.3,63.1,68.4)
xbar = mean(data)
s = sd(data)
```

It is "likely" that μ falls between `xbar-.62*s` = 65.3 ft^2 and `xbar-.62*s` = 74.9 ft^2.

The Central Limit Theorem says, roughly, that even if the underlying population is not normally distributed, it is still reasonable to follow this procedure to estimate μ.

12.3 The Central Limit Theorem

Theorem 12.3 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$ for*

$i = 1, 2, \ldots, n$. Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad U_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Then the distribution function of U_n converges to a standard normal distribution function as $n \to \infty$.

The Central Limit Theorem (CLT) is the mathematical basis for the statistical analysis coming in the next chapter.

Sketch of Proof

TODO

Example 12.3 (Practical Use of the CLT)**.** For large n,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

for a random sample taken from any distribution. That is, for any distribution (Poisson, binomial, gamma, uniform, …) with mean μ and standard deviation σ, if we take a simple random sample (SRS) of decent size, compute the sample mean, then this mean lives in a distribution that is approximately $N(\mu, \sigma/\sqrt{n})$. Consequently,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

will be approximately $t(n-1)$.

Example 12.4 (Can we really walk straight?)**.** Data on cadence (strides/sec) from a 1992 article in the American Journal of Physical Anthropology, for a sample of size $n = 20$ “randomly selected healthy men.”

```
data = c(0.95, 0.85, 0.92, 0.95, 0.93, 0.86, 1.00, 0.92, 0.85, 0.81,
0.78, 0.93, 0.93, 1.05, 0.93, 1.06, 1.06, 0.96, 0.81, 0.96)
```

The sample mean and standard deviation for these data are

- $\overline{x} = 0.925$
- $s = 0.081$.

We know that $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ has a t(19) distribution (assuming the underlying population is normal), and we can find c such that

$$P(|(\overline{x} - \mu)/(s/\sqrt{n})| < c) = .95.$$

12.4 Normal Approximation to a binomial distribution

If X is binom(n, p), we can view X as a sum of Bernoulli random variables: $X = \sum_{i=1^n} Y_i$ where each Y_i is binom$(1, p)$ [so $P(Y_i = 1) = p$ and $P(Y_i = 0) = 1 - p$, and

$$\mu_{Y_i} = p \quad \text{and} \quad \sigma_{Y_i} = \sqrt{p(1-p)}.$$

And

$$\frac{1}{n}X = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

By the Central Limit Theorem, for large n, it follows that $\frac{1}{n}X$ is approximately $N(p, \sqrt{p(1-p)/n})$.

> If X is binom(n, p) then X is approximately $N(np, \sqrt{np(1-p)})$ for large n.

Let's look at an example and then fine tune the approximation with a continuity correction.

Example 12.5. Suppose 44% of a voting population actually plan to vote for candidate A (though we don't know this :)). If we draw a random sample of $n = 100$ voters, what is the approximate probability that 51 or more of the 100 sampled plan to vote for candidate A?

If we know the size of the population we can answer this question precisely with the hypergeometric distribution:

For instance, suppose the population consists of 10000 voters, and X equals the number of voters in a sample of size 100 that plan to vote for candidate A. Then for any $x = 0, 1, \ldots, 100$,

$$p(x) = \frac{\binom{4400}{x}\binom{5600}{100-x}}{\binom{10000}{100}},$$

and

$$P(X \geq 51) = \sum_{x=51}^{100} p(x),$$

and this sum can be calculated in R by:

```
1-phyper(50,4400,5600,100)
```

```
## [1] 0.09442696
```

about a 9.4% chance.

Notice, if the populatioin is just 1000, the answer to this question would be `1-phyper(50,440,560,100)` = 0.0840868.

If we don't know the size of the population, but assume it's big, then the sampling process is close to that of 100 identical Bernoulli trials, where in each case, $p = .44$. In this case, X is binom$(n = 100, p = .44)$, and $P(X \geq 51)$ is found in R via

```
1-pbinom(50,100,.44)
```

```
## [1] 0.09553862
```

Notice that the binomial approximation here is closer to the actual probability calculated with the hypergeometric distribution for $n = 10000$ than for $n = 1000$.

Finally, let's approximate the likelihood with a normal distribution. According to the Central Limit Theorem, X is approximately $N(44, \sqrt{100(.44)(.56)})$, or

$N(44, 4.964)$. So $P(X \geq 51) = 1 - P(X < 51) =$ `1 - pnorm(51,44,4.964)` $=$ 0.079.

This normal estimate is a little low, and we can improve the estimate by making what is called a *continuity correction.*

Continuity Correction

Suppose X is `binom`$(100, .44)$, as in the voting example, and we want to estimate $P(51 \leq X \leq 55)$ by using the normal approximation $N(44, \sqrt{100(.44)(.56)})$.

The actual binomial probability can be represented as the sum of the 5 rectangle areas in Figure 12.3. Each rectangle has width 1, and the heights of the rectangles correspond to $P(X = x)$ (binomial probability) for each $x = 51, \ldots, 55$. We also see in the figure a portion of the $N(44, \sqrt{100(.44)(.56)})$ density curve $f(x)$. The area under f that best approximates the rectangle areas will be the integral with bounds [50.5,55.5] (whose area is shaded in the figure), as opposed to the integral with bounds [51,55].

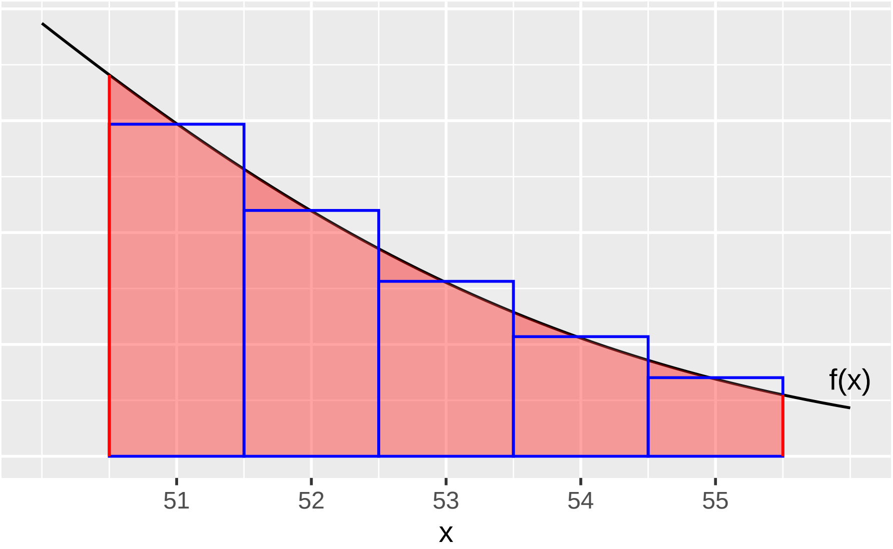


Figure 12.3: Continuity correction to estimate a binomial probability with a normal curve

In other words, to better approximate $P(51 \leq X \leq 55)$ with a normal distribution, instead of evaluating $\int_{51}^{55} f(x)\ dx$, we should use a continuity correction and evaluate

$$\int_{50.5}^{55.5} f(x)\ dx.$$

Observe:

- Actual value of $P(51 \leq X \leq 55)$:
  - `sum(dbinom(51:55,100,.44))` $= 0.08503$
- Normal approximation without continuity correction:
  - `pnorm(55,44,4.964)-pnorm(51,44,4.964)` $= 0.0659$
- Normal approximation with continuity correction:

  - `pnorm(55.5,44,4.964)-pnorm(50.5,44,4.964)` $= 0.08493$

Now, we return to our voting example and the normal approximation to the probability that at least 51 people in a sample of 100 people will vote for candidate A. With a continuity correction, $P(X \geq 51)$ is better approximated with:

```r
pnorm(100.5,44,4.964)-pnorm(50.5,44,4.964)
```

```
## [1] 0.09519473
```

Here's one more example.

Example 12.6. Use continuity correction to estimate $P(460 \leq X \leq 480)$ if X is `binom`$(1000, .5)$.

Well, with a continuity correction

$$P(460 \leq X \leq 480) \approx P(459.5 \leq Y \leq 480.5),$$

where $Y \sim N(500, \sqrt{1000(.5)(.5)})$.

- Actual probability: `pbinom(480,1000,.5)-pbinom(460,1000,.5)` $=$ 0.1025.

Estimated probability: `pnorm(480.5,500,sqrt(250))-pnorm(459.5,500,sqrt(250))` $= 0.1035$.

Chapter 13

Estimation

The Scene: We want to estimate some parameter θ of a population by gathering and analyzing an independent random sample drawn from the population.

13.1 Unbiased Estimators

Definition 13.1. A **statistic** is any function of a random sample $X_1, X_2, \dots X_n$ drawn from a population.

Definition 13.2. A statistic $\hat{\theta}$ based on a random sample $X_1, X_2, \dots X_n$ is an **unbiased estimator** of the population parameter θ if $E(\hat{\theta}) = \theta$.

The **bias of** $\hat{\theta}_n$ is $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

The **mean square error of** $\hat{\theta}_n$ is $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$.

A *good* estimator for the parameter θ is a statistic $\hat{\theta}$ that is unbiased with variance as small as possible. These features of $\hat{\theta}$ would ensure that for any random sample you happen to gather, the value $\hat{\theta}$ you compute from the data is likely to be close to θ (or at least likelier to be close to θ than some other statistic).

Example 13.1 (Two unbiased estimators for the upper bound of a uniform distribution)**.** Suppose $X_1, X_2, \dots, X_n$ is an independent random sample drawn from a uniform distribution $U(0, \theta)$, where θ is unknown. So each X_i is a random real number between 0 and θ.

> How can we estimate the unknown parameter θ from the data?

First estimator: Create an estimator from the sample mean:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

By properties of expected value,

$$\begin{aligned} E(\overline{X}) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\theta}{2} \end{aligned}$$

since each X_i is $U(0,\theta)$, so $E(X_i) = \frac{0+\theta}{2}$.

It follows that

$$E(\overline{X}) = \frac{1}{n} \cdot n \cdot \frac{\theta}{2} = \frac{\theta}{2}.$$

Since $E(\overline{X}) \neq \theta$, the sample mean $\overline{X}$ is not an unbiased estimator for θ. This makes sense. We shouldn't expect the average of the random numbers to be a good estimate of the upper bound of the interval from which the numbers were picked.

However, $E(\overline{X})$ *does* equal **a constant multiple** of θ, which means we can easily adjust $\overline{X}$ to a statistic that *is* an unbiased estimator for θ:

$$\hat{\theta}_1 = 2\overline{X} \tag{unbiased estimator 1}$$

Second Estimator: Create an estimator from the maximum value of the data, since this max is "closest" to θ of all the data points.

Let $Y = \max\{X_1, X_2, \ldots, X_n\}$. We prove below in 13.2 that

$$E(Y) = \frac{n}{n+1}\theta.$$

Assuming that for now, we can say that

$$\hat{\theta}_2 = \frac{n+1}{n} \cdot Y \tag{unbiased estimator 2}$$

is also an unbiased estimator for θ.

Let's see how these different estimators do for a particular random sample generated in R.

```
theta = 20 # we pretend we don't know this parameter
n = 10 # the size of the sample
X = runif(n,0,theta) # generate the random sample
est_1 = 2*mean(X)
est_2 = (n+1)/n*max(X)
print(round(X,2))
```

```
##  [1] 12.09 12.25 10.49  6.05  0.05 10.59 10.91 13.70  2.02  1.47
```

For this single sample the estimators takes on these values:

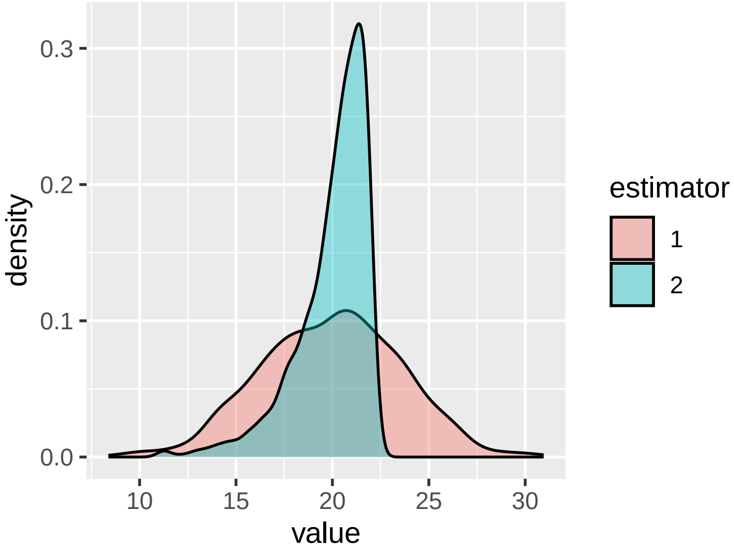
- Estimator 1: $\hat{\theta}_1 = 15.9$
- Estimator 2: $\hat{\theta}_2 = 15.1$.

The fact that both estimators are unbiased means that in the long run, the average of all the $\hat{\theta}_1$ estimates would approach θ, and the same is true for the average of the $\hat{\theta}_2$ estimates.

So, they're both good estimaors of θ in that regard. What makes one estimator better is if the *variation* of the estimates obtained from repeated sampling is smaller for one than the other.

Let's simulate drawing 1000 different samples of size $n = 10$, recording the distribution of values taken by the estimates θ_1 and θ_2, and seeing which distribution has smaller variance.

```
theta = 20;n = 10 # the size of the sample in each trial
trials = 1000 # number of times we take a sample of size n
dist_1 = c() # records estimator 1 values
dist_2 = c() # records estimator 2 values
# run the trials
for (i in 1:trials){
  X = runif(n,0,theta) # generate the random sample
  dist_1[i] = 2*mean(X)
  dist_2[i] = (n+1)/n*max(X)
}
# create data frame of results for ggplot
df = rbind(data.frame(estimator = "1",value = dist_1),
           data.frame(estimator = "2",value = dist_2))
# plot
ggplot(df)+
  geom_density(aes(x=value,fill=estimator),alpha = .4)+
  theme_get()
```



Both estimators have average value near 20. In fact,

- mean(dist_1)- $= 19.9$
- mean(dist_2) $= 19.99$.

But the estimator 2 distribution has visibly smaller variance. Indeed,

- `sd(dist_1)` = 3.6
- `sd(dist_2)` = 1.84.

It appears that the better estimator here is the one derived from the maximum value of the data as opposed to the mean of the data.

13.2 Order Statistics

If $X_1, X_2, \dots, X_n$ is a sample drawn from a distribution with density function $f_X(x)$, let

$$Y = \max\{X_1, X_2, \dots, X_n\}.$$

We can deduce the density function for Y by first writing down the distribution function. For any real number y,

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(\text{all } X_i \le y) \\ &= P(X_1 \le y, X_2 \le y, \dots, X_n \le y) \\ &= [F_X(y)]^n . \end{aligned}$$

We differentiate F_Y with the chain rule to find f_Y: Thus,

$$f_Y = n\,[F_X(y)]^{n-1} \cdot f_X(y). \qquad \text{(density for the max of sample)}$$

For X is $U(0,\theta)$ as in the previous example, $f_X(x) = 1/\theta$, and $F_X(x) = x/\theta$, for $0 \le x \le \theta$. So, the density function for $Y = \max(X_i)$, where X_i is $U(0,\theta)$ is

$$\begin{aligned} f_Y(y) &= n\left[\frac{y}{\theta}\right]^{n-1} \cdot \frac{1}{\theta} \\ &= \frac{n}{\theta^n} y^{n-1}, \end{aligned}$$

for $0 \le y \le \theta$, and

$$E(Y) = \int_0^\theta y \cdot \frac{n}{\theta^n} y^{n-1}\, dy = \cdots = \frac{n}{n+1}\theta,$$

giving us the result we assumed when defining estimator 2 in the previous example.

In the homework you derive the density function for the minimum of a random sample.

13.3 Common Unbiased Estimators

We have seen the following strategy for finding unbiased estimators: Try a simple estimator (e.g., $\overline{X}$ or $\max(X_i)$) and tweak it so that it becomes unbiased.

13.3.1 Estimating μ, a population mean

If $X_1, X_2, \dots, X_n$ is a sample drawn from a distribution with mean μ and standard deviation σ we have seen that the sample mean $\overline{X}$ has $E(\overline{X}) = \mu$ and standard deviation $\sigma/\sqrt{n}$. That is,

$\overline{X}$ is an unbiased estimator for μ, and its standard deviation is $\sigma/\sqrt{n}$.

13.3.2 Estimating p, a population proportion

If $X_1, X_2, \dots, X_n$ is a sample drawn from a $b(1, n)$ distribution (Bernoulli trial!) and $X = X_1 + X_2 + \cdots + X_n$ equals the number of successes in n trials, then we have seen that $\hat{p} = X/n$ has $E(\hat{p}) = p$ and standard deviation $\sqrt{\dfrac{p(1-p)}{n}}$. That is,

$\hat{p}$ is an unbiased estimator for p, and its standard deviation is $\sqrt{\dfrac{p(1-p)}{n}}$.

Example 13.2. In a sample of 65 Linfield students, 24 are first-generation students. Estimate p, the proportion of all Linfield students that are first-generation, and place a 2 standard deviation bound on the error of estimation.

From our sample, our point estimate for p is $\hat{p} = 24/65 \approx .369$. We know by the CLT that

$$\hat{p} \sim N(p, \sqrt{p(1-p)/n}),$$

and in a normal distribution, about 95% of the distribution is within two standard deviations of the mean. In other words,

$$P\left(\ |\hat{p} - p| < 2 \cdot \sqrt{p(1-p)/n}\ \right) \approx 0.95.$$

Now we don't know p (in fact, we're trying to estimate it!), so we can't know the value of the standard deviation $\sqrt{p(1-p)/n}$. However, for large n, the expression $\sqrt{x(1-x)/n}$ doesn't change much for nearby inputs, except when the inputs are close to 0 or 1 (try it!). In other words, we can reasonably expect

$$\sqrt{\hat{p}(1-\hat{p})/n}\ \approx \sqrt{p(1-p)/n},$$

in which case we can estimate that

$$P\left(\ |\hat{p} - p| < 2 \cdot \sqrt{\hat{p}(1-\hat{p})/n}\ \right) \approx 0.95.$$

In this problem $\hat{p} \approx .37$, so $2 \cdot \sqrt{p(1-p)/n} \approx 0.12$, so we can say that

$$|.37 - p| < 0.12$$

with probability about .95, and that

$$.25 < p < .49$$

gives a two standard deviation bound on the error of estimation.

13.3.3 Estimating $\mu_1 - \mu_2$, the difference of two population means

Suppose we have two independent random samples drawn from distinct normal distributions.

- $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1)$
- $Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2)$

Let $\overline{X}$ and $\overline{Y}$ denote the respective sample means, and consider the point estimate $\overline{X} - \overline{Y}$ of $\mu_1 - \mu_2$. Since $\overline{X} - \overline{Y}$ is a linear combination of normal random variables, we can show $\overline{X} - \overline{Y}$ is itself normal, with

- mean $= E(\overline{X} - \overline{Y}) = E(\overline{X}) - E(\overline{Y}) = \mu_1 - \mu_2$
- variance $= V(\overline{X} - \overline{Y}) = V(\overline{X}) + V(\overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

So $\overline{X} - \overline{Y}$ is an unbiased estimator for $\mu_1 - \mu_2$ with standard deviation equal to $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

13.3.4 Estimating $p_1 - p_2$, the difference of two population proportions

An unbiased estimator for $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_i$ equals the sample proportion from a sample of size n_i drawn from population i (which has population proportion p_i).

One can show that the point estimate $\hat{p}_1 - \hat{p}_2$ is approximately normal with mean $p_1 - p_2$ and standard deviation $\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$ for reasonably sized samples (more on this later). For now lets consider an example.

Example 13.3 (Two Large Buckets)**.** Each of two large buckets is full of two types of marbles, orange and green. Let p_i denote the proportion of orange marbles in bucket i ($i = 1, 2$).

Estimate $p_1 - p_2$, and place a 2 standard deviation bound on the error of estimation.

- Sample 1: $n_1 = 120$ marbles from bucket 1, of which 45 are orange, so $\hat{p}_1 = .375$
- Sample 2: $n_2 = 80$ marbles from bucket 2, of which 36 are orange, so $\hat{p}_2 = .45$.

Point estimate:

$$\hat{p}_1 - \hat{p}_2 = -.075.$$

Also, it is reasonable to assume that the sampling distribution for $\hat{p}_1 - \hat{p}_2$ is approximately

$$N\left(p_1 - p_2, \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

(again, plugging in $\hat{p}_i$ for p_i in the standard deviation is “ok” for larger sample sizes and values of $\hat{p}_i$ not too close to 0 or 1.).

For a normal distribution, about 95% of the distribution falls within 2 standard deviations of the mean, so

$$P\left(|(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)| < c\right) \approx .95,$$

where

$$c = 2 \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

So, we can say that

$$\hat{p}_1 - \hat{p}_2 - c < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + c$$

with probability about .95. Plugging in numbers in this example, we obtain the interval

$$-0.217 < p_1 - p_2 < 0.067.$$

Hmm... this interval contains 0. While the sample proportions here are such that we may suspect the second bucket has a higher proportion of orange marbles, when we take into account the error of estimation due to sampling variability, our samples do not provide overwhelming evidence that the buckets have different proportions of orange marbles.

Chapter 14

Confidence Intervals

In many of the examples in Chapter 13 we built confidence intervals, interval estimators that specify the method for using a sample to calculate two numbers that form the endpoints of an interval that likely (with some pre-assigned probability) contains the paramter of interest.

14.1 Pivotal Quantities

Here is the general scene for a **confidence interval** for a parameter θ. We use a sample to determine the **lower** and **upper confidence limit estimators**, $\widehat{\theta}_L$ and $\widehat{\theta}_U$. If

$$P(\widehat{\theta}_L \leq \theta \leq \widehat{\theta_U}) = 1 - \alpha$$

the probability $1 - \alpha$ is called the **confidence level**.

One method for finding confidence intervals is called the **pivotal method**, which leverages a **pivotal quantity**, which is a quantity with two features:

1. It is a function of the data $X_1, \dots, X_n$ and the parameter of interest θ, and
2. its probability distribution does not depend on θ.

Example 14.1. If a population mean μ is the parameter of interest, and we have sample $X_1, \dots, X_n$, then a good pivotal quantity, assuming X_i is approximately $N(\mu, \sigma)$, is

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

It meets both requirements here: it is a function of the data and the parameter μ, and it lives in a $t(n-1)$ distribution, which is independent of μ.

Example 14.2. If a population proportion p is the parameter of interest, and we have sample proportion $\widehat{p}$ (from a sample of size n), then the pivotal quantity

$$Z = \frac{\widehat{p} - p}{\sqrt{\widehat{p}(1-\widehat{p})/n}}$$

is approximately $N(0,1)$, a distribution independent of p.

Example 14.3. Consider $X_1, \dots, X_n$ drawn from the uniform distribution $U(0, \theta)$ and $Y = \max(X_1, \dots, X_n)$. We proved in Example 13.1 that

$$F_Y(y) = \left(\frac{y}{\theta}\right)^n \qquad \text{and} \qquad f_Y(y) = \frac{n}{\theta^n} y^{n-1},$$

for $0 \le y \le \theta$. Now, let $U = \frac{1}{\theta} Y$, which is just a function of the data and θ.

The distribution function for U is

$$\begin{aligned} F_U(u) &= P(U \le u) \\ &= P(Y/\theta \le u) \\ &= P(Y \le u \cdot \theta) \\ &= F_Y(u\theta) \\ &= \left(\frac{u\theta}{\theta}\right)^n && \text{for } 0 \le u \le 1 \\ &= u^n && \text{for } 0 \le u \le 1, \end{aligned}$$

which doesn't depend on θ. So $U = Y/\theta$ is a pivotal quantity.

Let's use U to build a 95% confidence interval for θ based on a sample $X_1, X_2, \dots, X_n$ drawn from the uniform distribution $U(0, \theta)$.

In particular, we find estimators $\hat\theta_L$ and $\hat\theta_U$ so that $P(\hat\theta_L \le \theta \le \hat\theta_U) = .95$.

Well, we know the distribution of the pivotal quantity $U = Y/\theta$, where Y is the max of the data, so one solution here is to find constants a and b between 0 and 1 so that

- $P(U < a) = 0.025$
- $P(U > b) = 0.025$

$$.025 = P(U < a) = F_U(a) = a^n \implies a = \sqrt[n]{.025}$$

and

$$.025 = P(U > b) = 1 - F_U(b) = 1 - b^n \implies b = \sqrt[n]{.975}.$$

With this we obtain a general form for a 95% confidence interval for θ:

$$\begin{aligned} .95 &= P\left(\frac{1}{\sqrt[n]{.025}} < U < \frac{1}{\sqrt[n]{.975}}\right) \\ &= P\left(\frac{1}{\sqrt[n]{.025}} < \frac{Y}{\theta} < \frac{1}{\sqrt[n]{.975}}\right) \\ &= P\left(\frac{1}{\sqrt[n]{.025}} > \frac{\theta}{Y} > \frac{1}{\sqrt[n]{.975}}\right) \\ &= P\left(\frac{Y}{\sqrt[n]{.975}} < \theta < \frac{Y}{\sqrt[n]{.025}}\right). \end{aligned}$$

We see that $\hat\theta_L = \frac{Y}{\sqrt[n]{.975}}$ and $\hat\theta_U = \frac{Y}{\sqrt[n]{.025}}$.

Let's put this interval formula to use. Here's a random sample of size 25 drawn in R from a $U(0, 64)$ distribution (though let's pretend we don't actually know $\theta = 64$ here.)

```
theta = 64
n = 25
x = runif(n,0,theta)
y = max(x)
L = y/(.975)^(1/n) #lower bound estimator for 95% CI
U = y/(.025)^(1/n) #upper bound estimator for 95% CI
c(L,U) #confidence interval
```

```
## [1] 63.37079 73.37223
```

Did our confidence interval actually capture the value of the parameter (64)? The confidence level of 95% gives us confidence that it does, in this sense: If we were to repeat the sampling procedure a large number of times, each time using our new sample to determine a new confidence interval for θ, then we should expect about 95% of the intervals to contain the parameter θ.

We can check this using R, first writing a little function out of the code above for producing the interval.

```
conf_int_for_theta <- function(n=25,theta=64){
  x = runif(n,0,theta)
  y = max(x)
  L = y/(.975)^(1/n)
  U = y/(.025)^(1/n)
  check = ifelse(((L <= theta)&(theta <= U)),1,0)#1 if interval captures theta, 0 e
  return(c(L,U,check)) #returns the interval and whether it captured theta
}

results = c() # a vector storing whether interval generated from data captures thet
for (i in 1:100){
  results[i]=conf_int_for_theta()[3]
}
table(results)
```

```
## results
##  0  1
##  4 96
```

In this simulation we drew 100 different random samples of size 25, each time generating a confidence interval for θ and found that 96 of the 100 intervals captured θ. We would expect in the long run that 95% of such confidence intervals would capture θ.

We have, generally,

- **2-sided confidence intervals for** θ: $[\hat{\theta}_L, \hat{\theta}_U]$, where $P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1-\alpha$,
- **1-sided confidence intervals for** θ:
    - $[\hat{\theta}_L, \infty)$, where $P(\hat{\theta}_L \leq \theta) = 1-\alpha$,
    - $(-\infty, \hat{\theta}_U]$, where $P(\theta \leq \hat{\theta}_U) = 1-\alpha$.

The interval $[\hat{\theta}_L, \infty)$ is called a **lower 1-sided** confidence interval, and in this case $\hat{\theta}_L$ is called the **lower confidence limit**, and $(-\infty, \hat{\theta}_U]$ is called an **upper 1-sided** confidence interval, and in this case $\hat{\theta}_U$ is called the **upper confidence**

limit

Goals for a good confidence interval:

1. It captures θ with high probability
2. It is narrow!

14.2 Large sample confidence intervals

Suppose θ is a parameter and $\hat{\theta}$ is an unbiased estimator of θ such that $\hat{\theta}$ is $N(\theta, \sigma_{\hat{\theta}})$. Then

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

is $N(0, 1)$ a nice pivotal quantity we can use to construct confidence intervals.

Suppose the desired confidence level is $1 - \alpha$ (it is common to let $\alpha = .05$, which corresponds to 95% confidence). Let $\pm z_{\alpha/2}$ denote the values in the tails of the $N(0, 1)$ distribution such that

$$P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha,$$

as pictured in Figure 14.1, where $\pm z_{\alpha/2}$ are denoted z_low and z_high.

Then

$$\begin{aligned} P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) &= 1 - \alpha \\ P\left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}\right) &= 1 - \alpha \\ P\left(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right) &= 1 - \alpha \end{aligned}$$

In this setting, a level $(1 - \alpha)$ confidence interval for θ is the interval

$$(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \ , \ \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}).$$

Note that in R,

- $-z_{\alpha/2} =$ `qnorm`$(\alpha/2)$, and
- $z_{\alpha/2} =$ `-qnorm`$(\alpha/2) =$ `qnorm`$(1 - \alpha + \alpha/2)$.

Similarly, a level $(1 - \alpha)$ **lower bound for** θ is

$$\hat{\theta} - z_{\alpha} \cdot \sigma_{\hat{\theta}}$$

(which defines the lower one-sided confidence interval $[\hat{\theta} - z_{\alpha} \cdot \sigma_{\hat{\theta}}, \infty)$) (see Figure 14.2).

A level $(1 - \alpha)$ **upper bound for** θ is

$$\hat{\theta} + z_{\alpha} \cdot \sigma_{\hat{\theta}}$$

(which defines the upper one-sided confidence interval $(-\infty, \hat{\theta} + z_{\alpha} \cdot \sigma_{\hat{\theta}}]$) (see Figure 14.3). Note

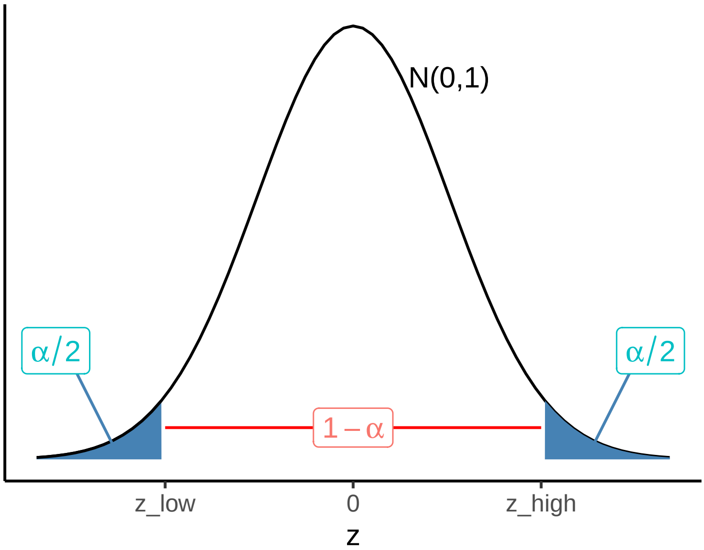


Figure 14.1: Finding z scores used to build a 2-sided confidence interval with desired confidence level

- $z_\alpha = \texttt{qnorm}(\alpha)$

The scene outlined above applies to many situations, thanks to the Central Limit Theorem, including the four common scenarios mention in Chapter 13:

- 1 mean μ, estimated with sample mean $\overline{X}$,
- 1 proportion p, estimate with sample proportion $\hat{p}$,
- difference of two means $\mu_1 - \mu_2$, estimated with $\overline{X}_1 - \overline{X}_2$,
- difference of two proportions $p_1 - p_2$, estimated with $\hat{p}_1 - \hat{p}_2$.

We summarize these confidence intervals here:

Common level $(1-\alpha)$ confidence intervals

1. For μ:

$$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

2. For p:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

3. For $\mu_1 - \mu_2$:

$$(\overline{X}_1 - \overline{X}_2) \pm z_{\alpha/2} \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

4. For $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

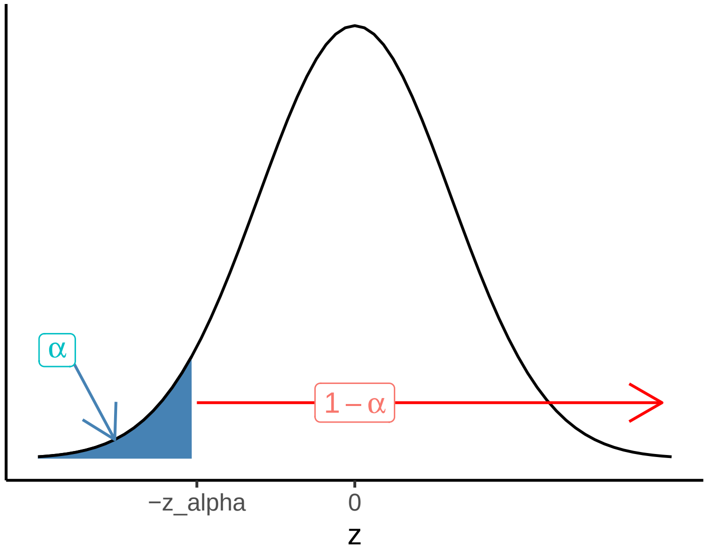


Figure 14.2: z scores corrsponding to lower 1-sided confidence interval

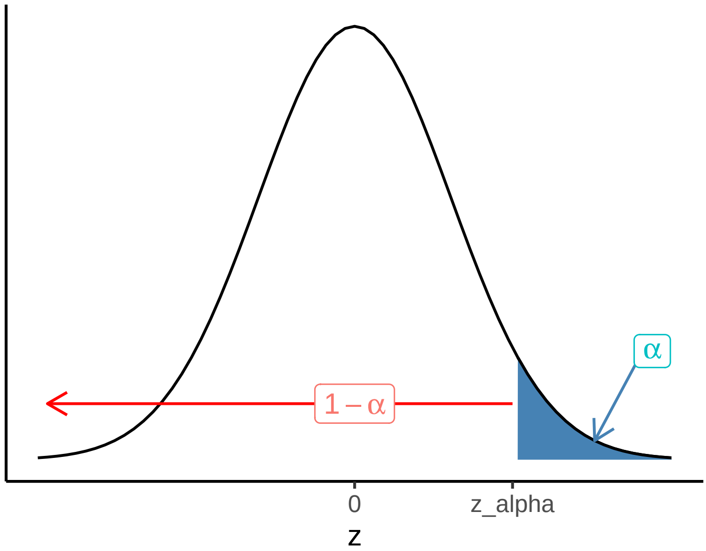


Figure 14.3: z scores corrsponding to upper 1-sided confidence interval

For large sample sizes ($n \geq 30$ is often sufficient), S can be used for unknown σ in confidence intervals for means, and sample proportions $\hat{p}$ can be used for unknown p. For small sample sizes, when σ is unknown, we shall use $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ when estimating means, and we have the following confidence interval for μ:

For small sample sizes, when σ is unknown, we estimate μ with

$$\overline{X} \pm t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

Let's work through several examples.

Example 14.4. Suppose we gather a sample of size $n = 20$ from a $N(\mu, 2)$ population, and find $\overline{x} = 16.3$. Find a 98% confidence interval for μ.

Here 98% confidence $\leftrightarrow$ $\alpha = 0.02$, and $z_{\alpha/2} = z_{.01} =$ `qnorm(.99)` $= 2.326$.

So our 98% confidence interval for the population mean μ is

$$\overline{x} \pm 2.326 \cdot \frac{2}{\sqrt{20}},$$

or

$$16.3 \pm 1.04,$$

or

$$15.26 \text{ to } 17.34.$$

Example 14.5. Suppose in this example we don't know $\sigma = 2$, but in our sample of size 20 from $N(\mu, \sigma)$, $\overline{x} = 16.3$ and $s = 2.13$. Find a 98% confidence interval for μ.

Using a sample standard deviation s in place of σ requires us to use t instead of z. In particular, instead of using $z_{.01} = 2.326$, we must use $t_{.01}(19)$, the value in the t(19) distribution that has 1% of the distribution greater than it.

So $t_{.01}(19) =$ `qt(.99,19)` $= 2.539$. (This t-score will always be larger than the corresponding z-score, making the confidence interval wider than it would have been if we knew σ.

Our confidence interval looks like:

$$\overline{x} \pm 2.539 \cdot \frac{2.13}{\sqrt{20}},$$

which simplifies to

$$16.3 \pm 1.21,$$

or, equivalently

$$15.09 \text{ to } 17.51.$$

Example 14.6. In a poll of 500 likely voters, 260 say they support a particular local measure. Based on this sample, find a 90% confidence interval for p, the proportion of all likely voters in favor of this measure.

We use the large sample confidence interval for p. From the data, $n = 500$, $\hat{p} = 260/500 = .52$, and for 90% confidence $\alpha = 0.1$, so

$$z_{\alpha/2} = z_{.05} = 1.645,$$

(computed in R with `qnorm(.95)`). Now evaluate:

$$\begin{aligned} \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= .52 \pm 1.645\sqrt{\frac{(.52)(.48)}{500}} \\ &= .52 \pm .037 \\ &= .483 \text{ to } .557. \end{aligned}$$

It looks like it will be a close vote! We do not have convincing evidence, really, that $p > .5$ since .5 lands inside the interval.

Example 14.7 (Angles of spider webs)**.** The Handbook of Small Data Sets published in 1994 has lots of interesting, small data sets. Here's one small data set:

```
X = c(25,12,31,26,17,15,24,10,16,12)
```

Spider webs' angles made with the vertical to the earth's surface have a von Mises circular distribution with known mean direction, μ, and μ varies from species to species. For instance,

- *Isoxya cicatricosa* has $\mu = 28.12°$, while
- *Araneus rufipalpus* has $\mu = 15.66°$, obviously.

The question arose (in the article "Sequential analysis for angular data", by Gadsden and Kanji, (1981), *The Statistician*, **30**, 119-129) of which species had constructed the 10 webs whose angles were listed above in the vector `X`.

Treating the data as a simple random sample we can construct a 95% confidence interval for μ. Since the sample size is small, we use $t_{\alpha/2}$, rather than $z_{\alpha/2}$, where, here $\alpha = .05$. Let's crunch out the interval in R.

```
#summary statistics
n = length(X)
xbar = mean(X)
s = sd(X)
tstar = qt(.975,9) #we let tstar denote t_{alpha/2}
xbar + c(-1,1)*tstar*s/sqrt(n)
```

```
## [1] 13.67688 23.92312
```

Based on the interval, which contains the known mean for *Araneus rufipalpus*, but not the known mean *Isoxya cicatricosa*, we have good evidence that these webs were made by the former type of spider.

Example 14.8 (Annual Snowfall in Buffalo, NY)**.** Here's another data set from the Handbook of Small Data Sets: The annual snowfall in Buffalo, NY (in inches) for the 63 years from 1910 to 1972.

```
##   year  snow
## 1 1910 126.4
## 2 1911  82.4
## 3 1912  78.1
## 4 1913  51.1
## 5 1914  90.9
```

Let's find a 95% confidence interval for the mean annual snowfall in Buffalo, taking the above snowfall column as a simple random sample of annual snowfall. Here are the summary statistics:

```
X = snowfall$snow #snow fall column as vector X
n = length(X)
xbar = mean(X)
s = sd(X)
```

For 95% confidence, $\alpha = .05$, and $z_{\alpha/2} = z_{.025} \approx 1.96$.

Since n is large, we use the confidence interval formula below with s plugged in for σ:

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

and we arrive at the interval

```
xbar + c(1,-1)*qnorm(.025)*s/sqrt(n)
```

```
## [1] 74.43965 86.15400
```

Confidence Intervals for $\mu_1 - \mu_2$

Suppose we have two samples from two populations: -an independent sample from population 1 which yields n_1, $\overline{x}_1$, s_1, and -an independent sample from population 1 which yields n_2, $\overline{x}_2$, s_2. Moreover, the two samples are independent.

Recall, if Population 1 is $N(\mu_1, \sigma_1)$ and Population 2 is $N(\mu_2, \sigma_2)$, then a 95% confidence interval for $\mu_1 - \mu_2$ is

$$\left(\overline{X_1} - \overline{X_2}\right) \pm z_{\alpha/2} \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

If σ_1 and σ_2 are unknown, but n_1 and n_2 are fairly large, it is reasonable to replace σ_1 and σ_2 with s_1 and s_2, respectively.

If n_1, n_2 are small (≤ 30) it's better to use a t-distribution.

Let's consider the special case that σ_1 and σ_2 are unknown but equal. Letting σ^2 denote this common value. In this case we consider the pooled sample variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

as our (best) estimate of σ^2. Then our level $(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $t_{\alpha/2}$ live in a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Example 14.9. The silver content (% Ag) of a number of Byzantine coins discovered in Cyprus was determined. Nine of the coins came from the first coinage of the reign of King Manuel, I, Comnenus (1143-1180); and 7 of the coins came from a coinage many years later.

The data

```
Ag1 <- c(5.9,6.8,6.4,7.0,6.6,7.7,7.2,6.9,6.2)
Ag2 <- c(5.3,5.6,5.5,5.1,6.2,5.8,5.8)
```

These data appear in *The Handbook of Small Data Sets* (p. 118), and are based on this article:

Hendy, M.F. and Charles J.A. (1970), *The production techniques, silver content and circulation history of the twelfth-century Byzantine Trachy.* Archaeonetry, 12. 13-21)

The question

Is there a significant difference in the silver content of coins minted early and late in Manuel's reign?

Let's conduct a small sample confidence interval for the difference in the two population means $\mu_1 - \mu_2$ from the summary statistics:

sample	n	xbar	s
1	9	6.74	0.543
2	7	5.61	0.363

The pooled sample standard deviation is then

$$s_p = \sqrt{\frac{8 \cdot s_1^2 + 6 \cdot s_2^2}{14}} \approx .474,$$

and for 95% confidence we use $t^* =$ `qt(.975,14)` $= 2.145$. With all these values we have the following 95% confidence interval for $\mu_1 = \mu_2$:

0.62 to 1.64.

The entire interval lies above 0, so we have good evidence here of a difference between μ_1 and μ_2.

Was it reasonable in the previous example to assume the two populations have equal variance? Might differences in silver content in these two eras make it likely that other production differences existed as well (that might make variation in silver content from coin to coin also change)?

If σ_1^2 and σ_2^2 are unknown, sample sizes are samll, and its unreasonalbe to assume σ_1^2 and σ_2^2 are equal, one often sees the following (approximate) level $(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$:

$$\overline{X}_1 - \overline{X}_2 \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{\alpha/2}$ lives in a t-distribution with degrees of freedom equal to the minimum of $n_1 - 1$ and $n_2 - 1$. So for the coin example, we would use $t^* =$ `qt(.975,6)` $=$

2.447, giving us a wider confidence interval in the end:

$$0.574 \text{ to } 1.686,$$

but still one that suggests a difference between μ_1 and μ_2.

Example 14.10 (Recent Buffalo Snowfall)**.** With climate change, it may be unreasonable to assume the average annual snowfall now is the same as it was 100 years ago.

Let μ_1 denote the "old" mean average annual snowfall (and we assume the data from 1910 to 1972 is a random sample from the "old" population"), and μ_2 denots the "new" mean average annual snowfall in Buffalo (from which the annual snowfall from 2000 to 2021, found at buffalo.or/snow, is a random sample.) Do these data provide evidence that $\mu_1 \neq \mu_2$?

Here are the 2001-2022 data:

```
Y = c(158.7,132.4,111.3,100.9,109.1,78.2,88.9,103.8,100.2,74.1,111.8,
      36.7,58.8,129.9,112.9,55.1,76.1,112.3,118.8,69.2,77.2,97.4)
```

A 95% confidence interval for the difference $\mu_X - \mu_Y$:

TODO

Example 14.11 (A one-sided confidence interval)**.** The measures of the outside diameter x_i (in inches) for 9 grains of the same type:

```
diam = c(2.021,2.002,2.001,2.005,1.990,1.990,2.009,1.983,1.987)
```

Assume the distribution is normal. Is $\mu \geq 2$?

Let's find a 1-sided lower bound 95% confidence interval for μ, and see what that gets us.

Since $n = 9$ is small and σ is unknown, we use the t-distribution.

Summary statistics:

- $n = 9$,
- $\overline{x} =$ `mean(diam)` $= 1.9987$.
- $s =$ `sd(diam)` $= 0.0122$

For a lower bound 95% confidence interval, $\alpha = .05$ and we use

$$t_\alpha = t_{.05}(8) = qt(.95, 8) = 1.86,$$

and the lower 1-sided interval is

$$\begin{aligned}(\overline{x} - t_{.05}(8) \cdot \frac{s}{\sqrt{n}}, \infty) &= (1.9987 - 1.86 \cdot \frac{0.0122}{\sqrt{9}}, \infty)\\ &= (1.991, \infty)\end{aligned}$$

This confidence interval does not convince me that $\mu \geq 2$. (If the interval had been something like $[2.01, \infty)$, then I would be more convinced since we're confident that μ is within the confidence interval, but with the interval $[1.991, \infty)$, best we can say is we're confident $\mu \geq 1.99$.

In MATH 340 we use R, a programming language designed for statistical analysis. In this tutorial we use R to estimate probability via simulation (repeated sampling from some set). We begin with data vectors, the structures we use to define sets in R.

Appendix A

Sampling in R

A.1 Data vectors

Use the `c()` command to enter an ordered list of elements. Separate entries with commas. The resulting object in R is called a **data vector**, or **vector**.

vector types

We see vectors of three types: **numeric**, **character**, and **logical**.

A character vector consists of a list of strings. Strings are entered with quotes.

```
animals = c("cat","rabbit","horse","boar","lynx")
```

The vector `x` below is numeric. No quotes, just numbers.

```
x = c(79.3,51.1,93.6,62.3,61.8)
```

A **logical** vector consists of a list of `TRUE`or `FALSE` elements (all caps!):

```
L = c(TRUE,FALSE,FALSE,FALSE,FALSE)
```

We can check the vector type with the `typeof()` command:

```
typeof(animals)
```

```
## [1] "character"
```

If you mix numbers and strings in a vector, R treats it as a character vector:

```
typeof(c(1,2,"potato"))
```

```
## [1] "character"
```

We may wish to place data vectors into a two-dimensional structure such as a matrix or a data frame.

Matrices

Create a matrix from a vector with the `matrix()` command, specifying how many rows, and whether we enter the data in the matrix by row, or by column.

```
matrix(c("a","a","a","b","b","b","c","c","c","d","d","d"),nrow = 4,byrow=TRUE)
```

```
##      [,1] [,2] [,3]
## [1,] "a"  "a"  "a" 
## [2,] "b"  "b"  "b" 
## [3,] "c"  "c"  "c" 
## [4,] "d"  "d"  "d"
```

```
matrix(c("a","a","a","b","b","b","c","c","c","d","d","d"),nrow = 3,byrow=FALSE)
```

```
##      [,1] [,2] [,3] [,4]
## [1,] "a"  "b"  "c"  "d" 
## [2,] "a"  "b"  "c"  "d" 
## [3,] "a"  "b"  "c"  "d"
```

data frames

A data frame links related vectors as columns in an array via the `data.frame()` command.

```
a = c("McMinnville","Denver","Minneapolis","Charleston")
x = c(45.21,39.74,44.98,32.78)
y = c(123.19,104.99,93.26,79.93)
df = data.frame(city = a, lat = x, long = y)
df
```

```
##          city   lat   long
## 1 McMinnville 45.21 123.19
## 2      Denver 39.74 104.99
## 3 Minneapolis 44.98  93.26
## 4  Charleston 32.78  79.93
```

Data frames are the most common way to manage related data vectors in R.

common vector commands

Here's a vector of Hank Aaron's home run totals in each of his MLB seasons:

```
hr = c(13,27,26,44,30,39,40,34,45,44,24,32,44,39,29,44,38,47,34,40,20,12,10)
```

With `hr` loaded into your session, you can refer to it by name when you want to extract features of it. Here are some commonly used commands on numeric vectors:

- `length(hr)`, number of elements in the vector (number of seasons Hank played)
- `sum(hr)`, the sum of the vector (total career home runs)
- `mean(hr)`, the mean of the vector (average HR per season)
- `max(hr)`, the max (best HR total in a season)
- `sd(hr)`, standard deviation
- `diff(hr)` returns a vector whose elements are the differences between consecutive elements in the vector `hr`
- `cumsum(hr)` returns a vector whose elements are the cumulative sum of the vector `hr`

- `rev(hr)` returns the vector in the reverse order

Behold:

```
diff(hr)
```

```
##  [1]  14  -1  18 -14   9   1  -6  11  -1 -20   8  12  -5 -10  15  -6   9 -13   6
## [20] -20  -8  -2
```

```
cumsum(hr)
```

```
##  [1]  13  40  66 110 140 179 219 253 298 342 366 398 442 481 510 554 592 639 673
## [20] 713 733 745 755
```

Comparison Operators

We compare things in R with various comparison operators, each one returning TRUE or FALSE:

- equal to `==`
- not equal to `!=`
- less than `<`
- less than or equal to `<=`
- greater than `>`
- greater than or equal to `>=`

A few examples:

```
12 >= 5
```

```
## [1] TRUE
```

Use **double equal signs** `==` to see whether two things are equal:

```
16 == 2*8
```

```
## [1] TRUE
```

```
"ab"=="ba"
```

```
## [1] FALSE
```

```
x = 3 # this defines the variable
x^2+3*x == 12 #this asks whether x^2 + 3*x equals 12 for the currently stored value
```

```
## [1] FALSE
```

Logical vectors arise when we give R a proposition involving a vector:

```
c(1,8,4,6) > 5
```

```
## [1] FALSE  TRUE FALSE  TRUE
```

`sum()` and `which()`

The `sum()` command on a numeric vector adds the elements of the vector, as we saw above with `sum(hr)`.

The `sum()` command on a logical vector returns the number of TRUE elements in the vector.

```
sum(c(TRUE,FALSE,TRUE,FALSE,TRUE,TRUE))
```

```
## [1] 4
```

We can thus easily count the number of elements in a vector meeting some condition:

```
sum(hr >= 40)
```

```
## [1] 8
```

8 seasons with at least 40 HR?!! Of course! 8! Ok, which seasons?

```
which(hr >= 40)
```

```
## [1]  4  7  9 10 13 16 18 20
```

The `which()` command returns the indices of the vector at which the condition being tested has been met. So Hank hit 40 or more HR in seasons 4, 7, 9, 10, 13, 16, 18, and 20.

extracting elements

Recall Hank Aaron's home runs by season:

```
hr
```

```
##  [1] 13 27 26 44 30 39 40 34 45 44 24 32 44 39 29 44 38 47 34 40 20 12 10
```

We can **extract an element** of a vector by indicating its [position]:

```
hr[3]
```

```
## [1] 26
```

Or we can specify several elements:

```
hr[c(1,3,5)]
```

```
## [1] 13 26 30
```

comparing vectors

We can count the number of positions in which two vectors of the same length agree

```
v = c(3,2,6,8); w = c(2,3,1,8)
sum(v==w) # how often they match
```

```
## [1] 1
```

We can find the position(s) at which they agree

```
which(v==w) #where they match
```

```
## [1] 4
```

and list the matching value(s):

```
v[which(v==w)]
```

```
## [1] 8
```

vector arithmetic

We can do element-wise arithmetic on two vectors of equal length, such as addition, subtraction, multiplication, divsiion, and exponentiation

```
v = c(-1,1,3)
w = c(1,4,5)
```

Operation	Result
`v + w`	0, 5, 8
`v - w`	-2, -3, -2
`v * w`	-1, 4, 15
`v / w`	-1, 0.25, 0.6
`v^w`	-1, 1, 243

We also have scalar multiplication,

```
8*v
```

```
## [1] -8  8 24
```

and, don't tell your vector calc prof, but you can add a scalar to each element in a vector:

```
8 + v
```

```
## [1]  7  9 11
```

concatenate vectors

The `c()` command allows you to concatenate vectors:

```
u = c(1,2,3)
v = c(4,5,6)
c(u,v)
```

```
## [1] 1 2 3 4 5 6
```

We can add an element to a vector `A` via concatenation:

```
A = c("Will","Lucas","Mike","Dustin")
A = c(A,"Eleven")
A
```

```
## [1] "Will"   "Lucas"  "Mike"   "Dustin" "Eleven"
```

Notice, the vector `A` currently has 5 elements. We can add a 6th element can also to `A` directly:

```
A[6]="Max" # creates new position after the last previous position
A
```

```
## [1] "Will"   "Lucas"  "Mike"   "Dustin" "Eleven" "Max"
```

A.2 Special vectors

consecutive integers

The integers 1 to n can be entered by typing `1:n`. For instance, we could define a 20 sided die by entering

```
die = 1:20
```

More generally, entering `a:b` creates a vector of consecutive integers starting with `a` and ending with `b` (even if `a` is greater than or equal to `b`).

```
8:2
```

```
## [1] 8 7 6 5 4 3 2
```

letters

```
letters
```

```
##  [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q" "r"
## [20] "t" "u" "v" "w" "x" "y" "z"
```

`LETTERS` is the capitalized version of the `letters` vector.

`rep()`

The `rep()` command lets us build a vector with lots of repeated elements.

Example A.1. Let's say we want to create a **bag of skittles** with this color distribution: 40 red, 30 orange, 25 yellow, 60 green, and 20 purple.

The `rep()` command let's us do this quickly: - first enter the distinct items (as a vector with the `c()` command!), - then enter how many times each occurs (as a vector!):

```
skittles = rep(c("red","orange","yellow","green","purple"),
               c(40,30,25,60,20)
               )
```

`table()`

The `table()` command is a handy way to see which unique values are contained in a vector and how often each unique value occurs:

```
table(skittles)
```

```
## skittles
##  green orange purple    red yellow
##     60     30     20     40     25
```

`seq()`

The `seq()` command lets us enter an arithmetic progression by entering (first, last, increment).

```
seq(0,1,by=.1)
```

```
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

Ok, now let's get to **sampling**!

A.3 Sampling

We use the `sample(x,...)` command to sample from vector `x`.

For instance, we can draw a random sample of size 2 from `hr`:

```
sample(hr,2)
```

```
## [1] 13 29
```

Here's another example. Let's grab 20 skittles at random from the bag `skittles` we created in example A.1 and count how many orange ones we get:

```
grab=sample(skittles,20)
table(grab)
```

```
## grab
##  green orange purple    red yellow 
##      7      2      1      9      1
```

The `table()` command counts how many of each color :).

We could have found the orange count directly with

```
sum(grab=="orange")
```

```
## [1] 2
```

`sample()` options

Typically, we provide the `sample()` command with 3 or 4 arguments, in this order:

- `x`, the vector we sample from
- `size`, the size of the sample
- `replace`, whether you sample with or without replacement (default = FALSE)
- `prob`, custom probabilities for the sampling of elements (default = equal probability for all elements in `x`)

If you enter your arguments in the order `x=, size=, replace=, prob=` then you do not need to specify the variable names.

If you do not specify their value, the `sample()` command assumes the following **default values**:

- `size` = the length of the vector
- `replace` = FALSE
- `prob` is set so all elements in the vector have equal probability of being chosen.

Here are handy special cases, illustrated with this vector:

```
animals = c("cat","dog","hedgehog","rabbit")
```

Permutations

Use `sample(x)` to generate a *random permutation* of `x`:

```
sample(animals) #default size is the length of the vector
```

```
## [1] "hedgehog" "dog"      "rabbit"   "cat"
```

Repeated sampling of 1 element

Use `sample()` to simulate picking one elemnt of `animals` n times by setting`size = n` and `replace = TRUE`.

Example; Draw one animal from the set 1000 different times and summarize the picks with a table.

```
picks=sample(animals,1000,replace=TRUE)
table(picks)
```

```
## picks
##       cat       dog hedgehog    rabbit
##       271       255       240       234
```

And the winner is… cat!

Or, since we fear rabbits and love dogs, we can do repeated sampling of a single element with **custom probabilities**:

```
picks2 = sample(animals,
                size=1000,
                replace=TRUE,
                prob=c(.2,.4,.3,.1))
table(picks2)
```

```
## picks2
##       cat       dog hedgehog    rabbit
##       202       419       294        85
```

Nice!

Remember, the default option for `sample()` is to sample without replacement, and with equal probabilities.

Sample without replacement

Task Pick 4 students at random from a class of 9 to race around Taylor Hall. (Assumes we have numbered the students 1-9).

```
sample(1:9,4)
```

```
## [1] 3 5 6 2
```

Sample with replacement

Task
On twelve consecutive days, ask one student at random, in a class of size 9, to write a solution on the board.

```
sample(1:9,12,replace=TRUE)
```

```
##  [1] 3 3 3 7 1 2 9 3 2 8 8 4
```

Sample with custom probabilities

Task
Roll a weighted 6 sided die with the following probability distribution 100 times and summarize the results.

x	1	2	3	4	5	6
$p(x)$	.2	.1	.05	.4	.1	.15

```
rolls = sample(1:6,
               size=100,
               replace=TRUE,
               prob=c(.2,.1,.05,.4,.1,.15))
table(rolls)
```

```
## rolls
##  1  2  3  4  5  6
## 14  5  6 52 11 12
```

Example: Lefties

Task
8% of a population is left-handed. Draw a random sample of 45 people from the population and record the number of lefties.

One approach: build a large population with these features and then draw 45 people from it without replacement.

```
pop=rep(c("L","R"),c(800,9200)) # a population of 10,000 people, 800 of them leftie
table(sample(pop,size=45))
```

```
##
##  L  R
##  2 43
```

A second approach: sample with replacement 45 times from a “two-sided” die with customized probabilities

```
die=c("L","R")
table(sample(die,size=45,replace=TRUE,prob=c(.08,.92)))
```

```
##
##  L  R
##  3 42
```

A third approach: Use a binomial distribution (later)

A.4 Repeated Sampling

Let’s say we have a huge urn full of orange and blue marbles, and 42% of them are orange. We can use repeated sampling to approximate the **sampling distribution** for the number of orange marbles we would draw in a sample of, say, 50 marbles. The sampling distribution provides information about what sorts of orange marble counts should we expect, and how often should we expect these counts?

Repeated sampling can estimate this sampling distribution. Here are two methods for achieving repeated sampling in R.

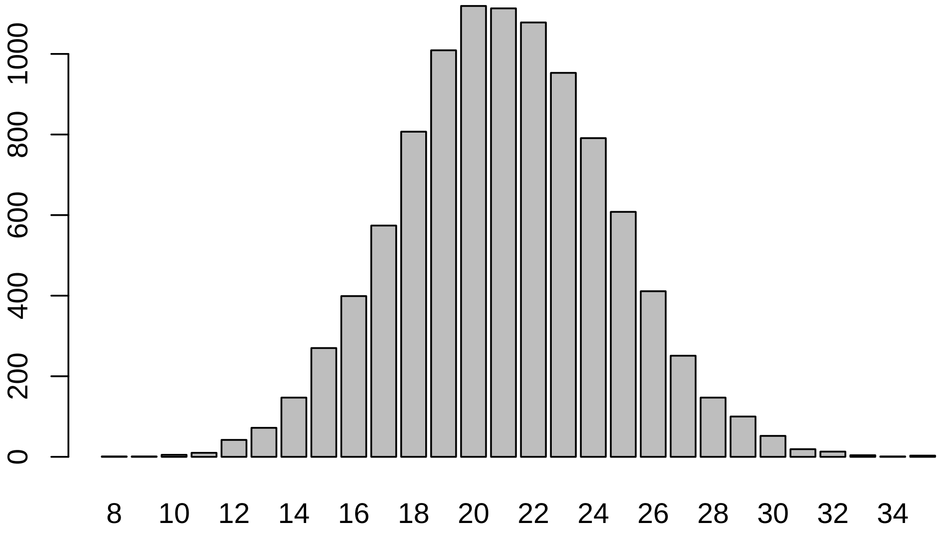
using a `for` loop

The code below creates a vector called *orange_counts* that, eventually, after the for loop has completed, has 10000 entries. Each entry in this vector gives the number of orange marbles drawn from the urn from a random sample of 50 marbles.

```
colors=c("orange","blue")
orange_counts=c() #a vector for storing the results of each trial
for (i in 1:10000){
  orange_counts[i]=sum(sample(colors,50,replace=TRUE,prob=c(.42,.58))=="orange")
}
```

We know that `table(orange_counts)` would display the counts of each of the unique values occuring in `orange_counts`. We can visualize these counts with a `barplot()`:

```
barplot(table(orange_counts))
```



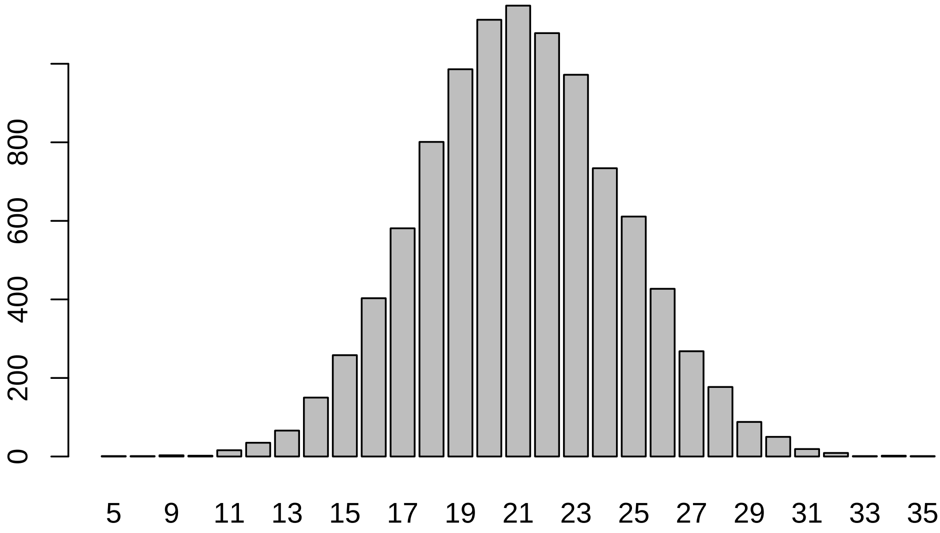
using replicate()

The `replicate()` command essentially does the above for loop for us :) The command `replicate(n,expr)` will evaluate `expr` `n` times, and store the results.

```
colors=c("orange","blue")
orange_counts =
  replicate(10000,
            sum(sample(colors,50,replace=TRUE,prob=c(.42,.58)) == "orange")
            )
```

Again, we can summarize the *frequency* with which each value of orange_counts occurs with `table()`, and visualize this frequency table with a `barplot`:

```
barplot(table(orange_counts))
```



In addition, we can calculate summary statistics to put numbers to qualitative descriptions of the distribution of values in `orange_counts` such as center and spread. These statistics help us answer the question of what sorts of orange counts to expect.

statistic	command	result
mean	`mean(orange_counts)`	21.0631
standard deviation	`sd(orange_counts)`	3.4885148
five number summary	`fivenum(orange_counts)`	5, 19, 21, 23, 35

A.5 Summary of R commands

defining vectors

Command	Description	Example
`c()`	List the elements	`x = c("a","c","c","z","z","z")`
`a:b`	Consecutive integers from `a` to `b`	`8:4` returns the vector 8, 7, 6, 5, 4
`rep()`	Build a vector from a frequency table	`rep(c("y","n"),c(3,2))` returns y, y, y, n, n
`seq()`	Arithmetic progression (first,last,step)	`seq(0,1,.2)` returns 0, 0.2, 0.4, 0.6, 0.8, 1

summarizing vectors

Command	Description
`typeof(x)`	the vector type of `x` (usually *character*, *numeric*, or *logical*)
`length(x)`	the length of `x` (how many elements it has)
`table(x)`	the frequency table (which values occur in `x` along with how often each value occurs)

sampling from vectors

Sampling Options	Example with `x = 1:6`
permutation of `x`	`sample(x) =` 5, 3, 6, 1, 4, 2
sample n elements without replacement	`sample(x,3) =` 1, 3, 6
sample n elements with replacement	`sample(x,5,replace=T) =` 3, 5, 2, 6, 2
sample with custom probabilities	`sample(x,10,replace=T,prob= c(0,.2,0,.5,.1,.2)) =` 4, 4, 4, 2, 5, 5, 4, 4, 6, 4

Appendix B

Simulating Probability in R

We can use repeated sampling of a chance experiment to estimate the probability of some event. The examples here were chosen to provide an introduction to sampling scenarios that you may find useful in this class (rolling dice, flipping coins, drawing names from a hat, marbles from a box,...). These examples also assume no previous knowledge of R beyond what was covered in Appendix A, and have been ordered roughly by my view of the complexity of the coding involved.

B.1 Difference of two dice

> If you roll 2 6-sided dice, what's the likelihood that their values are 1 apart.

Our Strategy: roll the two dice, record the absolute value of their difference, repeat!

Code:

```
trials = 10000
die_1 = 1:6
die_2 = 1:6
results = c() #stores the difference of the two dice each trial
for (i in 1:trials){
  roll_1 = sample(die_1,1)
  roll_2 = sample(die_2,1)
  difference = abs(roll_1-roll_2)
  results[i] = difference
}
```

Results:

```
table(results)
```

```
## results
##    0    1    2    3    4    5
## 1614 2855 2197 1725 1085  524
```

Conclusion: By computing the ratio `sum(results==1)/trials`, we obtain our estimate of the probability the two dice to be 1 apart: 0.2855.

Referring to the 6x6 grid recording the difference for each of the 36 possible outcomes in Table 3.4, we would find the actual probability equal to $10/36 \approx 0.278$.

B.2 Oregon License Plates

> Classic Oregon license plates consist of 3 letters (A-Z) followed by 3 numbers (0-9). Find the probability that a randomly selected plate has two 8s.

This is not such an interesting probability question - at least not one we need to simulate - but it's fun to build a random Oregon license plate!

Strategy:

1. Build a random plate. We want three random letters (repeats ok), followed by three random numbers (repeats ok). So, we sample from `LETTERS` 3 times with replacement, then 0:9 three times with replacement.

```
plate = c(sample(LETTERS,3,replace=TRUE),
          sample(0:9,3,replace=TRUE))
plate
```

```
## [1] "E" "N" "C" "4" "1" "0"
```

Fun!

2. Count how many "8"s are in the plate

```
sum(plate=="8")
```

```
## [1] 0
```

Final Code

```
trials = 10000
results = c()
for (i in 1:trials){
  plate = c(sample(LETTERS,3,replace=TRUE),
            sample(0:9,3,replace=TRUE)) #build a plate
  eights = sum(plate=="8") # count the 8s
  results = c(results,eights) # update the results vector
}
table(results)
```

```
## results
##    0    1    2    3
## 7222 2486  282   10
```

Based on this simulation, we estimate the probability of having a plate with exactly 2 "8"s

```
sum(results=="2")/trials
```

```
## [1] 0.0282
```

Using our counting tools to calculate the probability (and we really don't need to keep track of the letters in the plate, but we do):

$$\frac{26 \cdot 26 \cdot 26 \cdot \binom{3}{2} \cdot 1 \cdot 1 \cdot 9}{26^3 \cdot 10^3} = \frac{27}{1000} = 0.027.$$

B.3 Rolling a 10-sided die

> We're rolling a fair 10-sided die. Use simulation to estimate the probability that the first time we roll an 8 or higher is on the 5th roll.

Scratch Work

1. Simulate rolling a 10-sided die 5 times. (Sample 5 times with replacement from the vector 1:10.)

```
die = 1:10
rolls = sample(die,5,replace=TRUE)
```

2. We want code to check whether a random sequence of 5 rolls has these five features: Rolls 1-4 are less than 8, and roll 5 is an 8, 9, or 10.

We can ask these five questions in R and store the answers in a logical vector:

```
c(rolls[1:4]<8,rolls[5]>=8)
```

```
## [1]  TRUE  TRUE  TRUE FALSE FALSE
```

Recall, the `sum()` of a logical vector counts the number of TRUE values. We need the sum to be 5 to have the kind of sequence we want to count.

```
sum(c(rolls[1:4]<8,rolls[5]>=8)) #does this sum equal 5?
```

```
## [1] 3
```

I think we're ready!!

Final Code

```
trials = 10000
results = c() #stores how many dice "do the right thing" in each trial

die = 1:10
for (i in 1:trials){
  rolls = sample(die,5,replace=TRUE) # roll the 10-sided die 5 times
  x=sum(c(rolls[1:4]<8,rolls[5]>=8)) #how many dice "do the right thing"
  results=c(results,x) # update the results vector
}
```

The results of the simulation:

```
table(results)
```

```
## results
##    0    1    2    3    4    5 
##   54  531 2050 3658 3016  691
```

Based on our simulation, we estimate the probability in question by the ratio: `sum(results==5)/trials` = 0.0691.

This estimated probability is likely very close to what we calculate by our counting tools:

$$\frac{7^4 \cdot 3}{10^5} \approx .07203.$$

Final note: In this example we added to the `results` vector each iteration by concatenation (`results=c(results,x)`) rather than by specifying `results[i]=x` in each trial, as we did in the difference of two dice example. Either approach works.

B.4 Marbles from an urn

> An urn contains 100 orange and 200 green marbles. If you draw 8 marbles from the urn at random (without replacement), how likely is it that more than 5 of them are orange?

```
urn = rep(c("orange","green"),c(100,200))
trials = 10000
results = c()
for (i in 1:trials){
  grab = sample(urn,8,replace=FALSE)
  orange_count = sum(grab=="orange")
  results[i] = orange_count
}
table(results)
```

```
## results
##    0    1    2    3    4    5    6    7 
##  380 1506 2737 2804 1719  674  158   22
```

We can enter `sum(results > 5)` to see how often we grabbed more than 5 orange marbles in our sample of size 8, and `sum(results > 5)/trials` is a good estimate of the likelihood of this happening.

Conclusion: It appears we should expect the more than 5 orange marbles in our sample of 8 about 1.8% of the time:

```
sum(results > 5)/trials
```

```
## [1] 0.018
```

This question is a classic "good potatoes/bad potatoes" problem, and by our counting techniques, we know the probability is

$$\frac{\binom{100}{6}\cdot\binom{200}{2}}{\binom{300}{8}}+\frac{\binom{100}{7}\cdot\binom{200}{1}}{\binom{300}{8}}+\frac{\binom{100}{8}\cdot\binom{200}{0}}{\binom{300}{8}},$$

which we can evaluate in R as a check:

```
sum(choose(100,6:8)*choose(200,8-6:8)/choose(300,8))
```

```
## [1] 0.01830405
```

Note: After studying common named discrete probability distributions, we will see that R has a nice built-in command for doing these sorts of computations.

B.5 Tracking runs of Heads in coin flips

> If you flip a coin 20 times, how likely is it to have a run of at least 5 Heads in a row?

One approach:

- Build a coin vector: `coin = c("H","T")`
- Build a vector for recording the outcomes of twenty flips: `flips = sample(coin,size=20,replace=TRUE)`
- Record the longest run of Heads in the sequence.
- Repeat many, many times.

The following code plays this 'flip a coin 20 times and record the longest run of heads' game for 10000 trials.

```
coin = c("H","T")
trials=10000
results=c() #stores longest run of Heads each trial
for (i in 1:trials){
  flips = sample(coin,20,replace=TRUE) #generate the 20 flips
  run = 0 #marker for current streak of Heads
  max_run = 0 #stores the longest streak of Heads
  for (k in 1:20){
    if (flips[k] == "H"){
      run = run + 1 #current run increases by 1 if flip k is "H"
      max_run = max(max_run,run) #checks for new max run
      }
    if (flips[k] == "T"){
      run = 0 #resets current run to 0 if we flip k is "T"
    }
    results[i]=max_run
  }
}
```

Summary of results:

```
table(results)
```

```
## results
##    1    2    3    4    5    6    7    8    9   10   11   12   13   15
##  190 1942 3090 2270 1270  663  296  148   65   30   14   17    3    2
```

Are you kidding me? In 2 of the trials we saw 15 Heads in a row!? Based on these trials, our estimate for the likelihood of seeing a run of at least 5 Heads in 20 flips is `sum(results >=5)/trials` = 0.251.

B.6 Splitting a set into multiple subsets

> A class has 12 people: 6 juniors, 4 sophomores, and 2 first-years. The class is randomly divided into 3 subgroups of size 5, 4, and 3. What is the probability that the 2 first-years are in the same subgroup?

One approach:

1. Build the class: `class = rep(c("J","S","F"),c(6,4,2))`
2. Partition the members into three subgroups of size 5, 4, and 3. Our approach: shuffle the `class` vector (find a random permutation), take the first five in this permutation for subgroup 1, the next 4 for subgroup 2, and the last 3 for subgroup 3.

```
shuffle = sample(class)
sub1 = shuffle[1:5]
sub2 = shuffle[6:9]
sub3 = shuffle[10:12]
# the code below displays the subgroups as a check
cat("Subgroup 1: ", sub1, "\n",
    "Subgroup 2: ", sub2, "\n",
    "Subgroup 3: ", sub3, sep = "")
```

```
## Subgroup 1: FJFSS
## Subgroup 2: SJJJ
## Subgroup 3: JSJ
```

3. Count the number of first-years in each subgroup, and record these numbers as a vector of length 3:

```
count = c(sum(sub1=="F"),
            sum(sub2=="F"),
            sum(sub3=="F"))
print(count)
```

```
## [1] 2 0 0
```

4. The two first-years are in the same subgroup if and only if `count` contains a 2. The following code uses an `ifelse()` command to record 1 if both first-years are in the same group, and 0 if not.

```
ifelse(2 %in% count,1,0)
```

```
## [1] 1
```

Final Code

We put it all together now. In particular, we repeat the following process for 10000 trials: Partition the class into the three subgroups, count the “F”s in each subgroup, record 1 if both “F”s find themselves in the same group, and 0 otherwise. We store these 1s and 0s in the vector `results`.

```
class = rep(c("J","S","F"),c(6,4,2))

trials = 10000
results = c()
for (i in 1:trials){
  shuffle = sample(class) #randomly shuffles the 12 people.
  sub1 = shuffle[1:5]  # first 5 in the random shuffling go to subgroup 1
  sub2 = shuffle[6:9] # next 4 to subgroup 2
  sub3 = shuffle[10:12] # last 3 to subgroup 3
  count = c(sum(sub1=="F"),sum(sub2=="F"),sum(sub3=="F"))
  #are both "F"s in the same subgroup? We record 1 if "yes", and 0 if "no"
  results[i] = ifelse(2 %in% count,1,0)
}
table(results)
```

```
## results
##    0    1
## 7088 2912
```

Based on this simulation, we estimate the probability that both first-years end up in the same subgroup as

```
sum(results==1)/trials
```

```
## [1] 0.2912
```

Using our counting tools to calculate the probability:

$$\frac{\binom{10}{3\ 4\ 3}+\binom{10}{5\ 2\ 3}+\binom{10}{5\ 4\ 1}}{\binom{12}{5\ 4\ 3}}$$

which evaluates to $7980/27720 \approx 0.288$:

```
(factorial(10)/(6*24*6)+factorial(10)/(120*2*6)+factorial(10)/(120*24*1))/(factoria
```

```
## [1] 0.2878788
```

B.7 Pollsters

In an upcoming election for mayor of a large city, a pollster plans to predict the winner of the popular vote by taking a random sample of 1000 voters and declaring that the winner will be the one obtaining the most votes in his sample. Suppose that 48 percent of the voters plan to vote for the Republican candidate and 52 percent plan to vote for the Democratic candidate. To get some idea of how reasonable the pollster's plan is, write a program to make this prediction by simulation. Repeat the simulation 1000 times and see how many times the pollster's prediction would come true.

First, let's create and summarize a single poll of 1000 people from a population in which 52 percent are "D", and 48 percent are "R".

```
one_poll = sample(c("D","R"),1000,replace=TRUE,prob=c(.52,.48))
table(one_poll)
```

```
## one_poll
##   D   R
## 531 469
```

Of course, the goal is to use the poll to predict the winner of the election. We use the `sum()` command to count how many elements in `one_poll` equal "D" (use those double equal signs), and the `ifelse()` command to record the predicted winner.

```
ifelse(sum(one_poll=="D") > 500,"Dem wins","Tie or Rep wins")
```

```
## [1] "Dem wins"
```

Now, our goal is to repeat this sampling and prediction procedure 10000 times, and keep track of the predicted winner in each trial.

```
dem = .52 #proportion voting "D"
rep = 1 - dem # proportion voting "R"
poll_size = 1000 # poll sample size
trials = 10000
results = c()

for (i in 1:trials){
  poll = sample(c("D","R"),
                size = poll_size,
                replace = TRUE,
                prob = c(dem,rep))
  results=c(results,
            ifelse(sum(poll == "D") > poll_size/2,"D","Tie or R"))
}
table(results)
```

```
## results
##        D Tie or R
##     8924     1076
```

This table gives us a sense of the likelihood that the pollsters plan will result in an accurate prediction of which candidate will win the election. Based on our simulation, that likelihood is about 0.892.

B.8 Matching Birthdays

Suppose you ask n random people their birthday (month and day, disregarding year). What is the probability that at least one of them shares your birthday? What is the probability that at least two of them share the same birthday? (Assume 365 days in a year - ignore leap days.)

We do not use R to estimate these probabilities in this example. We find the probabilities exactly, and then use R to analyze them as n changes.

We tackle the first question by first computing the probability that none of the n people share my birthday. The probability that a random person does not have my birthday is $\frac{364}{365}$, and the probability that n random people all do not have my birthday is $\left(\frac{364}{365}\right)^n$.

The *complement* of this event is that at least one of the n people has my birthday, and so the probability of this will be $1 - \left(\frac{364}{365}\right)^n$.

We can write a function in R to compute this probability for various values of n.

```r
prob_my_bday <- function(n){
  return(1-(364/365)^n)
}
```

For instance, in a group of 15 people here's the probability that someone shares my birthday:

```r
prob_my_bday(15)
```

```
## [1] 0.04031703
```

Not too likely!

For the second question, we begin by calculating the probability that no one in a group of n people shares the same birthday as anyone else.

So, we need all n people to have a different birthday. This probability is found in essentially the same way that you answered #6 in Homework 2, and it equals

$$\frac{P_r^n}{365^n} = \frac{365 \cdot 364 \cdot \cdots \cdot (365 - n + 1)}{365^n}.$$

So, the probability that at least two people share the same birthday, which is the complement of "no one shares the same birthday" is

$$1 - \frac{365 \cdot 364 \cdot \cdots \cdot (365 - n + 1)}{365^n}.$$

For instance, in a group of 6 people, the probability that at least two people share a birthday is

$$1 - 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \cdot \frac{361}{365} \cdot \frac{360}{365} \approx 0.04.$$

Here's a function that will compute this probability for any group size n.

```
prob_shared_bday <- function(n){
  p=1
  for (i in 1:n){
    p=p*(365-i+1)/365
    }
  return(1-p)
}
```

For instance, in a group of 15 people here's the probability that at least two people share a birthday

```
prob_shared_bday(15)
```

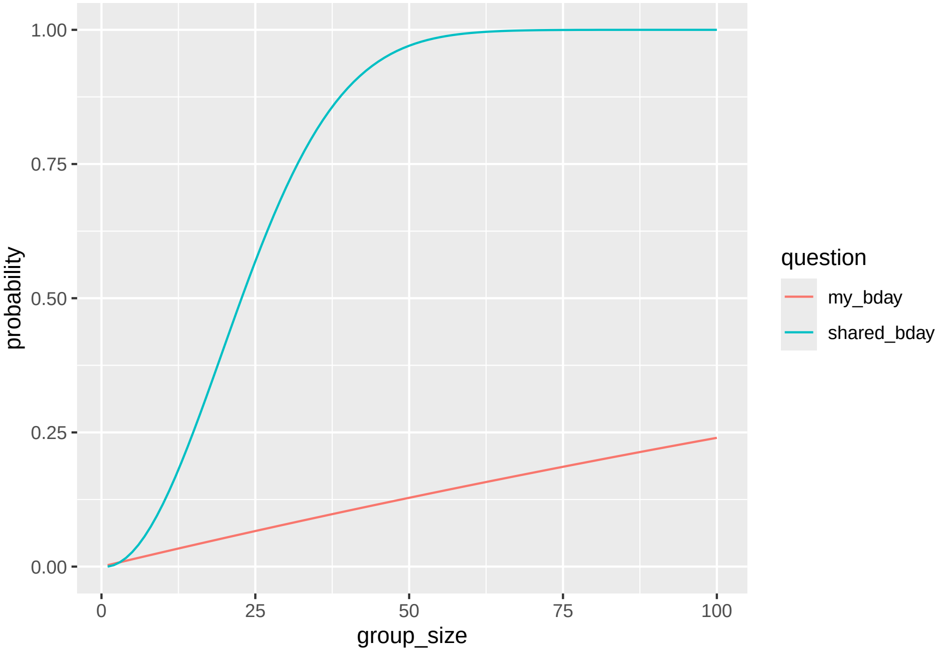
```
## [1] 0.2529013
```

Whoa! A 25 percent chance! In fact, it turns out we only need 23 people gathered in a room to have a 50 percent chance that two of them share a birthday! Also, with a group of 59 people we have a 99 percent chance of a birthday match: `prob_shared_bday(59)`=0.993.

Here's a graph of these two probabilities for values of $1 \leq n \leq 100$.

```
sh=c()
my=c()
for (i in 1:100){
  sh=c(sh,prob_shared_bday(i))
  my=c(my,prob_my_bday(i))
}
df <- data.frame(group_size=1:100,my_bday=my,shared_bday=sh)
df_long <- df%>% pivot_longer(cols=c(my_bday,shared_bday), names_to = "question",

ggplot(df_long)+
  geom_line(aes(x=group_size,y=probability,col=question))
```



B.9 Flipping Coins with Fibonacci

Let X equal the number of flips required to observe heads on consecutive flips. For instance, $X = 6$ in the flip sequence "T H T T H H". The random variable X is discrete, taking on countably infinite values 2, 3, 4, … .

The probability function for X is

$$p(x) = \frac{F_{x+1}}{2^x} \text{ for } x = 2, 3, 4, \dots,$$

where F_n is the nth Fibonacci number. (F_n is defined recursively: $F_1 = F_2 = 1$, and for $n \geq 3$, $F_n = F_{n-1} + F_{n-2}$.)

Deriving this function requires some satisfying work, which we work through in class.

We can also approximate the probability function by simulation.

```
trials = 10000
results = c() #stores result
for (i in 1:trials){
  flips = 0
  consecutive_H = 0
  while (consecutive_H<2){
    flips = flips + 1 #recording flips made in this trial
    consecutive_H = ifelse(sample(c("H","T"),1)=="H",consecutive_H+1,0)
  }
  results[i]=flips #updates results to include flips required for consec heads in t
}
```

We note that the maximum number of flips it took to get consecutive Heads in these trials was `max(results)` = 49! (that's not a factorial symbol, just me

x	2	3	4	5	6	7	8	9	10	1
Rel_freq	0.2535	0.1261	0.1235	0.0892	0.0785	0.0631	0.0508	0.0426	0.0307	0
p(x)	0.2500	0.1250	0.1250	0.0938	0.0781	0.0625	0.0508	0.0410	0.0332	0

exclaiming). So we don't list the full table of results here, just the first 10 relative frequencies, which provide estimates for the probability function:

```
table(results)[1:10]/trials
```

```
## results
##      2      3      4      5      6      7      8      9     10     11
## 0.2535 0.1261 0.1235 0.0892 0.0785 0.0631 0.0508 0.0426 0.0307 0.0254
```

These values compare closely to the actual probability values $F_{x+1}/2^x$:

B.10 Seats on an airplane

> A small airplane has 4 rows of 3 seats. 8 people board and sit randomly among the 12 seats. Then a couple boards. What is the probability they will be able to sit next to each other in the same row?

Strategy

It is convenient to use a matrix with 4 rows and 3 columns to represent the seats in the plane. Each entry in the matrix is 0 or 1 depending on whether the seat is empty or occupied.

We randomly assign 8 1s to the 12 entries in the matrix, leaving the remaining 4 spots 0, and then we check to see whether one of the rows has adjacent 0s.

First, here's one way to randomly assign 8 1s and 4 0s to 12 seats (we make a vector with 8 1s and 4 0s, and then take a random permutation of it):

```
x = sample(c(rep(1,8),rep(0,4)))
x
```

```
##  [1] 0 1 1 0 1 1 0 1 1 0 1 1
```

And here's how we can store this vector of length 12 in a 4×3 matrix:

```
m=matrix(x,byrow=TRUE,nrow=4)
m
```

```
##      [,1] [,2] [,3]
## [1,]    0    1    1
## [2,]    0    1    1
## [3,]    0    1    1
## [4,]    0    1    1
```

As the display above suggests, we can extract a value from matrix `m` by indicating its [row,column]: `m[3,2]` returns the element in row 3, column 2.

Here's one way to check for empty seats next to each other in a given row: See whether the middle entry is 0 AND the sum of the three numbers in the row is less than or equal to 1.

```
for (row in 1:4){
  result=ifelse(m[row,2]==0 & sum(m[row,])<= 1,"seats together!","nope")
   print(paste("Row",row,result))
}
```

```
## [1] "Row 1 nope"
## [1] "Row 2 nope"
## [1] "Row 3 nope"
## [1] "Row 4 nope"
```

Ok, we've got a plan! In the code below, we use s to record, in each trial, whether the couple can find seats together (s = 1) or they can't (s = 0).

Final code

```
trials = 10000
results = c()
for (i in 1:trials){
  m=matrix(sample(c(rep(1,8),rep(0,4))),byrow=TRUE,nrow=4)
  s=0
  for (r in 1:4){
    if (m[r,2]==0&sum(m[r,])<=1){
      s = 1
      break
    }
  }
  results[i]=s
}
sum(results)/trials
```

```
## [1] 0.6126
```

This probability seems higher than I first expected it would be, but we can also prove using our counting techniques -we do this in class - that the probability is, in fact, 20/33.

B.11 Drawing names for Homemades

Growing up, I enjoyed a family tradition of exchanging homemade gifts at Christmas. In part, my enjoyment stemmed from the fact that I was the youngest person, by 6 years, in our family of 6. While I would earnestly make and give to some unlucky brother or parent a clay boot I made in the kiln at school, I was likely to receive a far superior gift in return, perhaps a lightweight springboard from my Dad that was just sturdy enough to propel my little body, flipping, into a sea of beanbags in the basement, but not sturdy enough to withstand the weight of a big brother.

I also enjoyed the secrecy of the homemades: the locked doors behind which people worked, the occasional curious scrap of material left behind on a workbench, or a splash of paint in the utility room that wasn't there earlier.

Then there was the celebration of the exchange itself, the pride of giving your gift, the excitement of receiving one, and, after months of speculation, the big reveal as to who everyone drew.

We would usually draw names in late October. If anyone drew their own name, we would immediately stop the process, return all names to the hat, and try again. We kept redrawing, as necessary, until everyone drew a name that wasn't theirs. Then we were off and running, guarding the secret of our givee's identity at all costs.

In this example we pursue three questions related to the homemades:

a) In our family of 6, what is the probability we will not have to redraw.

Our approach: Create a vector called `family`, and a permutation of `family` (called `draw`). For each i, Person i in `family` draws person i in `draw`. So, we have to redraw if the two vectors ever agree at the same location.

```
family = c("Dad","Mom","Tom","John","Dave","Mike")
draw = sample(family)
matches = sum(family==draw) # number of locations at which the two vectors agree
redraw = ifelse(matches==0,"No.","Yes.")
print(paste("Redraw?",redraw))
```

```
## [1] "Redraw? Yes."
```

Let's check who each person drew to be sure (remember, our code is assuming the ith person in the `family` vector drew the ith person in the `draw` vector)

```
for (i in 1:6){
  print(paste(family[i],"drew",draw[i]))
  if (family[i]==draw[i]){
    print("Dang it! Need to redraw. Names back in the hat.")
    break}
}
```

```
## [1] "Dad drew Dad"
## [1] "Dang it! Need to redraw. Names back in the hat."
```

We now estimate the probability of not needing to redraw as follows:

```
n=6 #size of family
family = 1:n #we use numbers, to protect the innocent
trials=10000
results = c() #record for each trial a 0 for redraw, 1 for no redraw
for (i in 1:trials){
  draw = sample(family)
  results[i] = ifelse(sum(family == draw)==0,1,0)
}
sum(results)/trials
```

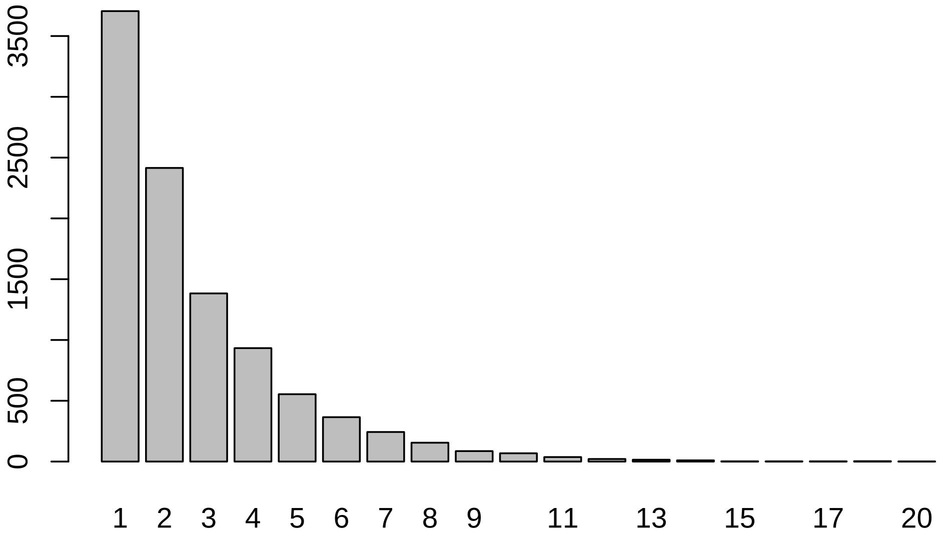
```
## [1] 0.3697
```

About a 37 percent chance of not needing to redraw.

b) What is the expected number of draws needed until we do not need to redraw?

Now in our simulation we want to redraw until we don't have to, and record how many attempts it took.

```
n=6 #size of family
family = 1:n #we use numbers, to protect the innocent
trials=10000
results = c() #record for each trial how many attempts it took to get a good draw
for (i in 1:trials){
  attempts = 0
  draw_again = TRUE
  while (draw_again){
    attempts = attempts+1
    draw = sample(family)
    draw_again = ifelse(sum(family == draw)>0,TRUE,FALSE)
  }
  results[i] = attempts
}
barplot(table(results))
```



Cripes! In 1 of the trials we needed 20 draws until no one drew their own name! Anyway, our estimate for the expected number of draws needed to get the gift-making off the mark is:

```
sum(results)/trials
```

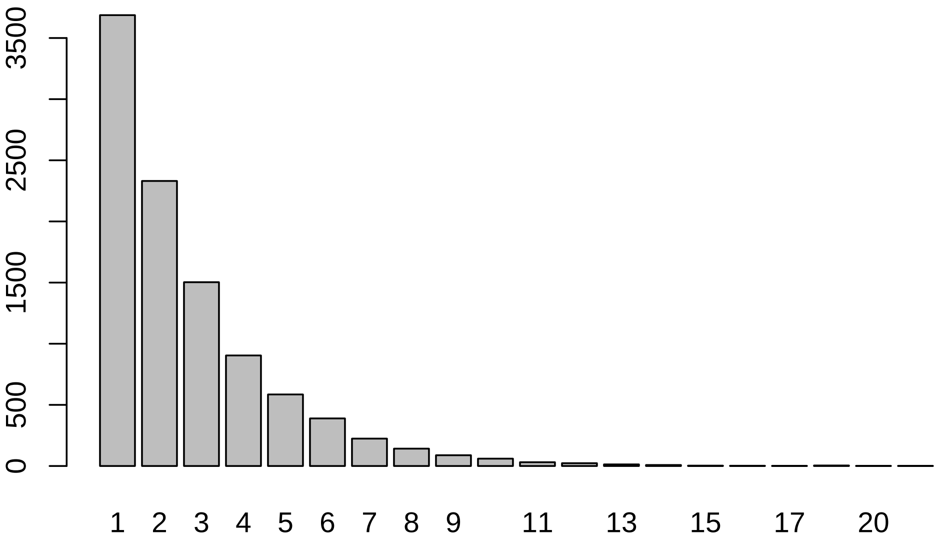
```
## [1] 2.6935
```

and if that number looks somewhat familiar, I encourage you to read the mathematical addendum at the end of this example.

c) Over time the family grew as families do until we had 16 in the gift exchange. Does the probability of needing to redraw change substantially with the larger family size?

Let's just repeat **(b)** with $n = 16$.

```
n=16 #size of family
family = 1:n #we use numbers, to protect the innocent
trials=10000
results = c() #record for each trial how many attempts it took to get a good draw
for (i in 1:trials){
  attempts = 0
  draw_again = TRUE
  while (draw_again){
    attempts = attempts+1
    draw = sample(family)
    draw_again = ifelse(sum(family == draw)>0,TRUE,FALSE)
  }
  results[i] = attempts
}
barplot(table(results))
```



Based on the trials, we estimate the probability of not needing a redraw in the larger family to be

```
sum(results==1)/trials
```

```
## [1] 0.3687
```

and our estimate for the expected number of draws to be

```
sum(results)/trials
```

```
## [1] 2.6934
```

Not much movement in the estimates, it seems.

With a function and a for loop we can streamline the process for estimating both of these values for many different family sizes.

```
hat_draw_sim <- function(n,trials){
  # inputs: n - family size; trials - number of trials in simulation
  # output: a vector with two values based on a number of trials:
       # - first value: estimated prob of redraw for family size n
       # - second value: estimated expected number of draws for family size n
  family = 1:n
  results = c() #record for each trial how many attempts it took to get a good dra
  for (i in 1:trials){
    attempts = 0
    draw_again = TRUE
    while (draw_again){
      attempts = attempts+1
      draw = sample(family)
      draw_again = ifelse(sum(family == draw)>0,TRUE,FALSE)
    }
    results[i] = attempts
  }
  return(c(sum(results==1)/trials,sum(results)/trials))
}
```

```
# run simulation for various family sizes
family_size = c(2:10,seq(15,60,5))
prob_redraw = c()
expected_draws = c()
for (i in 1:length(family_size)){
  v=hat_draw_sim(family_size[i],trials=10000)
  prob_redraw[i] = v[1]
  expected_draws[i] = v[2]
}
#table of results
kbl(data.frame(family_size,prob_redraw,expected_draws)) %>% kable_styling(full_widt
```

It seems the probability of not needing a redraw and the expected number of draws needed both converge fairly quickly to about 0.37ish and 2.7 ish, respectively.

In fact, these values ought to be closely related of one another since "drawing until our first success" sounds like a geometric distribution to me. The probability of success on any given draw somewher in the neighborhood of $p = .37$, and the expected value for a geometric distribution is $1/p$, which would be about $1/.37$, which puts us in the neighborhood of 2.7. If these numbers seem frustratingly imprecise, I'm glad, and I invite you to read the following addendum.

Mathematical addendum to the question of drawing names.

The actual expected number of draws, as $n \to \infty$, is e, and the probability of not needing to redraw approaches $1/e$ as $n \to \infty$.

In part (a) of this question, we are counting derangements of a vector. A **derangement** of a vector is a permutation of it in which no element is in its

family_size	prob_redraw	expected_draws
2	0.4988	2.0005
3	0.3380	2.9587
4	0.3636	2.7123
5	0.3704	2.7430
6	0.3649	2.7235
7	0.3736	2.7132
8	0.3764	2.7052
9	0.3665	2.7505
10	0.3626	2.7146
15	0.3775	2.7089
20	0.3647	2.7142
25	0.3743	2.6862
30	0.3694	2.7259
35	0.3649	2.7212
40	0.3796	2.6798
45	0.3668	2.7160
50	0.3700	2.6865
55	0.3669	2.7206
60	0.3659	2.7305

original position. An element of a permutation is called *fixed points* if it appears in its original position, so a derangement of a vector is a permutation of it with no fixed points. In part (a) we are estimating the probability that a permutation of the `family` vector is a derangement. There is a nice recursive formula for the number of derangements of a vector of size n, and it can be used to prove that the probability of drawing a derangement converges to $1/e$ as $n \to \infty$.

B.12 Idiot's Delight

Idiot's delight is a simple card game without strategy or game play options. Here are the rules:

At the start of the game your *hand* is empty, and the full deck is the *draw pile*. Your hand will always be an ordered list of cards.

Each turn consists of two steps.

1. *Draw one card* from the draw pile. This card becomes the first, or "top" card in your hand.
2. *Check for hand reductions.* If you have fewer than 4 cards in your hand you can make no reductions. If you have at least four cards, compare the top card with the fourth card in your hand (fourth from the top). If these cards have the same rank, remove the top four cards from your hand. If these cards have the same suit, remove the two cards in between (so, the 2nd and 3rd cards in your hand). Repeat step 2 until no reduction can be made. When no reduction can be made begin the next turn.

The game ends after you have drawn all cards from the deck and you can make no reductions. Your *game score* is the size of your hand at the end. For a regular 52 card deck you *win* if your score is less than 10.

We have two questions about this game.

a) Estimate the probability that you win.

b) Estimate how likely a player will have a reduction on each of the 52 draws that constitute a game.

Creating this game in R requires some work, beginning with the creation of a deck of cards, and continuing with the creation of some functions to manage the game play. Playing the game for a while might help the reader make sense of the code below.

```
#Build a Deck
rank = rep(c(2:10,"J","Q","K","A"),4)
suit = c(rep("clubs",13),rep("diamonds",13),rep("hearts",13),rep("spades",13))
deck = paste0(rank,"-",suit) #standard deck of cards
sample(deck,4) #display four random cards
```

```
## [1] "6-clubs"    "J-spades"   "K-clubs"    "Q-diamonds"
```

Here's code for playing the game with a lot of printed statements to increase the drama of each turn. We do not run the code here, but recommend you copy and paste it into a script in your own session and play a few times to get the feel for it.

```
#build a deck
rank = rep(c(2:10,"J","Q","K","A"),4)
suit = c(rep("clubs",13),rep("diamonds",13),rep("hearts",13),rep("spades",13))
deck = paste0(rank,"-",suit)

#game functions
rank <- function(card){ #extracts the rank from the card string
  return(unlist(strsplit(card,"-"))[1])
}
suit <- function(card){ #extracts the suit from the card string
  return(unlist(strsplit(card,"-"))[2])
}
scan <- function(hand){ #scans the hand for hand reductions
  if (length(hand) <= 3){return(hand)}
  if (rank(hand[1])==rank(hand[4])){return(hand[-(1:4)])}
  if (suit(hand[1])==suit(hand[4])){
      hand <- hand[-(2:3)]
      hand <- scan(hand)
  }
  return(hand)
}

#play game with commentary
hand = c()
```

```r
draw_pile = sample(deck)#shuffles the deck
for (k in 1:length(deck)){
  print(paste("Turn",k,"hand after draw but before scanning:"))
  hand <- c(draw_pile[k],hand) #add card to hand
  print(hand)
  hand <- scan(hand)
  print('=================')
}
print(paste("Game score:",length(hand)))
```

Here's a play game function that plays the game (without commentary) and returns the size of your hand after each turn.

```r
play_game <- function(){
  #build a deck
  rank = rep(c(2:10,"J","Q","K","A"),4)
  suit = c(rep("clubs",13),rep("diamonds",13),rep("hearts",13),rep("spades",13))
  deck = paste0(rank,"-",suit)

  #game functions
  rank <- function(card){ #extracts the rank from the card string
    return(unlist(strsplit(card,"-"))[1])
    }
  suit <- function(card){ #extracts the suit from the card string
    return(unlist(strsplit(card,"-"))[2])
    }
  scan <- function(hand){ #scans the hand for hand reductions
    if (length(hand) <= 3){return(hand)}
    if (rank(hand[1])==rank(hand[4])){return(hand[-(1:4)])}
    if (suit(hand[1])==suit(hand[4])){
      hand <- hand[-(2:3)]
      hand <- scan(hand)
      }
    return(hand)
    }
  #play game without comment, returning vector of hand_sizes
  hand = c()
  draw_pile = sample(deck)#shuffles the deck
  hand_size = c()
  for (k in 1:length(deck)){
    hand <- c(draw_pile[k],hand) #add card to hand
    hand <- scan(hand) #look for reductions
    hand_size[k]=length(hand) #record length of hand after turn k
    }
  return(hand_size)
}
```

Now we do our simulations. The result of running the `play_game()` function is a vector of length 52, indicating the size of our hand after each of our 52 draws. With the trials below, we concatenate these vectors into one long vector `results` (of length `52*trials`) which we then convert into matrix with 52 columns in such a way that each row of the matrix represents the hand sizes for a single game.

```
trials = 1000
results = c()
for (i in 1:trials){
  v = play_game()
  results = c(results,v)
}
M = matrix(results,byrow = TRUE,ncol = 52)
```

The last column of this matrix records each of our final game scores, so the fraction of those values less than 10 answer part (a) for us.

Our estimated probability of winning in a game of Idiot's delight:

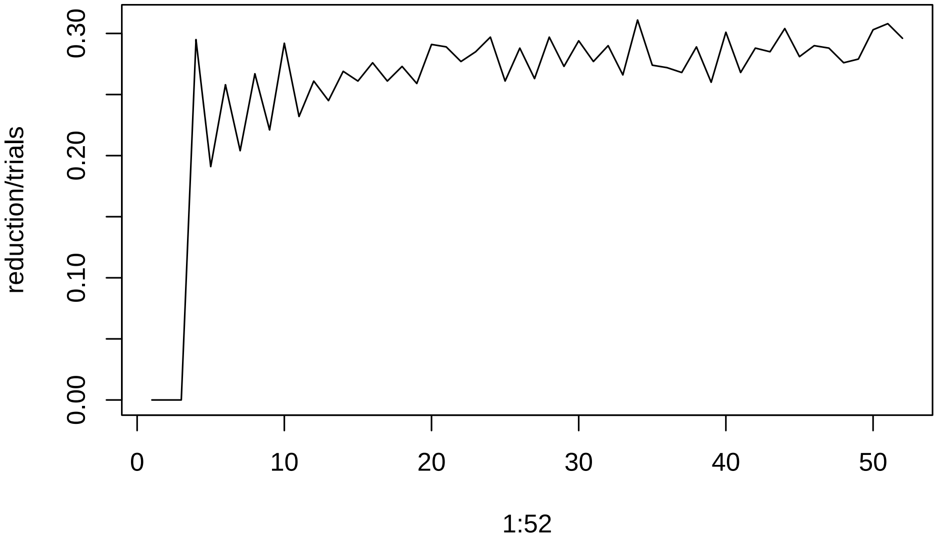
```
sum(M[,52]<10)/trials
```

```
## [1] 0.32
```

For (b), the likelihood of having a reduction on turn k (for each $k = 1, 2, \dots 52$), we can compare our hand size after turn k with our hand size after turn $k - 1$.

We know turns $k = 1, 2, 3$ can have no reductions, but starting with $k = 4$ we might:

```
reduction = c(0,0,0) # start by recording 0 reductions for turns 1,2,3
for (k in 4:52){
  reduction[k] = sum(M[,k]<M[,k-1])
}
plot(x=1:52,y=reduction/trials,type='l')
```



Starting with turn 4 it looks like the likelihood of having a reduction stays roughly constant, at about 0.27.

Appendix C

Discrete Random Variables in R

Here we investigate in R the common, named discrete random variables we encounter in MATH 340:

- binomial | `binom`
- geometric | `geom`
- negative binomial | `nbinom`
- hypergeometric | `hyper`
- Poisson | `pois`

We use four commands to work with the named distributions. For a distribution named `___`:

- `d___(x,...)` | probability function, $p(x)$
- `p___(q,...)` | Cumulative probability, $P(X \leq q)$
- `q___(p,...)` | Quantiles, finds x such that $P(X \leq x) = p$
- `r___(n,...)` | Random sample of size n from the distribution

We also discuss below how to build and analyze homemade discrete random variables in R.

C.1 Binomial `binom`

The Scene

Recall, $X \sim$ `binom`(n, p) means X counts the number of successes in n independent, identical Bernoulli trials, when probability of success on any given trial is p.

The space of X is $x = 0, 1, \dots, n$.

Probability function
For $x = 0, 1, \dots, n$,

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

The binomial distribution in R

`dbinom()` - probability function

`dbinom(x,n,p)` returns the probability $P(X = x)$ for $X \sim \texttt{binom}(n, p)$.

For instance, `dbinom(2,5,.3)` returns

$$\binom{5}{2}(.3)^2(.7)^3.$$

```
dbinom(2,5,.3)
```

```
## [1] 0.3087
```

As a check:

```
choose(5,2)*(.3)^2*(.7)^3
```

```
## [1] 0.3087
```

`pbinom()` - cumulative probability

`pbinom(q,n,p)`returns the cumulative probability $P(X \le q)$ for $X \sim \text{binom}(n, p)$:

$$\sum_{x=0}^{q} p(x) = \sum_{x=0}^{q} \binom{n}{x} p^x (1-p)^{n-x}.$$

So `pbinom(2,5,.3)` returns $P(X \le 2)$ when X is $\texttt{binom}(5, .3)$:

```
pbinom(2,5,.3)
```

```
## [1] 0.83692
```

As a check:

```
dbinom(0,5,.3)+dbinom(1,5,.3)+dbinom(2,5,.3)
```

```
## [1] 0.83692
```

`qbinom()` - quantiles

Recall the definition of quantile (9.4): If $0 < p < 1$, the pth quantile of X, denoted ϕ_p, is the smallest value such that $P(X \le \phi_p) \ge p$. In other words, the value ϕ_p marks the smallest value below which one finds 100*p percent of the distribution of X.

`qbinom(q,n,p)` returns the quantile ϕ_q for $X \sim \texttt{binom}(n, p)$

For instance, what value marks the 95th percentile of the binom$(100, .5)$ distribution?

```
qbinom(.95,100,.5)
```

```
## [1] 58
```

So, if you flip a fair coin 100 times and count how many heads you get, about 95% of the time you would flip less than or equal to 58 heads.

We can check this:

```
pbinom(58,100,.5)
```

```
## [1] 0.955687
```

rbinom() - sampling

> rbinom(10,20,.4) will generate a vector that stores a random sample of size 10 drawn from a binom$(20, .4)$ distribution.

```
rbinom(10,20,.4)
```

```
##  [1]  7  8  9  5  6  9 11 11 11  9
```

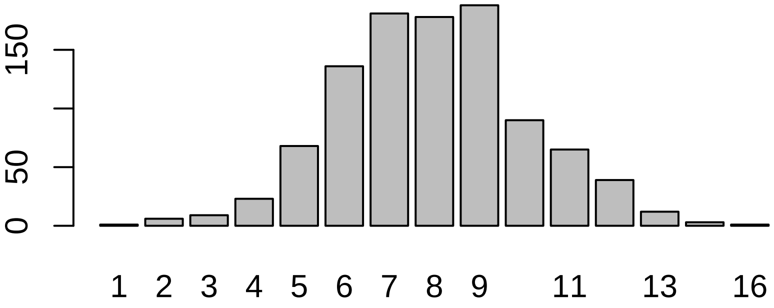
We can use r___ to run simulations, and to visualize the shape of a distribution.

Two useful commands for summarizing data: table() presents the frequency table for the sample, and barplot(table()) is a quick way to visualize this frequency table.

```
sim_data = rbinom(1000,20,.4) # sample of size 1000 from binom(20,.4).
table(sim_data)
```

```
## sim_data
##   1   2   3   4   5   6   7   8   9  10  11  12  13  15  16 
##   1   6   9  23  68 136 181 178 188  90  65  39  12   3   1
```

```
barplot(table(sim_data))
```



C.2 Geometric `geom`

The Scene
Let the random variable X denote the number of identical, independent Bernoulli trials (with probability of success p, probability of failure $q = 1 - p$) up to and including the first success. Then X is called a *geometric random variable* with parameter p.

Notation
X is `geom`(p).

The space of X is $x = 1, 2, \ldots$

Probability function
For $x = 1, 2, 3, \ldots,$

$$p(x) = q^{x-1}p.$$

NOTE: The geometric distribution in R counts failures, not total trials.

In R `geom` counts the number of failures until the first success, not the total number of trials up to and including the first success.

As with the `binom` distribution, we can use the `d___`, `p___`, `q___`, and `r___` commands to determine probability for particular values of x, cumulative probabilities, quantiles, and random samples, respectively.

> `dgeom(4,.3)` gives the probability of seeing 4 failures before the first success in a Bernoulli trial in which $p = .3$

```
dgeom(4,.3)
```

```
## [1] 0.07203
```

> `pgeom(4,.3)` gives the probability of seeing 4 or fewer failures before the first success in a sequence of Bernoulli trials in which $p = .3$

```
pgeom(4,.3)
```

```
## [1] 0.83193
```

and the following line gives the probability of seeing more than 4 failures prior to the first success:

```
1-pgeom(4,.3)
```

```
## [1] 0.16807
```

Example C.1.

> Roll a fair 6-sided die until a four comes up, and let X denote the rolls up needed to see that first four. Repeat this game 10,000 times, and plot the

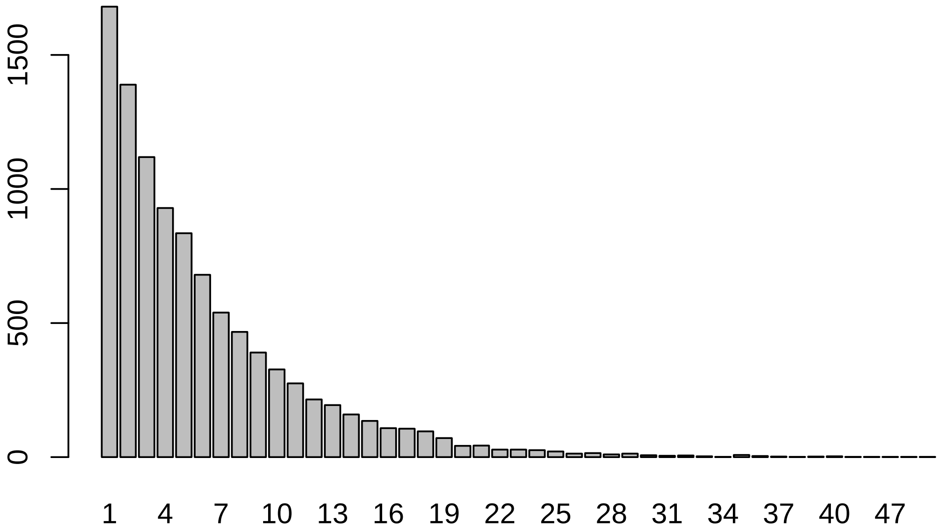
frequency distribution for X.

Strategy:

1. Note that this game is a Bernoulli trial, where "success" means rolling a 4 and "failure" means not rolling a four. So $p = 1/6$, and $q = 5/6$.

2. Take a random sample of size 10000 from the `geom` distribution in R with the `rgeom()` method (which records the number of failures, not the number of trials).
3. Add one to each value in the sample to get the number of trials.
4. barplot the table!

```
results=rgeom(10000,1/6)+1
barplot(table(results))
```



OMG notice from the bar plot that one depressing game required 57 rolls to see my first 4.

C.3 Negative Binomial `nbinom`

The Scene
Again, we consider a sequence of Bernoulli trials (probability of success is p, probability of failure is $q = 1 - p$).

We let X denote the number of trials in the sequence up to and including the rth success, where $r \geq 1$ is a positive integer. Then X is called a *negative binomial random variable* with parameters p and r.

Notation: X is `nbinom`(r, p)

The space of X is $x = r, r + 1, r + 2, ...$

Probability function

For $x = r, r+1, r+2, \dots,$

$$p(x) = \binom{x-1}{r-1} p^r q^{x-r}.$$

Example C.2. A study indicates that an exploratory oil well drilled in a particular region should strike oil with probability 0.2. Find the probability that the third oil strike comes on the 10th well drilled.

Here, if X equals the number of wells drilled until the company gets its third strike, then X is Nb(3,.2), and the answer to this question is $P(X = 10)$ which is

$$P(X = 10) = \binom{9}{2} 0.2^3 .8^7.$$

```
round(choose(9,2)*.2^3*.8^7,4)
```

```
## [1] 0.0604
```

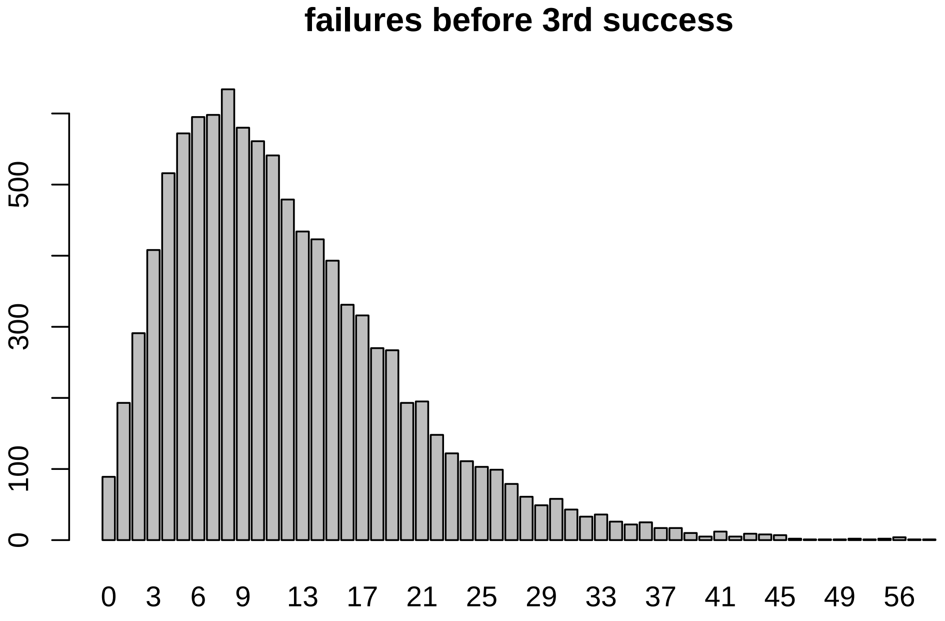
In R this distribution is accessed using `nbinom`, but this distribution, like `geom`, focuses on the number of failures, not total trials. If we want to know the probability that our third success occurs on the 10th trial, this is equivalent to the probability of having 10-3 = 7 failures before getting our third success, which can be computed in R as

```
dnbinom(7,3,.2) # 7 failures to get 3rd success, p = .2
```

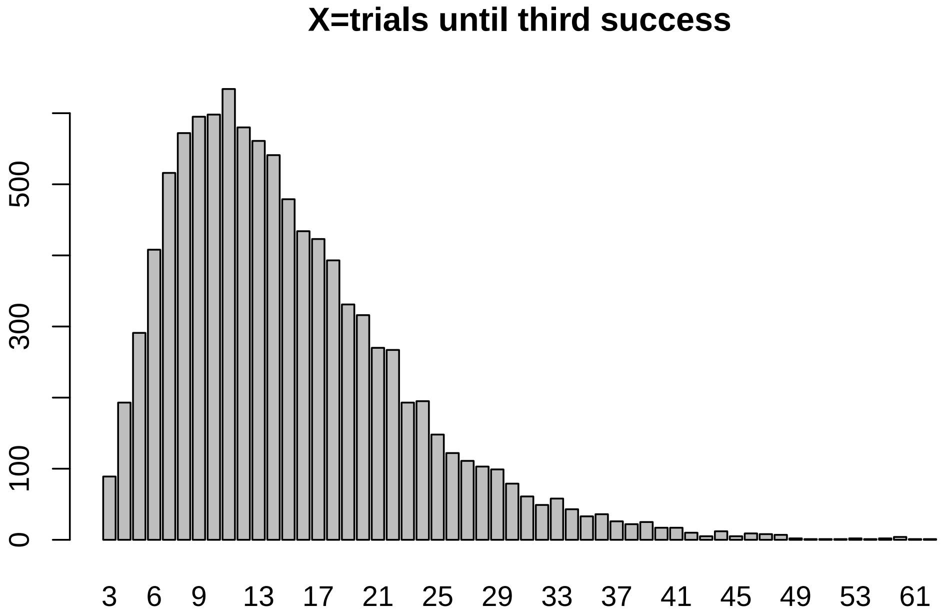
```
## [1] 0.06039798
```

Visualizing $X \sim$ `nbinom`$(3, .2)$

```
r = 3 #going until we get 3rd success
p = .2 #probability of success on any given Bernoulli trial
trials = 10000 #trials in this simulation
failure_count = rnbinom(trials,r,p)
barplot(table(failure_count),main="failures before 3rd success")
```



```
trial_count=failure_count+r
barplot(table(trial_count),main="X=trials until third success")
```



C.4 Hypergeometric `hyper`

The Scene
A finite population has N elements, each of which possesses one of two possible characteristics. Say we have a jar of N marbles, each is either red or black. Let's say m of them are red and n of them are black (so $m + n = N$). We draw a sample of size k, and let X denote the number of red marbles in the jar.

Then X is called a *hypergeometric random variable* with parameters m, n, and k.

Notation: X is `hyper`(m, n, k)

The space of X is $x = 0, 1, 2, \dots, k$ subject to the restriction that $x \leq m$ and $k - x \leq n$.

Probability function
The probability function is

$$p(x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}.$$

In R Use `hyper`.

Example C.3.

A group of 6 seals and 4 pelicans hang at the beach, and they select a random subset of size 5 to play beach volleyball. Let X = the number of pelicans chosen.
Here, X is hypergeometric with parameters $m = 4$ (4 pelicans), $n = 6$ (6

seals) and $k = 5$ (sample size).
The probability that $X = 2$ is

```
choose(4,2)*choose(6,3)/choose(10,5)
```

```
## [1] 0.4761905
```

We can also use the built in command `dhyper(x,m,n,k)`

```
dhyper(x=2,m=4,n=6,k=5)
```

```
## [1] 0.4761905
```

Example C.4 (Good Potatoes Bad Potatoes in R)**.**

A truck has 500 potatoes, 50 of which are bad, the rest are good. We sample 10. What is the probability that more than 3 are bad?

If X equals the number of bad potatoes in the sample, then X is hypergeometric with parameters $m = 50$, $n = 450$, and $k = 10$. So

$$P(X > 3) = 1 - P(X \leq 3)$$

which can be calculated with the cumulative probability command `phyper`:

```
1-phyper(3,50,450,10)
```

```
## [1] 0.01186118
```

C.5 Poisson `pois`

The Scene
The Poisson probability distribution can provide a good model for the number of occurrences X of a rare event in time, space, or some other unit of measure. A Poisson random variable X has one parameter, λ, which is the average number of occurrences of the rare event in the indicated time (or space, etc.)

Notation: X is `Poisson`(λ).

The space of X is $x = 0, 1, 2, \ldots,$ (countably infinite!)

Probability function
The probability function is

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

In R use `pois`.

Example C.5.

Suppose X is Poisson(5). Determine $P(X \geq 10)$.

Note: $P(X \geq 10) = 1 - P(X < 10) = 1 - P(X \leq 9)$. So, using `ppois()` we have

```
1-ppois(9,5)
```

```
## [1] 0.03182806
```

Example C.6.

> The number X of typos on a page in a textbook follows a Poisson distribution with an average number of 2 typos per page. (a) If you pick a page at random, what is the probability it contains 0 typos? (b) According to this model, 99% of the pages have no more than how many typos?

(a)

```
dpois(0,2)
```

```
## [1] 0.1353353
```

(b)

```
qpois(.99,2)
```

```
## [1] 6
```

Example C.7 (Rutherford/Geiger Data)**.**

> In a paper published in 1910 entitled “The Probability Variations in the Distribution of α-particles”, Rutherford and Geiger reported data that counted the number of “scintillations” in 72 second intervals caused by radioactive decay of a quantity of the element polonium.

Here are the data:

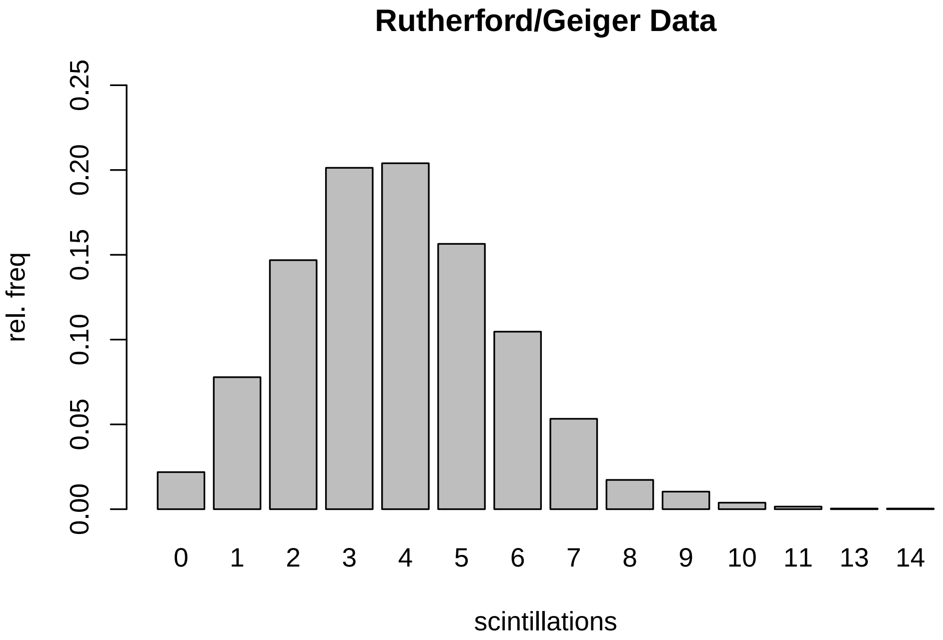
```
results=rep(0:14,c(57,203,383,525,532,408,273,139,45,27,10,4,0,1,1))
trials=length(results)
table(results)
```

```
## results
##   0   1   2   3   4   5   6   7   8   9  10  11  13  14
##  57 203 383 525 532 408 273 139  45  27  10   4   1   1
```

```
barplot(table(results)/trials,
        ylim=c(0,.25),
        ylab="rel. freq",
        xlab="scintillations",
        main="Rutherford/Geiger Data")
```

Table C.1: Fitting data with a Poisson distribution

x	rel_freq	pois_prob
0	0.0219	0.0209
1	0.0778	0.0807
2	0.1469	0.1562
3	0.2013	0.2015
4	0.2040	0.1949
5	0.1564	0.1509
6	0.1047	0.0973
7	0.0533	0.0538
8	0.0173	0.0260
9	0.0104	0.0112
10	0.0038	0.0043
11	0.0015	0.0015
12	0.0000	0.0005
13	0.0004	0.0001
14	0.0004	0.0000



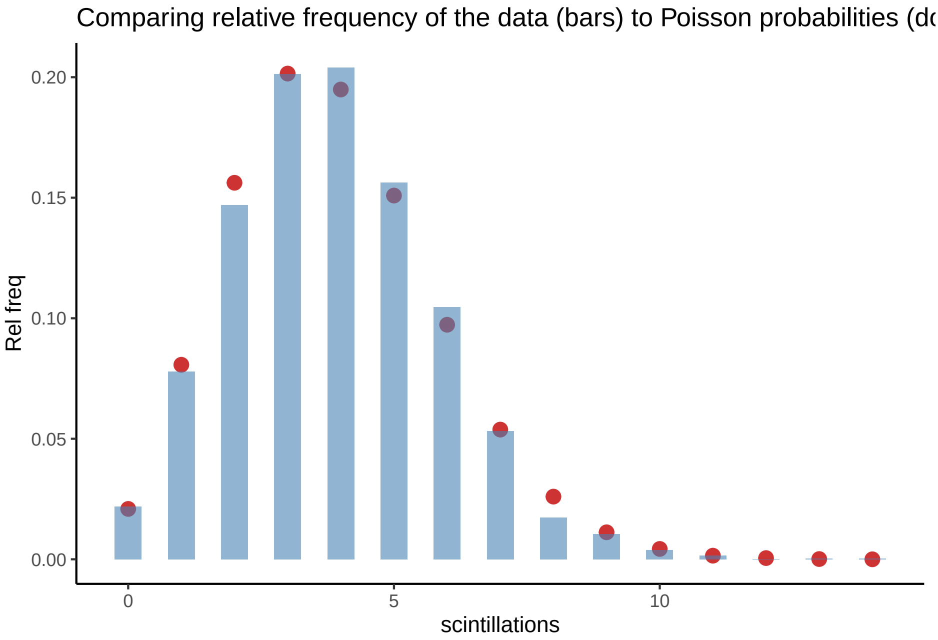
Here's the mean of the data (which gives average # of scintillations in 72 seconds):

```r
mean(results)
```

```
## [1] 3.871549
```

Let's compare the observed relative frequencies to the theoretical probabilities associated with a `Poisson`(3.87) distribution:

```r
ggplot(df)+
  geom_point(aes(x,pois_prob),col="brown3",size=3)+
  geom_col(aes(x,rel_freq),fill="steelblue",alpha=.6, width=.5)+
  ylab("Rel freq")+
  xlab("scintillations")+
  ggtitle("Comparing relative frequency of the data (bars) to Poisson probabilities
  theme_classic()
```



C.6 Homemade Discrete Random Variables

Let's say a discrete random variable X has finite sample space and known probability function $p(x)$. We often display this type of probability model via a table:

x	5	6	7	8	9
$p(x)$	0.1	0.1	0.3	0.4	0.1

We can input this model into an R session by defining two vectors:

```
X = c(5,6,7,8,9)
Px = c(0.1,0.1,0.3,0.4,0.1)
```

We can check in R that the two conditions for a *valid* probability have been met by this assignment:

- Each probability is non-negative: `Px >= 0` = TRUE, TRUE, TRUE, TRUE, TRUE
- The probabilities add to 1: `sum(Px)`= 1

Expected Value of X

Recall if X is a discrete random variable with probability function $p(x)$, then the expected value of X is

$$E(X) = \sum_{\text{all } x} x \cdot p(x)$$

Having defined vectors X and Px in R, we calculate $E(X)$ by running

```
sum(X*Px)
```

```
## [1] 7.3
```

Note: For those who have taken vector calculus `sum(v*w)` returns the dot product of `v` and `w`, aka the inner product. R has an alternative command for this dot product, which is `v %*% w`. So, `sum(v*w)` and `v %*% w` do the same thing, but I prefer the first option to remind me that the expected value is obtained as a sum over all x of some things.

Variance of X

Recall the variance of X is

$$V(X) = E[(X - \mu)^2],$$

where $\mu = E(X)$. Alternatively, the variance can be computed via

$$V(X) = E(X^2) - \mu^2.$$

So we can compute the variance of X in R as follows:

```
mu=sum(X*Px)
Vx=sum((X-mu)^2*Px)
print(Vx)
```

```
## [1] 1.21
```

Or, alternatively, as follows:

```
mu=sum(X*Px)
Vx=sum(X^2*Px)-mu^2
print(Vx)
```

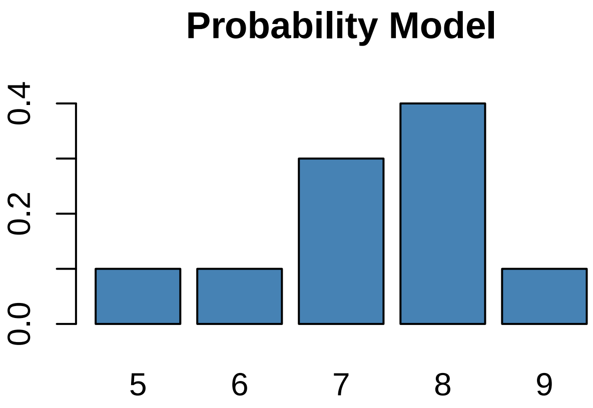
```
## [1] 1.21
```

Distribution Plots

R can offer some quick visualizations of probability distributions.

The following code will give the shape of the probability distribution (with a splash of color and plot title:)

```
barplot(Px,names.arg=X, col="steelblue", main="Probability Model")
```



Sampling

The following code draws a sample of size 10 from our distribution using the weighted probabilities assigned by the probability function:

```
sample(X,10,replace=TRUE,prob=Px)
```

```
##  [1] 7 5 8 8 8 8 9 7 8 8
```

If we take a large sample, and make a relative frequency table of the results, it should be close to the probability table:

```
round(table(sample(X,10000,replace=TRUE,Px))/trials,3)
```

```
##
##     5     6     7     8     9
## 0.383 0.373 1.163 1.526 0.389
```

C.6.1 Discrete Uniform Distribution

Definition C.1. If X is a finite set with size $|X| = n$. The probability distribution defined by

$$p(x) = \frac{1}{n}$$

for all $x \in X$ is called **uniform**.

In a uniform distribution, we will find over a large number of trials that each name comes up with about the same frequency.

Example C.8.

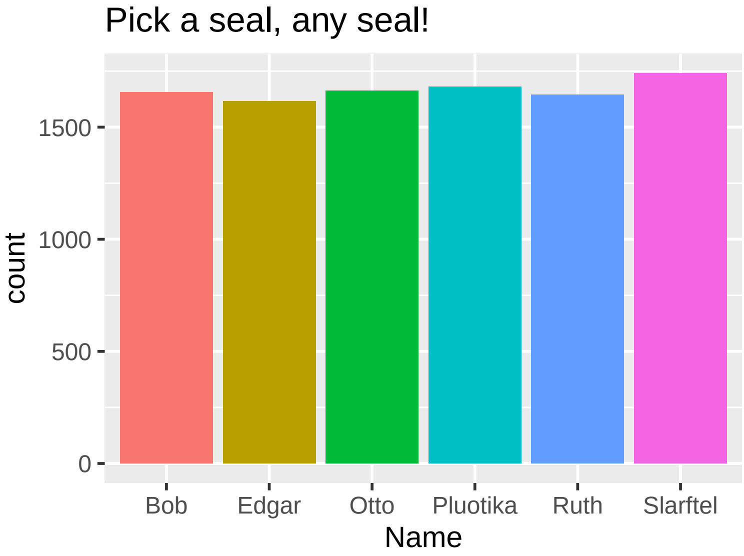
> Pick a random seal from the famous Eddington family: Otto, Ruth, Pluotika, Slarftel, Edgar and Bob.

To simulate the process of picking one seal at random from the family, a large number of times, we sample 1 element with replacement, a large number of times.

```
family=c("Bob", "Edgar", "Pluotika", "Otto", "Ruth", "Slarftel")
results=sample(family,10000,replace=TRUE)
```

The resulting frequency plot should look uniform:

```
ggplot(data.frame(results))+
  geom_bar(aes(x=results,fill=results))+xlab("Name")+ggtitle("Pick a seal, any se
```



Way to go Slarftel, you over achiever!

Appendix D

Continuous Random Variables in R

Here we investigate in R the common, named continuous random variables we encounter in MATH 340:

- Uniform probability distribution | `unif`
- Normal probability distribution | `norm`
- Exponential probability distribution | `exp`
- Gamma probability distribution | `gamma`
- Chi-square probability distribution | `chisq`
- Beta probability distribution | `beta`

For each of these distributions we may use the 4 associated commands we used in the discrete case:

- `d___()` gives the density function
- `p___()` gives cumulative probability
- `q___()` gives quantiles
- `r___()` gives random samples

We also discuss below how to build and analyze homemade continuous random variables in R.

D.1 Uniform Distribution

The uniform distribution is so very useful, it deserves top-billing here. With it we can generate random numbers, and from it we can build other interesting distributions.

A uniform random variable X over the interval $[a, b]$ has density function

$$f(x) = \frac{1}{b-a}, \quad \text{for all} \quad a \leq x \leq b.$$

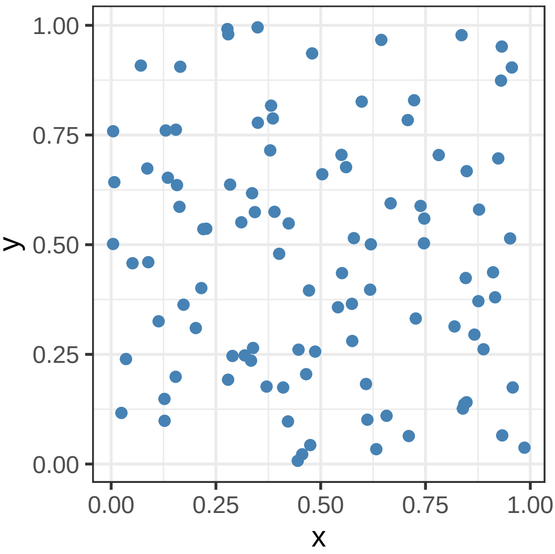
Picking random numbers

```
runif(10,0,1) #pick 10 random numbers between 0 and 1.
```

```
## [1] 0.40278356 0.42582576 0.60715240 0.61041372 0.01538657 0.45759926
## [7] 0.21881637 0.34626135 0.65106480 0.67295856
```

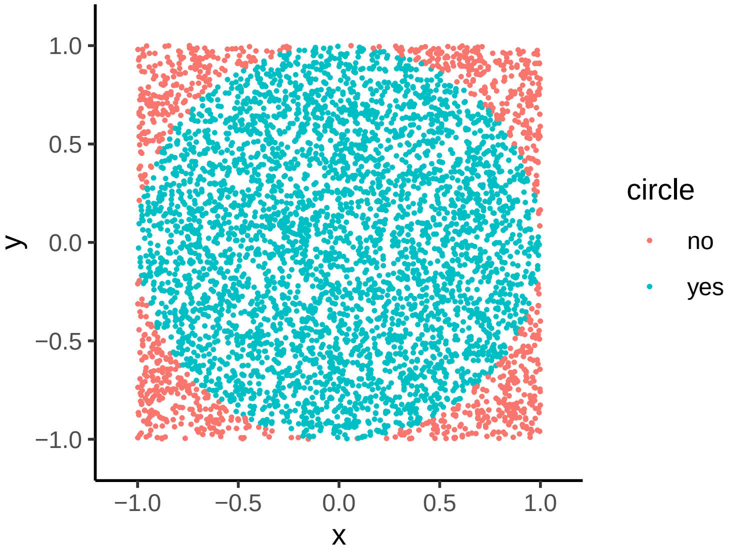
Picking random points in the unit

```
ggplot(data.frame(x=runif(100,0,1),
                  y=runif(100,0,1)))+
  geom_point(aes(x,y),col="steelblue")+
  theme_bw()
```



Estimate the value of

```
points=5000
df <- data.frame(x=runif(points,-1,1),
                  y=runif(points,-1,1))
df$circle <- ifelse(sqrt(df$x^2+df$y^2)<1,"yes","no")
ggplot(df)+
  geom_point(aes(x,y,col=circle),size=.3)+
  xlim(c(-1.1,1.1))+ylim(c(-1.1,1.1))+
  theme_classic()
```



The area of the square is 2*2 $= 4$. The area of the circle is $\pi(1)^2 = \pi$. So the ratio

$$(\text{area of circle})/(\text{area of square}) = \pi/4,$$

and we can estimate π as follows:

$$\pi \approx 4 \cdot \frac{\text{points in circle}}{\text{total points}}$$

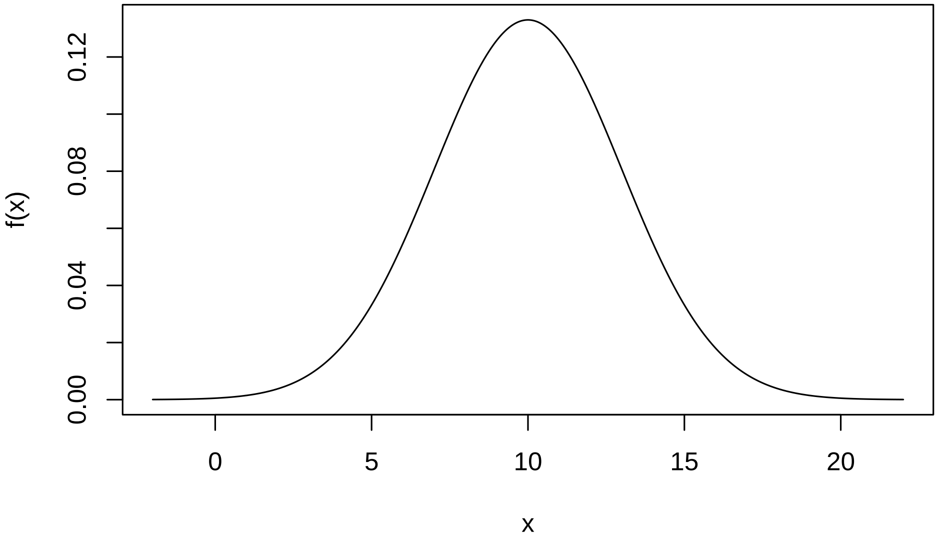
```
4*sum(df$circle=="yes")/points # our estimate of pi
```

```
## [1] 3.172
```

D.2 Normal Distribution norm

Thanks to the Central Limit Theorem this distribution has a central role in statistics.

```
mu=10; sigma=3
x=seq(mu-4*sigma,mu+4*sigma,by=.01)
plot(x,dnorm(x,mu,sigma),type="l",ylab="f(x)")
```



Example D.1.

Suppose newborn birthweights are normally distributed with mean 7.8 pounds and standard deviation 0.95 pounds.
a) What proportion of newborns weight more than 10 pounds?
b) What proportion of newboard weigh between 7 and 9 pounds? b) Find the birth weight that marks the bottom 1% of all birthweights.

```
# part (a)
1-pnorm(10,mean=7.8,sd=0.95)
```

```
## [1] 0.01028488
```

```
# part (b)
pnorm(9,mean=7.8,sd=0.95)-pnorm(7,mean=7.8,sd=0.95)
```

```
## [1] 0.6968693
```

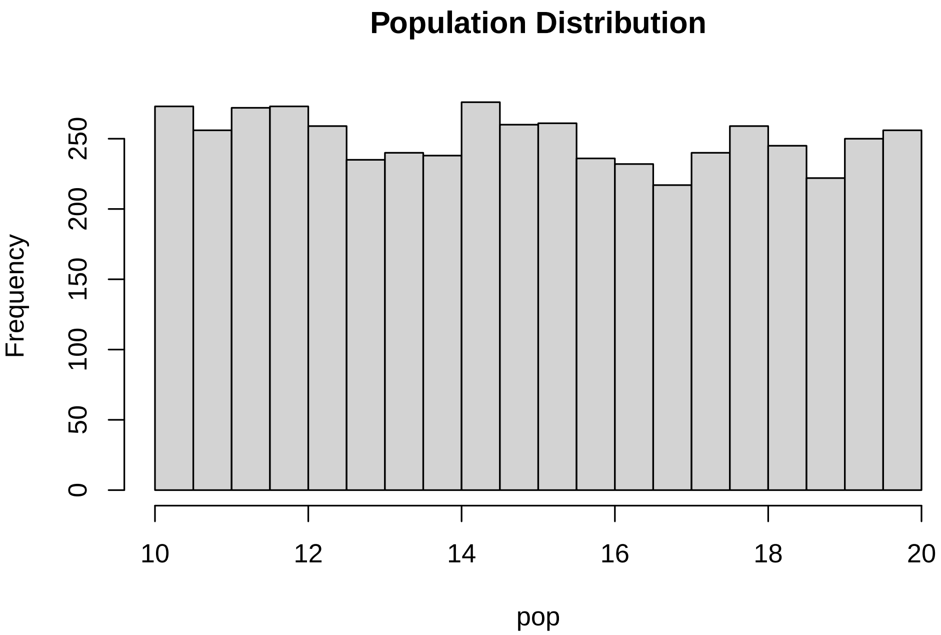
```
# part (c)
qnorm(.01,mean=7.8, sd=0.95)
```

```
## [1] 5.58997
```

Sampling Distribution of a sample mean

Suppose we have a population of 5000 random numbers between 10 and 20, which should have a uniform looking frequency distribution:

```
pop=runif(5000,10,20)
hist(pop,breaks=20, main="Population Distribution")
```



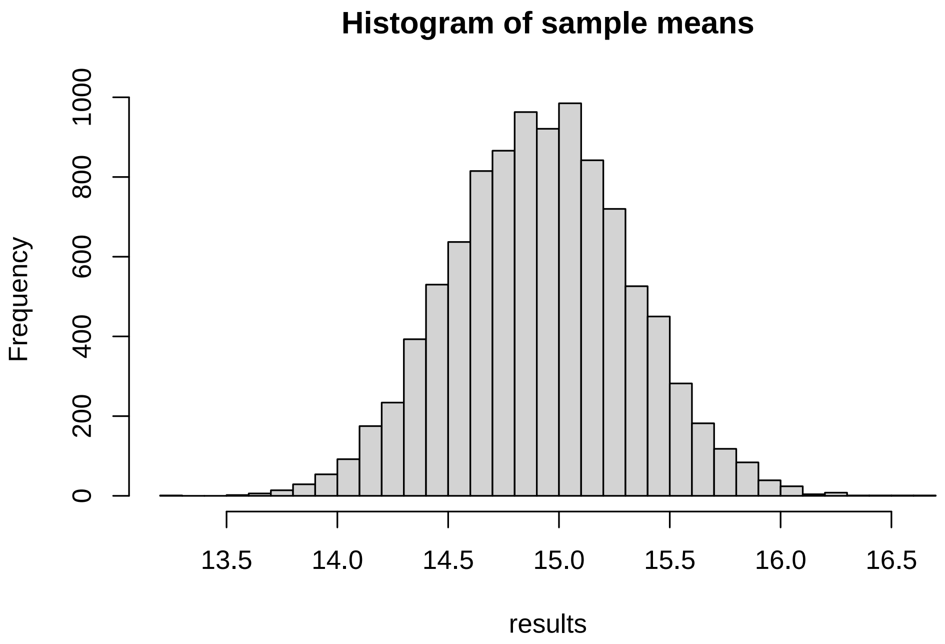
Now suppose we draw a sample of size 50 from this population, and compute the sample mean of these 50 values:

```
mean(sample(pop,50))
```

```
## [1] 14.85187
```

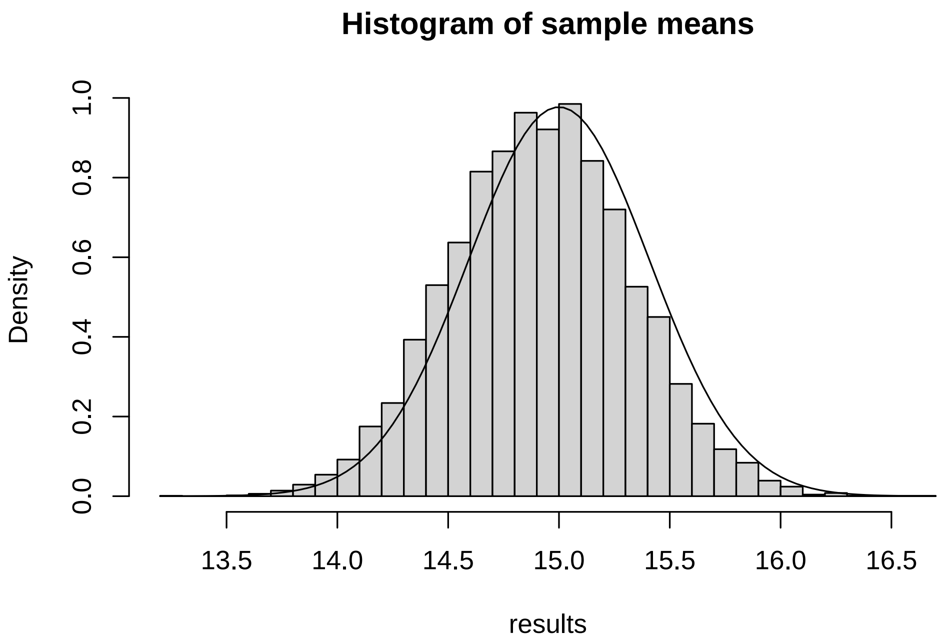
Now let's repeat this process for 10000 trials, and look at the distribution of the 10000 sample means:

```
trials=10000
results=c()
for (i in 1:trials){
  results=c(results,mean(sample(pop,50)))
}
hist(results,breaks=25, main="Histogram of sample means")
```



Look Normal?

```
x=seq(13.5,16.5,by=.05)
hist(results, main="Histogram of sample means",freq=FALSE,breaks=29)
curve(dnorm(x,15,10/sqrt(12)/sqrt(50)),add=TRUE)
```



D.3 Exponential Distribution exp

An exponential random variable X with parameter β has pdf

$$f(x) = \frac{1}{\beta} e^{-x/\beta} \quad \text{for} \quad x > 0$$

The **mean** of this distribution is $E(X) = \beta$ and the **rate** associated to this distribution is $1/\beta$. **In R**, we specify the exponential parameter by entering the rate $1/\beta$, not β itself.

Suppose X is Exp(b). **In R**, $P(X \leq q)$ is given by `pexp(q,1/b)`

Example D.2. The life of a lightbulb is exponentially distributed with mean 120 hours.

a) What is the probability that the lightbulb lasts more than 200 hours?

Here X is exponential with parameter $\beta = 120$. The rate associated with this distribution is 1/120, so $P(X > 200)$ can be computed with

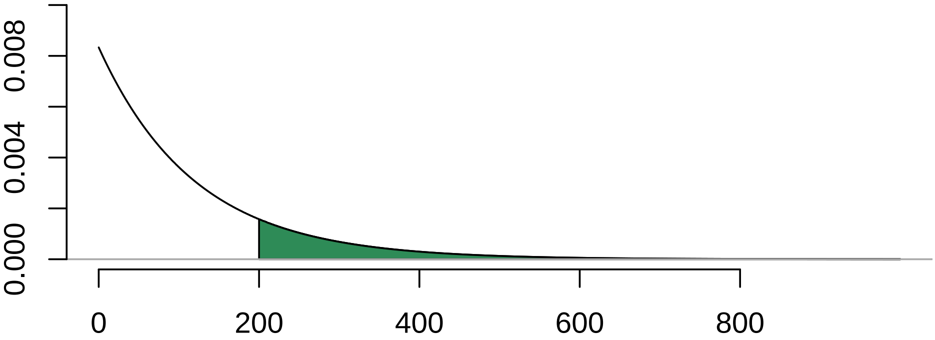
```
1-pexp(200,rate=1/120)
```

```
## [1] 0.1888756
```

As a reminder, this probability corresponds to the integral

$$\int_{200}^{\infty} \frac{1}{120} e^{-x/120}\, dx$$

which corresponds to the shaded area below



b) What proportion of lightbulbs last fewer than 5 hours?

```
pexp(5,1/120)
```

```
## [1] 0.04081054
```

c) Find the 5th percentile for this distribution.

```
qexp(.05,1/120)
```

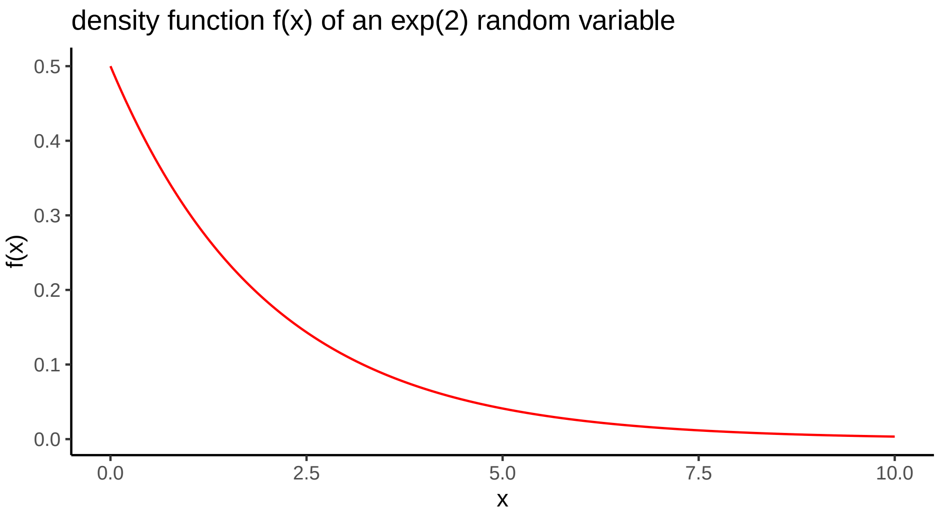
```
## [1] 6.155195
```

So, 5% of light bulbs last less than 6.16 hours.

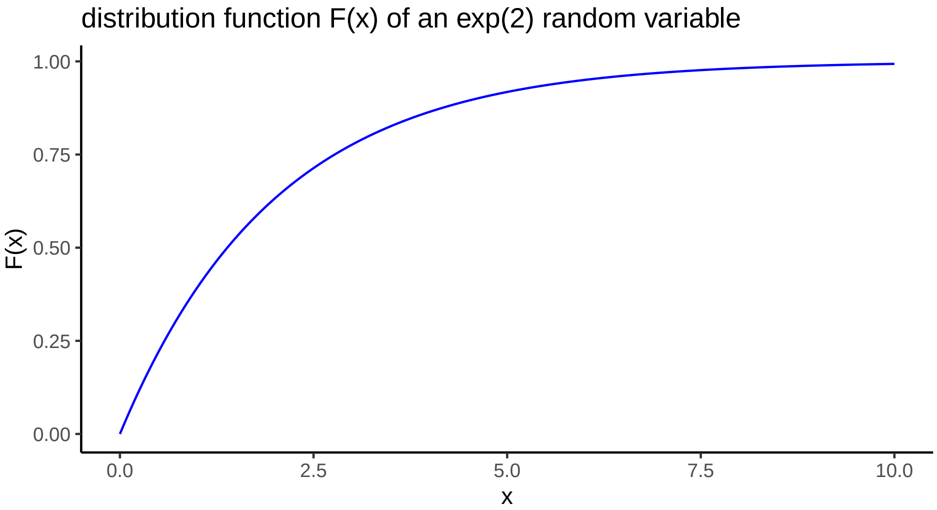
Example D.3.

Suppose X is an exponential random variable with parameter $\beta = 2$. Sketch the density function $f(x)$ as well as the distribution function $F(x)$.

The density function is $f(x) = \frac{1}{2}e^{-x/2}$ for $x > 0$, and we can sketch it by plotting an x vector of many inputs between, say, 0 and 10, and the corresponding values of `dexp()`:



The distribution function, which gives cumulative probability is found by plotting `pexp()`:


distribution function F(x) of an exp(2) random variable

A Memoryless distribution

Along with the geometric distribution, the exponential distribution is *memoryless* in this sense: For any $t, s > 0$,

$$P(X > t + s \mid X > s) = P(X > t).$$

For the geometric distribution we can interpret the above as follows: the probability of waiting more than t trials to see the first success is the same as waiting more than t additional trials after not seeing a success in the first s trials.

For the "lifetime of a light-bulb interpretation" of the exponential distribution: However long the light bulb has already lasted, the probability that the light-bulb lasts at least t more hours is the same.

We can estimate both $P(X > t)$ and $P(X > t + s \mid X > s)$ by checking a large random sample from an exponential distribution.

```
trials=10000
x=rexp(trials,rate=1/5)
s=2; t=3
p1=sum(x > t)/trials #P(X > t)
p2=sum(x[which(x > s)]>s+t)/sum(x>s) #P(X>t+s | X > s)
print(paste("Estimate for P(X>t):",round(p1,3)))
```

```
## [1] "Estimate for P(X>t): 0.549"
```

```
print(paste("Estimate for P(X>t+s | X>s):",round(p2,3)))
```

```
## [1] "Estimate for P(X>t+s | X>s): 0.558"
```

D.4 Gamma Distribution gamma

Recall, the gamma probability distribution, `gamma`(α, β) is a family of skewed right distributions. The parameter α is sometimes called the **shape** parameter, β

is called the **scale** parameter, and its reciprocal $1/\beta$ is called the **rate**. Figure 10.2 plots 3 different gamma density functions. In R we refer to a gamma distribution in our `p_`,`q_`,`d_`, and `r_` functions via the shape parameter α and either the rate $1/\beta$ or the scale β parameter. It's good practice to label the inputs.

> Suppose X is `gamma`(a, b). In R $P(X \leq q)$ is given by `pgamma(q,shape=a,rate=1/b)` or `pgamma(q,shape=a,scale=b)` or `pgamma(q,a,1/b)` (if you don't label the two parameters, R assumes (shape,rate)).

Example D.4. Suppose X has a gamma distribution with parameters $\alpha = 3$ and $\beta = 4$.

> a) Find $P(4 < X < 12)$.

This probability corresponds to the area pictured in Figure D.1, and can be computed in R, remembering to input the shape parameter α and the rate parameter $1/\beta$:

```
pgamma(12,shape=3,scale=4)-pgamma(4,shape=3,scale=4)
```

```
## [1] 0.4965085
```

Just about a 50% chance that a random value from a `gamma`$(3, 4)$ distribution is between 4 and 12.

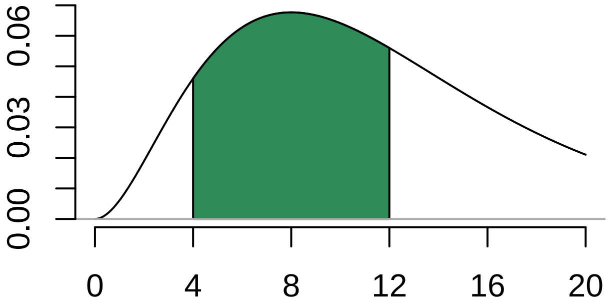


Figure D.1: Finding P(4<X<12) for a gamma(3,4) distribution

> b) Gather a random sample of 1000 values from this distribution, and determine what proportion of them live between 4 and 12.

```
x=rgamma(1000,3,1/4) # random sample of size 1000 (no parameter names <-> shape,r
sum(abs(x-8)<4) # values in the sample between 4 and 12
```

```
## [1] 476
```

Well, 476 is mighty close to half of the 1000 values!

Exponential distributions are special gamma distributions. In particular, if we set $\alpha = 1$, the gamma distribution gamma$(1,\beta)$ is exactly equal to the exponential distribution exp(β).

So, if X is exponential with mean 10, the following commands all compute $P(X \leq 5)$.

`pexp(5,rate=1/10)` $= 0.3934693$

`pgamma(5,shape=1,rate=1/10)` $= 0.3934693$

`pgamma(5,shape=1,scale=10)` $= 0.3934693$

D.5 Chi-square Distribution `chisq`

Like the exponential distribution, the chi-square distribution is a special gamma distribution. For a positive integer ν, the **Chi-square probability distribution with degrees of freedom** ν, denoted $\chi^2(\nu)$, is the gamma distribution with $\alpha = \nu/2$ and $\beta = 2$.

In R, `pchisq(x,df = v)` and `pgamma(x,shape = v/2,scale = 2)` will return the same value. For example, if $x = 7$ and $v = 10$, we have

`pchisq(7,df = 10)` $= 0.274555$, and

`pgamma(7,shape = 5,scale = 2)` $= 0.274555$

Here are plots of three different chi-square distributions.

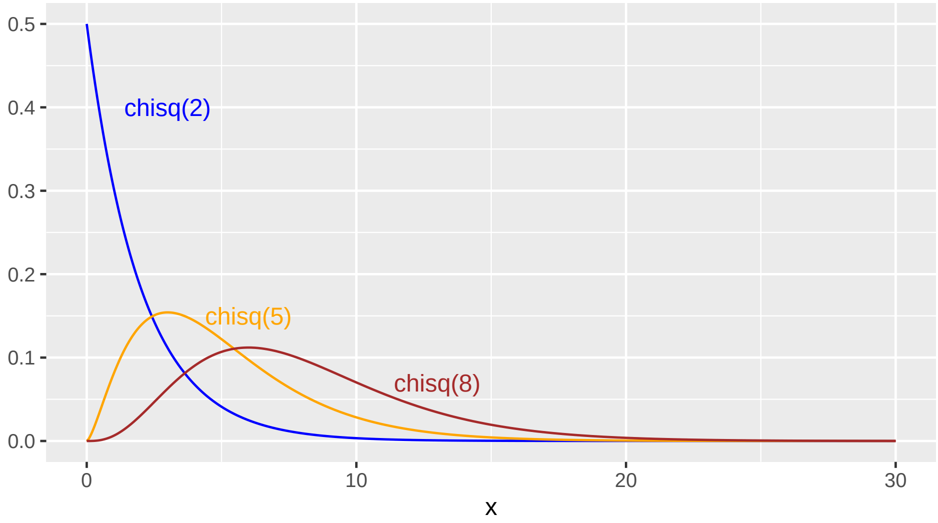


Figure D.2: Three chi-square distributions

D.6 Beta distribution `beta`

The beta(α, β) probability distribution provides a way to model random variables whose possible outcomes are all real numbers between 0 and 1. Such distributions are useful for modeling proportions. As with the gamma and normal distributions, this is a 2-parameter family of distributions.

Example D.5. Let X denotes the proportion of sales on a particular website that comes from new customers any given day, and suppose from past experience, X is well-modeled with a beta distribution with shape parameters $\alpha = 1$, and $\beta = 3.5$.

> Determine the probability that on any given day, over 1/2 the sales come from new customers.

```
1-pbeta(0.5,1,3.5)
```

```
## [1] 0.08838835
```

D.7 Homemade Continuous Random Variables

We may wish to study a continuous random variable X from a given probability density function such as $f(x) = \frac{3}{8}(2-x)^2$ for $0 \leq x \leq 2$.

In this case, probabilities such as $P(X > 1.2)$ correspond to areas under the density curve, which are calculated by integration, e.g.,

$$P(X > 1.2) = \int_{1.2}^{2} \frac{3}{8}(2-x)^2 \; dx.$$

If we can find an antiderivative of $f(x)$, we can find this probability using the fundamental theorem of calculus. If not, we can always estimate the value of the integral with Riemann sums. We do this below.

Input the density function

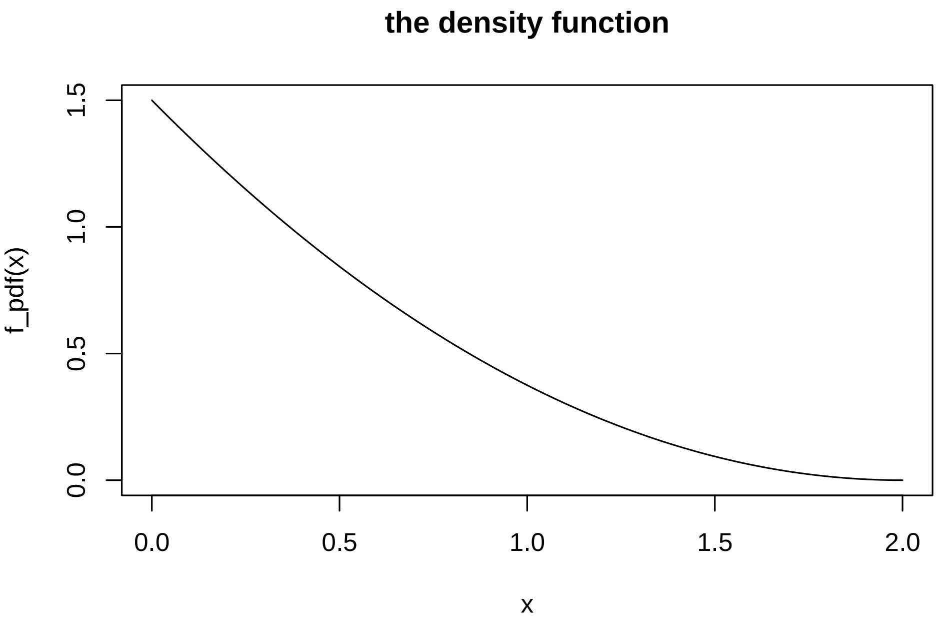
First build the probability density function (pdf) as a function in R.

```
f_pdf <- function(x){
  return(3*(2-x)^2/8)
  }
```

Visualize the density function

We create a vector of inputs `x` going from 0 to 2 in small increments (the increment is .01 below), to give us many points over the interval of interest [0,2]. Then we **plot the density curve** by plotting these x values against the function values f(x). (`type="l"` gives us a **l**ine plot instead of a point plot).

```
x=seq(0,2,by=.01)
plot(x,f_pdf(x),type="l",
     main="the density function")
```



Estimating Integrals with Riemann Sums

We know $P(X \geq 1.2)$ corresponds to the area under the density curve between 1.2 and 2. We can estimate areas by computing a Riemann Sum (a sum of many thin rectangle areas approximating the area under the density curve).

Here's a function for estimating $\int_a^b f(x)\ dx$ with a sum of n rectangle areas, generated using the midpoint rule.

```
mid_sum=function(f,a,b,n){
  #inputs:
      #f - function
      #a, b - lower and upper bounds of interval
      #n - number of subdivisions
  #output: The sum of the n rectangle areas whose heights are
  # determined by the midpoint rule
  dx=(b-a)/n
  ticks=seq(a+dx/2,b,dx)
  return(sum(f(ticks)*dx))
}
```

For instance, `mid_sum(f_pdf,a=0.4,b=1.2,n=4)` computes the area of the 4 rectangles in Figure D.3. We divide the interval [0.4,1.2] into n=4 equal-width subintervals, and build rectangles having height equal to the function height at the midpoint of each of these subintervals.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

The area of the four rectangles is

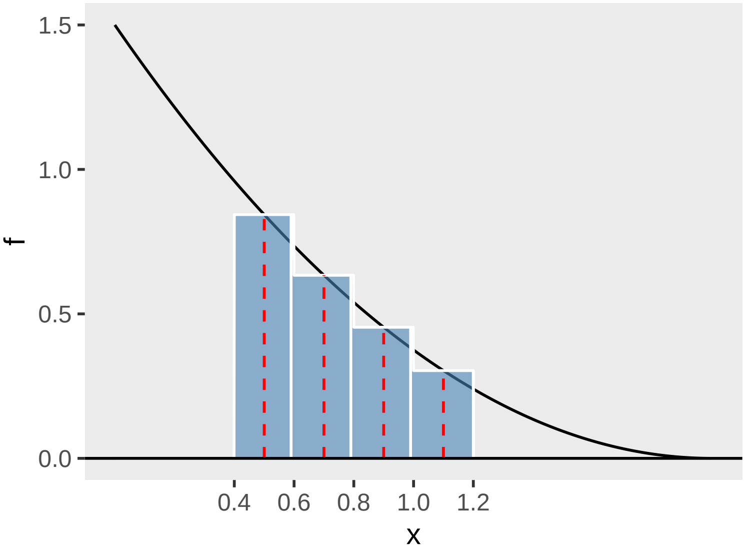


Figure D.3: Four midpoint rectangles

```
mid_sum(f_pdf,a=.4,b=1.2,n=4)
```

```
## [1] 0.447
```

Estimating Probabilities

So, getting back to our example, if we want to estimate $P(X > 1.2)$ we can compute a midpoint sum - the more rectangles the better. Let's start with $n = 100$:

```
mid_sum(f_pdf,1.2,2,100)
```

```
## [1] 0.0639984
```

What if we use $n = 1000$ rectangles?

```
mid_sum(f_pdf,1.2,2,1000)
```

```
## [1] 0.06399998
```

It seems as if our estimate hasn't changed much by going from 100 to 1000 subintervals, for this density function.

To estimate $P(0.5 < X < 1.1)$ we can evaluate

```
mid_sum(f_pdf,0.5,1.1,100)
```

```
## [1] 0.3307493
```

The distribution function $F(X)$

Recall, $F(x)$ gives cumulative probability. In particular, $F(x) = P(X \leq x)$.

Consider again the random variable X with pdf $f(x) = (3/8)(2-x)^2$ for $0 < x < 2$.

For any value of b between 0 and 2,

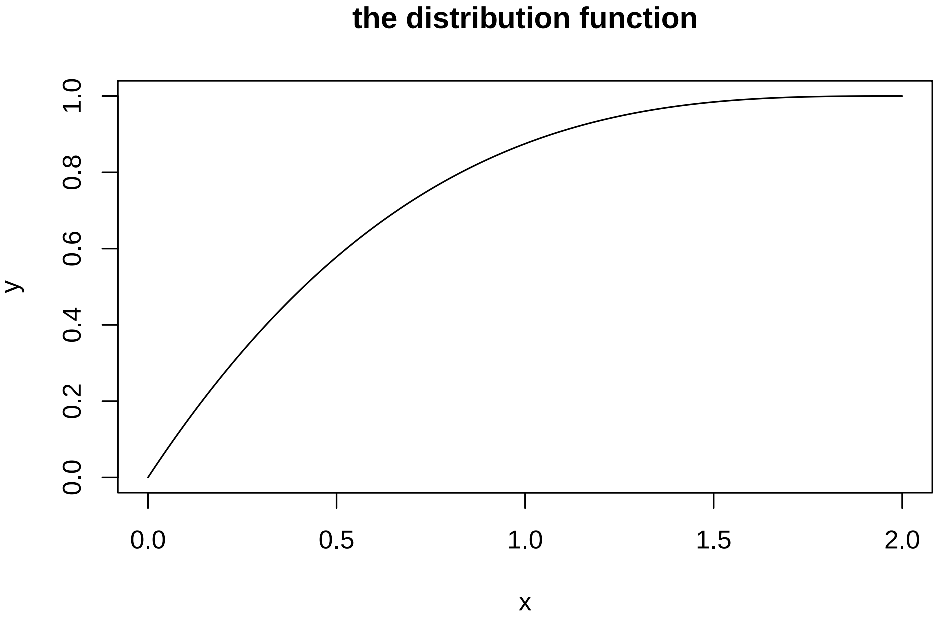
$$F(b) = \int_0^b f(x)\ dx,$$

which we can numerically approximate with

```
F_example <- function(b){
  return(mid_sum(f_pdf,0,b,100))
}
```

Then we can sketch the graph of the distribution function, for inputs between 0 and 2

```
x=seq(0,2,by=.01)
y=c()
for (i in 1:length(x)){
  y=c(y,F_example(x[i]))
}
plot(x,y,type="l",
     main="the distribution function")
```



Estimating Moments

Recall, $E(X^n)$ is called the *nth moment about 0* of the distribution. The first moment is the expected value $E(X)$, and the 2nd and 1st together determine the variance: $V(X) = E(X^2) - E(X)^2$.

For a continuous random variable X with pdf $f(x)$,

$$E(X^n) = \int_{-\infty}^{\infty} x^n \cdot f(x).$$

In R we can numerically estimate these integrals with the `mid_sum()` function defined above, applied to the integrand $x^n \cdot f(X)$.

```
moment.integrand<-function(f,n){
  #inputs:
      # f - a previously defined pdf
      # n - an integer
  #output: the integrand function for evaluating E(X^n)
  return(function(x){return(x^n*f(x))})
}
```

Expected Value

For a continuous random variable X,

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

To estimate this integral, we plug the first moment integrand $x \cdot f(x)$ into our Riemann sum function.

```
mu=mid_sum(moment.integrand(f_pdf,1),0,2,100)
mu
```

```
## [1] 0.500025
```

Note: The actual expected value is

$$\int_0^2 x \cdot (3/8)(2-x)^2\ dx = 0.5.$$

We estimate the **variance** knowing that $V(X) = E(X^2) - E(X)^2$.

```
EX2=mid_sum(moment.integrand(f_pdf,2),0,2,100)
EX2
```

```
## [1] 0.4
```

So the variance of X is

```
EX2-mu^2
```

```
## [1] 0.149975
```

Note: The actual value of $E(X^2)$ is

$$\int_0^2 x^2 \cdot (3/8)(2-x)^2\ dx = 0.4,$$

so $V(X) = 0.4 - (0.5)^2 = 0.15$.

Index