Introduction to Probability and Statistics

Ross Parker

August 7, 2020

These notes were used as the basis for APMA 1650 (Statistical Inference I), which I taught in the 2016 summer session at Brown University. Thanks especially to Katie Wu, Rebecca Santorella, Emily Winn, and Patrick Liscio for finding errors and typos in the text. If you find any more mistakes, please contact me so I can correct them. I hope you find this useful!

Contents

1 Probability Essentials

In general parlance, the term *probability* is used as a measure of a person's belief in the occurrence of an event. This is the *subjective* notion of probability. Take a sentence such as, "There is a 60% probability of rain tomorrow". Where did this come from? A professional meteorologist has used a sophisticated mathematical model to distill large amounts of atmospheric data down to a simple number indicating her belief that it is more likely to rain than to not rain tomorrow. We can interpret this number however we wish. We should be careful, however, to examine both our biases and those of the meteorologist. Who does this meteorologist work for? How reliable have their predictions been in the past? (We should be careful here since we might be more likely to recall rain when none was predicted than the other way around.) Are meteorologists more likely to err on the side of predicting rain, since that way fewer people will be upset if they are wrong? We can use this number to make important life decisions such as whether or not to carry an umbrella. The downside of this is that it is in essence a subjective opinion, even if that opinion comes from an expert. In addition, for subjective probabilities to make sense, they have to be internally consistent. In this example, if we believe there is a 60% chance of rain tomorrow, we must also believe that there is a 40% chance of it not raining tomorrow.

Let's look at a different example. What is the probability of rolling a 1 on a standard six-sided die? We can argue based on symmetry that since dice are cubical, there should be no reason for one face to be preferred over another. Thus this probability is 1/6. This is the *classical* notion of probability. In this interpretation, we use symmetry arguments to divide our experiment (a single die roll, in this case) into elementary events which are equally probable (the six distinct die rolls). We can compute the probability of any event using these elementary, equiprobable events. The limitation of this approach is that is requires a symmetry argument to be effective; thus it works for coin flips and die rolls, but not for more complicated scenarios.

We can look at the die roll experiment in another way. Imaging rolling a standard six-sided die repeatedly. The empirical probability of rolling a 1 is the ratio of the number of times a 1 is rolled to the total number of rolls. In general, we have:

$$\text{empirical probability of a certain event} = \frac{\text{number of times the event occurs}}{\text{total number of trials}}$$

Intuitively, as we perform more and more dice rolls, the empirical probability of rolling a 1 should approach some mythical quantity which we call the *true probability* of rolling a 1. This approach is the *empirical*, or *frequentist*, approach to probability. For a standard, six-sided die, it stands to reason that the empirical probability should approach the classical probability of 1/6 as the number of rolls approaches infinity. This result is called the *law of large numbers* and will be discussed later in the course. The empirical approach has a critical advantage over the classical approach in that we do not require symmetry to compute our probabilities. As an example of this, think of how you would determine the probability of rolling a 1 if someone handed you a loaded die. The empirical approach does, however, have its limits. Returning to the weather example, there is no way to think of the chance of rain

tomorrow as the limit of a sequence of independent experiments. Unless we are in the movie Groundhog Day, tomorrow can only happen once!

For our purposes, we require a more rigorous, mathematical construction of probability. This is known as *axiomatic probability* and will unify some of the aspects of the other approaches to probability. For this, we turn to the language of *set theory*.

1.1 Sample Spaces

A *set* is a collection of distinct objects. A *sample space*, denoted $\mathcal{S}$ is the set of all outcomes of a particular experiment. Here are some examples of sample spaces:

1. Single coin flip: $\mathcal{S} = \{H, T\}$
2. Roll of one standard, six-sided die: $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
3. Roll of two standard, six-sided dice: Here we represent the sample spaces as ordered pairs.

		second roll					
		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)
first roll	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)

4. Number of free throw attempts it takes for me to make a single basket: $\mathcal{S} = \{1, 2, 3, ...\}$. This set is often denoted $\mathbb{N}$ for the natural numbers (positive integers)
5. Number of minutes late my RIPTA bus arrives: $\mathcal{S} = [0, \infty)$.

Note that the first three sample spaces contain only a finite number of elements (2, 6, and 36 elements, respectively). These are called *finite sample spaces*. The fourth and fifth sample spaces both contain an infinite number of elements, but there is a fundamental difference between the two. The set $\mathbb{N}$ can be written out in its entirety in an infinitely long list; another way to think about this is that we can start at 1 and count up to any number in the set (as long as we have enough time!). A set with this property is called *countable*. For the set $[0, \infty)$, it makes intuitive sense that we cannot do this, i.e. we cannot list all the elements and, say, "count up to π". A proof of this fact is left for another course. Such an infinite set is called *uncountable*[1]. A sample space which is either finite or countable is called *discrete*.

[1] An uncountable set is a "larger infinity" than a countable set, which leads to the concept of "sizes of infinity". John Green alludes to this in his novel *The Fault in Our Stars*, but unfortunately gets the math wrong. If you find this interesting, I recommend the Vi Hart video `https://www.youtube.com/watch?v=23I5GS4JiDg`

1.2 Events and Subsets

An *event* is a subset of a sample space. Events are usually designated by capital letters, and we write the relationship "A is a subset of $\mathcal{S}$" by $A \subset \mathcal{S}$. For two events A and B, $A \subset B$ if every element in A is also contained in B. The *empty set*, denoted $\emptyset$, is the set containing no elements, and it is a subset of every set.

Let us consider the sample space $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, representing the roll of a single die. The following are examples of events:

1. $A = \{2, 4, 6\}$, the event that an even number is rolled
2. $B = \{1, 2, 3\}$, the event that the roll is less than or equal to 3
3. $C = \{1\}$, the event that a 1 is rolled

The event C consists of a single element in the sample space. Such an event is called a *simple event* and cannot be decomposed. The events A and B are each composed of three simple events.

Next, consider the sample space representing rolls of two dice. Let E be the event that the sum of the two dice is 7. We can represent this event graphically; in the figure below, the event E consists of the squares which are highlighted in yellow.

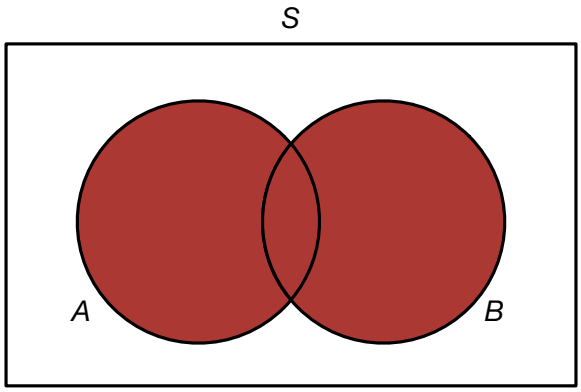
		second roll					
		1	2	3	4	5	6
first roll	1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
	2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
	3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
	4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
	5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
	6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

1.3 Basic Set Operations

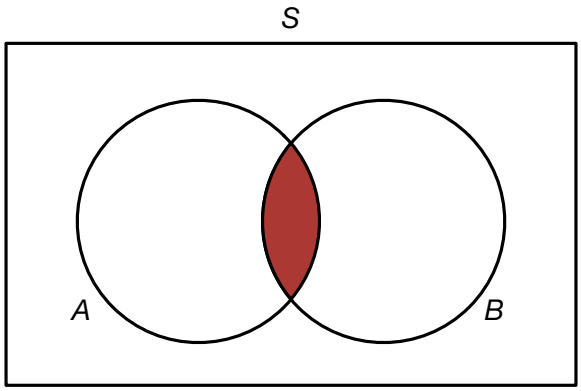
> "You have multiple core competencies with surprisingly minimal Venn. You can pivot from working on astrophysics problems, to teaching the young Arkers, to podcasting to folks on the ground, without skipping a beat!" - Neal Stephenson, *Seveneves*

Let $\mathcal{S}$ be our sample space, the set of all elements under consideration. Consider two events A and B which are subsets of $\mathcal{S}$. We have the following three basic set operations, which are handily illustrated using Venn diagrams.

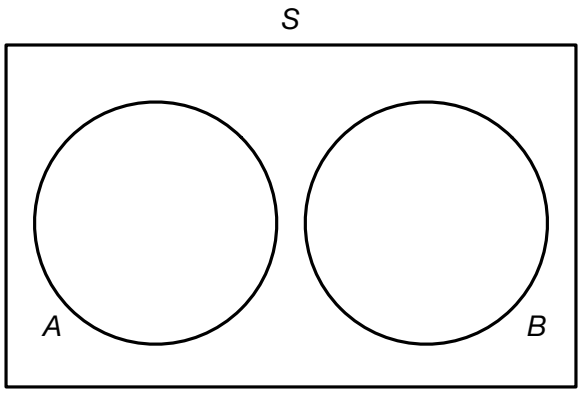
1. The *union* of A and B, denoted $A \cup B$, is the set of all elements which are in A or B (or both). That is, the union is all elements that are in at least one of the two sets.



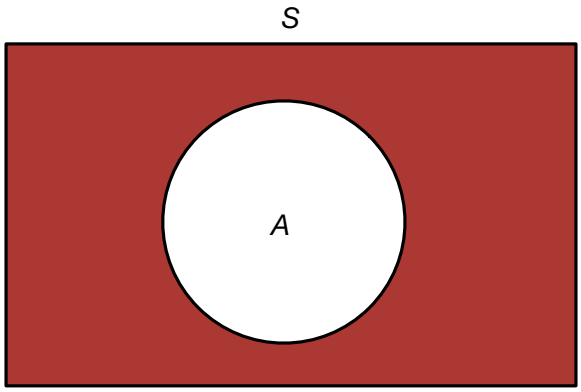
2. The *intersection* of A and B, denoted $A \cap B$, is the set of all elements which are in both A and B.



Two sets A and B are *disjoint* or *mutually exclusive* if they have no elements in common, i.e. if $A \cap B = \emptyset$.



3. The *complement* of A, denoted A^c, is the set of all points in $\mathcal{S}$ which are not in A. Note that A and A^c are disjoint, and $A \cup A^c = \mathcal{S}$.



There are many relationships between these operations which fall under the rubric of set algebra. Most of them we will not need, but we mention a few useful ones here:

1. Distributive laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

2. DeMorgan's laws

$(A \cap B)^c = A^c \cup B^c$ (the complement of an intersection is the union of the complements)
$(A \cup B)^c = A^c \cap B^c$ (the complement of a union is the intersection of the complements)

1.4 Axiomatic Definition of Probability

Equipped with our our knowledge of set theory, we can define probability axiomatically as follows. Given any event A in our sample space $\mathcal{S}$, we assign a probability $\mathbb{P}(A)$ to that event such that the following rules hold[2]:

1. $0 \leq \mathbb{P}(A) \leq 1$

   The probability of an event is a real number between 0 and 1, where a probability of 0 means that the event will never occur, and a probability of 1 means that the event will always occur.

2. $\mathbb{P}(\emptyset) = 0$

   The probability that nothing happens is 0, i.e. something must happen.

[2]You can construct a coherent notion of probability with fewer axioms and derive the remaining rules from these; I like this version of probability rules, since it codifies what we want to be true given our intuitive notion of probability.

3. $\mathbb{P}(\mathcal{S}) = 1$

   The probability of the whole sample space is 1, which is another way of saying that something must happen.

4. If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$

   If we make a set bigger, its probability can only increase (or stay the same); it cannot decrease.

5. If $A_1, A_2, \ldots, A_n$ are pairwise disjoint events, i.e. $A_i \cap A_j = \emptyset$ if $i \neq j$, then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

   This holds for an infinite sequence as well, i.e. if $A_1, A_2, A_3, \ldots$ are a sequence of pairwise disjoint events, then

$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

From these we can derive a very important rule:

$$\mathbb{P}(A) + \mathbb{P}(A^c) = 1, \quad \text{i.e. } \mathbb{P}(A) = 1 - \mathbb{P}(A^c)$$

Sometimes it is easier to calculate the probability of an event *not* happening than the probability of the event itself!

These rules tell us the properties that we want probability to have. However, given a sample space, they do not actually tell us how to assign probabilities to each event in the sample space. Doing that in a way that is consistent with the above rules can be a bit tricky [3], but luckily for a discrete sample space we can do this with no problem. Since a discrete sample space is composed of a finite (or countable) number of simple events, all we have to do is assign probabilities to each simple event in such a way that they all add up to 1.

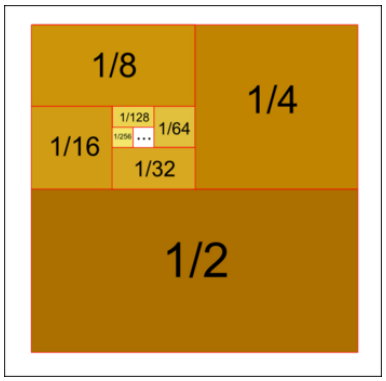
Example. Consider once again tossing a single die. The sample space for this is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$. This sample space contains 6 simple events: $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$, and $\{6\}$. We can assign any probabilities we want to these simple events, as long as they add up to 1. For example, assuming we have a fair die, we can let $\mathbb{P}(\{i\}) = 1/6$ for $i = 1, \ldots, 6$. If we like, we can check that all the above rules hold. If we have a loaded die, which rolls a 6, say, half the time, we could assign probabilities: $\mathbb{P}(\{6\}) = 1/2, \mathbb{P}(\{i\}) = 1/10$ for $i = 1, \ldots, 5$.

[3] In fact, for an uncountable sample space such as $\mathcal{S} = [0, 1]$, you can show that you cannot construct a notion of probability which is consistent with all the rules; this is the starting point for the development of measure theory, and is beyond the scope of this course.

Example. Consider this time a countable sample space $\mathcal{S} = \mathbb{N} = \{1, 2, 3, \dots\}$. One possibility is to assign probabilities $\mathbb{P}(\{i\}) = 1/2^i$ for $i = 1, 2, 3, \dots$, i.e. $\mathbb{P}(\{1\}) = 1/2$, $\mathbb{P}(\{2\}) = 1/4$, $\mathbb{P}(\{3\}) = 1/8$, etc. Perhaps you recall from calculus that this is a geometric series with first element $1/2$ and common ratio $1/2$, and so we know its sum is:

$$\sum_{i=1}^{\infty} \mathbb{P}(\{i\}) = \sum_{i=1}^{\infty} \frac{1}{2^i} = \frac{1}{2} \frac{1}{1 - 1/2} = 1$$

Since the sum of the probabilities of all the simple events is 1, we are all set! If you have not seen this before, we will cover this in more detail when we discuss the geometric distribution. In the meantime, here is a nice picture to convince you that the sum is indeed 1.



1.5 Discrete Uniform Distribution

The first probability distribution we will consider is the uniform distribution on a *finite* sample space. In the discrete uniform distribution, every simple event is equally likely to occur.

Suppose we have a finite sample space $\mathcal{S}$ with n simple events. The discrete uniform distribution assigns each simple event a probability of $1/n$. Why is this the case? If each simple event has probability p and there are n simple events $A_1, \dots A_n$ in our sample space, then using the fact that the sample space S must have probability 1 and that the simple events are pairwise disjoint:

$$\begin{aligned} 1 &= \mathbb{P}(S) \\ &= \mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \\ &= \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n) \\ &= np \end{aligned}$$

where in the last line we use the fact that the simple events each have probability p, and there are n of them. Solving for p, we get $p = 1/n$.

How do we find the probability of an event that is not a simple event? Again, consider a discrete sample space $\mathcal{S}$ with n simple events. Suppose we have an event A which is a subset

of $\mathcal{S}$. Since we have a finite sample space, A is composed of a finite number m of simple events, where m is an integer between 0 and n. The probability of A is:

$$\mathbb{P}(A) = \frac{\text{number of simple events in } A}{\text{number of simple events in } \mathcal{S}} = \frac{m}{n}$$

Example. Consider the finite sample space $\mathcal{S}$ representing the roll of two standard dice. This sample space has 36 simple events, so each simple event has a probability of 1/36. Note that for the simple events in this sample space, the order of the die rolls matters. $(1, 6)$ and $(6, 1)$ are two different events; even though both events contain the same die rolls, for the former, the 1 is rolled first, while for the latter, the 6 is rolled first.

1. What is the probability that the sum of the two dice is 7? Looking at the graphical depiction of this event above, we see that there are 6 simple events that give us a sum of 7. Thus the probability of a sum of 7 is $6/36 = 1/6$.

2. What is the probability that the sum of the two dice is less than 11. In this case, it is easier to compute the probability that the sum is 11 or greater and then subtract that probability from 1. (Why can we do this?) Let A be the event that the sum is 11 or greater. A is composed of 3 simple events: $(6, 6), (6, 5)$, and $(5, 6)$. Thus we have $\mathbb{P}(A) = 3/36$. The probability we want is:

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) = 1 - 3/36 = 33/36$$

In the previous example, it is relatively straightforward to draw the sample space, so we can essentially compute any probability we want simply by looking at the picture and counting which boxes comprise our event of interest. For more complicated problems, this is not as easy. Consider the following example:

Example. A communication system consists of n antennas arranged in a line. Exactly m out of the n antennas are defective. The system is functional if no two consecutive antennas are defective. Assuming that each linear arrangement of the antennas is equally likely, what is the probability that the system will be functional?

For small values of n and m, say $n = 4$ and $m = 2$, we can write out all of the possible configurations. Representing a functional antenna by `1` and a defective antenna by `0`, there are exactly six linear arrangements:

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

Take a moment to convince yourself that these are the only possible configurations. Configurations 1, 2, and 5 are functional, so there are 3 functional configurations out of 6 total

configurations. So in this case, the probability that the system is functional is 3/6.

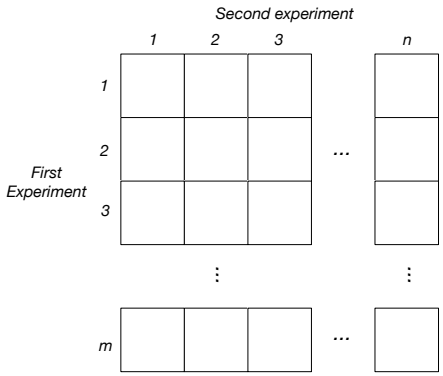
For general n and m, it is not immediately obvious how to perform the requisite counting of configurations. Taking a cue from the Count on Sesame Street, we need to learn more about counting. The mathematical theory of counting is known as *combinatorics*.

1.5.1 The Basic Principle of Counting

The basic principle of counting (mn-rule)

Suppose we are performing two experiments. If the first experiment has m possible outcomes and the second experiment has n possible outcomes, there are mn possible outcomes for the two experiments. We can generalize this to more than two experiments; in that case, we take the product of the number of outcomes from each experiment.

To see this, draw a grid of boxes with the m outcomes from the first experiment on the left and the n outcomes of the second experiment across the top. There are mn total boxes, which are all the possible outcomes of both experiments.



We have already seen the basic principle of counting in play when we looked at the sample space for two dice. Each of the two die rolls is a separate experiment. Since there are 6 outcomes for each die roll, the total number of outcomes is $6 \cdot 6 = 36$. We can extend this to three or more dice. For three dice, for example, the total number of outcomes is $6 \cdot 6 \cdot 6 = 6^3 = 216$. If you like, you can visualize this as a cube. For n dice, there are 6^n possible outcomes. I have trouble visualizing this for $n > 3$, but perhaps you can.

We can also think of the basic principle of counting in terms of choosing items from groups. If there are m items in Group 1 and n items in Group 2, there are mn pairs of items consisting of one item from each group. Again, this can be extended to any number of groups.

Example. When I was growing up in Virginia, standard (non-vanity) license plates were composed of three letters (A-Z) followed by three digits (0-9). Assuming that all such possibilities can exist:

1. How many possible license plates are there?

   We are choosing items from 6 groups. The first three groups contain 26 items, and the last three groups contain 10. Thus the number of possibilities is: $26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 = 17,576,000$. (That works out to a little more than two cars per person.)

2. If the Virginia DMV decided that letters and numbers could not be repeated, how many possible license plates would there be?

   The difference here is that the group sizes shrink as we choose items. For the letters, the first group has 26 items. The second group only has 25 items, since we cannot choose the letter we chose from the first group. The third group has 24 items, since we cannot choose the letter we chose from the first two groups. The digits are similar. Thus the number of possibilities is: $26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 = 11,232,000$

1.5.2 Permutations

An ordered arrangement of r distinct items is called a *permutation*. The number of ways of ordering r items drawn from a group of n distinct items is designated nPr.

Before we give the formula for the number of possible permutations, let's look at some examples so we can get an intuitive understanding of what is going on.

Example. How many different ordered arrangements are there of the letters a, b, and c?

In this case, we can actually write them all out: *abc*, *acb*, *bac*, *bca*, *cab*, *cba*.
From this we see that there are 6 possible ordered arrangements. We can also do this using the "choosing-from-groups" approach. For the first letter, we have 3 to choose from; for the second, we have only 2; and for the third, there is only one remaining letter. Thus the number of permutations is:

$$3 \cdot 2 \cdot 2 = 3! = 6$$

Similarly, the number of ordered arrangements of n distinct symbols is:

$$n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 = n!$$

The symbol $n!$ is the *factorial* operation, and it is defined exactly as written above.

Example. Every December (since 1996), the city of Ithaca, NY hosts the International Rutabaga Curl[4] (contestants must supply their own rutabaga). Gold, silver, and bronze medals are given to the top three finishers. If 100 contestants enter the rutabaga curl, how many possibilities are there for the winners?

[4] `http://www.rutabagacurl.com`

For the gold medalist, we have 100 contestants to choose from. Since you cannot win more than one medal, we choose from 99 contestants for the silver medal and 98 contestants for the bronze medal. The number of medal possibilities is:

$$100 \cdot 99 \cdot 98 = 970,200$$

We can generalize this in the following formula:

The number of permutations (ordered arrangements) of r items drawn from a group of n distinct items is:

$$nPr = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

To see this, there are:

- n ways to choose the first item
- $n-1$ ways to choose the second item

  ...

- $n-r+1$ ways to choose the rth item

Multiply these together to get the permutation formula. To get the "factorial form" of the permutation formula, we have:

$$nPr = n(n-1)(n-2)\cdots(n-r+1)\frac{(n-r)!}{(n-r)!} = \frac{n!}{(n-r)!}$$

where $n! = n(n-1)\cdots 2 \cdot 1$, and we define $0! = 1$.[5]

1.5.3 Combinations

Consider the following example, which is modification of the Rutabaga Curl example above.

Example. 100 students buy raffle tickets. Three names are chosen from a hat uniformly at random to win a free sandwich at Eastside Pockets[6]. How many possibilities are there for the winners?

How is this problem fundamentally different from the Rutabaga Curl? Here, the three prizes are identical, as opposed to the three distinct medals in the Rutabaga Curl. We can think of this problem as selecting 3 items from a group of 100, but *the order in which we select them does not matter*. Let's look at this in a few stages:

[5]This may seem a little arbitrary, but it is convenient. It also makes some sense that there is exactly one way to arrange no items.

[6]A popular eatery on Thayer St. near Brown University

1. Let's pretend for a moment that the order of selection matters. Then, as in the Rutabaga Curl, there are $100 \cdot 99 \cdot 98$ possibilities.

2. Compared to the the case where order matters, is the number of possibilities greater, fewer, or the same? There must be fewer possibilities since, for example, choosing the numbers $(1, 2, 3)$ from the hat in that order is the same as choosing them in the order $(3, 2, 1)$.

3. How many permutations correspond to choosing the numbers 1, 2, and 3 from the hat? We can write out all the permutations:

$$(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), \text{ and } (3, 2, 1)$$

   This gives us a total of 6 permutations. But this number is 3!, the number of ordered arrangements of 3 objects. Does this make sense that this should be the case.

4. Since we have shown that 6 permutations correspond to one possibility of winners, all we need to do is divide the number of permutations by 6. Thus the total number of possibilities for the winners is:

$$\frac{100 \cdot 99 \cdot 98}{6}$$

An unordered arrangement of r distinct items is called a *combination*. The number of combinations of r items drawn from a group of n distinct items is denoted $\binom{n}{r}$ (or sometimes nCr). We can think of this as:

1. The number of subsets of size r which can be formed from a group of n distinct objects

2. The number of ways to select r items from a group of n distinct items, where order does not matter

Using the logic from the previous example, we can deduce the formula for $\binom{n}{r}$. All we have to do is take the number of permutations and divide by $r!$, which is the number of ordered arrangements (permutations) of r items.

The number of combinations (unordered arrangements) of r items drawn from a group of n distinct items is:

$$\binom{n}{r} = \frac{nPr}{r!} = \frac{n!}{r!(n-r)!}$$

Example. The Brown crossword puzzle club consists of 5 undergraduate and 7 graduate students.

1. How many different committees of 2 undergrads and 3 graduate students can be formed?

   There are $\binom{5}{2}$ possible groups of 2 undergrads, and $\binom{7}{3}$ possible groups of 3 graduate students. We multiply these together (basic principle of counting) to get:

$$\binom{5}{2}\binom{7}{3} = \frac{5!}{2!3!}\frac{7!}{3!4!} = \frac{5\cdot 4}{2\cdot 1}\frac{7\cdot 6\cdot 5}{3\cdot 2\cdot 1} = 10\cdot 35 = 350$$

2. What if two of the graduate students refuse to be on the same committee?

   There are $\binom{7}{3} = 35$ possible groups of 3 graduate students. How many contain both of the two students who refuse to serve together? To make a three-person committee which includes these two, you need to include the two rival students plus one other student selected from the five remaining graduate students. Thus there are $\binom{2}{2}\binom{5}{1} = 5$ committees which include the two rival students. Subtracting from 35, there are 30 committees which don't include both rival students. As above, we multiply to get $10 \cdot 30 = 300$ possible committees.

Computing probabilities of poker hands are classic problems in probability. Since the order of the cards in a poker hand does not matter, these problems involve combinations.

Example. You are dealt a five-card poker hand. What is the probability of getting a full house (three cards of one number plus two cards of another number)

There are 52 cards in a poker deck, and a poker hand is 5 cards. Since the order of the cards dealt does not matter, there are $\binom{52}{5}$ possible five-card poker hands. To count the number of hands which give us a full house, we use the following procedure:

1. Choose the number for the three-of-a-kind. There are $\binom{13}{1}$ ways to do that. Now select three out of the four cards of this number. There are $\binom{4}{3}$ ways to do that. Thus there are $\binom{13}{1}\binom{4}{3}$ ways to choose the three-of-a-kind.

2. Choose the number for the pair. There are $\binom{12}{1}$ ways to do that, since we have already chosen one of the numbers for the three-of-a-kind. Now select two out of the four cards of this number. There are $\binom{4}{2}$ ways to do that. Thus there are $\binom{12}{1}\binom{4}{2}$ ways to choose the pair once the three-of-a-kind has been chosen.

3. Multiply all of these together to get $\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}$ poker hands which are full houses.

To get the probability of a full house, we divide by the number of poker hands (the size of the sample space) to get:

$$\frac{\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}}{\binom{52}{5}} = 0.00144$$

It is worth noting that the number of full houses is not $\binom{13}{2}\binom{4}{3}\binom{4}{2}$, i.e. choose two numbers, then choose a three-of-a-kind, then choose a pair. Why does this not work? This method does not distinguish between, say, `33366` and `33666` (i.e. it treats these as the same), and these are two distinct full houses. Since this undercounts by a factor of two, you can check that if you multiply this by two, you get the correct answer above.

Combinations are incredibly useful. We can use them in some rather surprising cases.

Example. Consider the binary string `11000`. How many distinct orderings are there of the digits in this string?

At first, this does not appear to involve combinations at all, since we are looking for orderings, and combinations are used where order does not matter. Let's look at this problem in a different way. Consider the five-element set $A = \{a, b, c, d, e\}$. There are $\binom{5}{2}$ two-element subsets of A. One way to describe subsets of A is to make a table. The columns of the table represent the elements of A: a, b, c, d, and e. Each row corresponds to a subset of A: a 1 indicates that the element is in the subset, and a 0 indicates that it is not. Here is an example table, where we depict two two-element subsets.

a	b	c	d	e	subset
0	0	1	1	0	$\{c, d\}$
1	0	1	0	0	$\{a, c\}$

Notice that two-element subsets match up exactly with orderings of `11000`! Thus we see that there are $\binom{5}{2}$ rearrangements of the binary string `11000`.

In general, if you have a string of length n composed of two symbols, and you have r of one symbol (thus $n - r$ of the other), the number of distinct orderings of the string is $\binom{n}{r}$. Equipped with this, let's return to the antenna example from the beginning of the section.

Example. A communication system consists of n antennas arranged in a line. Exactly m out of the n antennas are defective. The system is functional if no two consecutive antennas are defective. Assuming that each linear arrangement of the antennas is equally likely, what is the probability that the system will be functional?

First we need to figure out how many possible arrangements there are, i.e. the size of our sample space. If we use the digit `1` to represent a functional antenna, and the digit `0` to represent a defective antenna, then each arrangement can be represented by a binary string of length n consisting of m 0s and $n - m$ 1s. Thus from what we learned above, there are $\binom{n}{m}$ possible arrangements.

How many of these arrangements are functional? Let's draw a picture! Imagine we have the $n - m$ functional antennas in a line. We will represent each functional antenna by the symbol *, and the spaces between the functional antennas by a vertical bar |. We will also place a space to the far right and to the far right:

$$| * | * | * | * | * | \ldots | * | * |$$

For the system to be functional, each space | can contain at most one defective antenna. There are $n - m + 1$ spaces to choose from, and m defective antennas to place. Thus there are $\binom{n-m+1}{m}$ functional arrangements. Dividing this by the total number of arrangements

$\binom{n}{m}$, the probability that the system is functional is:

$$\frac{\binom{n-m+1}{m}}{\binom{n}{m}}$$

For a sanity check, let's calculate this for the case where $n = 4$ and $m = 2$, which we did manually above:

$$\frac{\binom{n-m+1}{m}}{\binom{n}{m}} = \frac{\binom{3}{2}}{\binom{4}{2}} = \frac{3}{6} = \frac{1}{2}$$

This agrees with what we found above!

Example. I buy an assortment of 10 bagels from Bagel Gourmet on Thayer St.[7]. There are five types of bagels to choose from: everything, onion, poppy, raisin, and sesame. How many different assortments of bagels can I bring home? Assume there are at least 10 of each kind of bagel in the store.

We can model the problem in a way that lets us use combinations. Imagine you buy the bagels in a special box. The bagels are arranged in a line, and are always in alphabetical order: everything, onion, poppy, raisin, sesame. There are four dividers, one between each bagel type, and if a bagel type is not present, we just put the dividers next to each other. Here is an example of three bagel boxes with their dividers. Bagels are designated by * and dividers by |.

$**\|*\|***\|**\|**$	2 everything, 1 onion, 3 poppy, 2 raisin, 2 sesame
$***\|\|***\|\|**$	3 everything, 3 poppy, 2 sesame
$**********\|\|\|\|$	10 everything

From the picture, we see that an assortment of 10 bagels can be represented as a linear string of 14 symbols: 10 * representing the bagels and 4 | representing the dividers. The number of assortments is the same as the number of distinct orderings of this linear string: $\binom{14}{4} = 1001$

One final application of combinations is to the binomial theorem. The binomial theorem gives us a nice expression for expanding the binomial $(x + y)^n$, where n is a positive integer.

The binomial theorem

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

[7]Having lived in Brooklyn for many years, I am only *slightly* obsessed with bagels.

Due to their appearance in this theorem, the coefficients $\binom{n}{r}$ are often called *binomial coefficients*. Rather than give the usual proof by induction, we will argue this is true using what we know about combinations. Consider the product:

$$(x_1 + y_1)(x_2 + y_2) \cdots (x_n + y_n)$$

If we multiply this out, we get a sum of 2^n terms, each of which is the product of n factors. To see this is the case, to get a term in the sum, we pick either x_i or y_i from each of the n binomials $(x_i + y_i)$ and multiply them together. There are 2^n such terms, since for each binomial we have two possible choices and there are n total binomials. How many of these 2^n terms have exactly r of the x_i terms and $n - r$ of the y_i terms? We can think of each term as a string of length n composed of r x's and $n - r$ y's. By what we learned above, there are $\binom{n}{r}$ such terms. Taking $x_i = x$ and $y_i = y$ for all i, we obtain the binomial theorem.

1.5.4 Multinomials

The number of ways of dividing n objects into k distinct groups containing $n_1, \ldots, n_k$ objects each, where each object appears in exactly one group, and $n_1 + \cdots + n_k = n$ is given by:

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

The term on the left is called a *multinomial coefficient*

Note that this is the same as the binomial coefficient when $k = 2$. To see why this is true, we will use the following argument:

1. For the first group, we choose n_1 out of n items. Since order does not matter, there are $\binom{n}{n_1}$ ways to do this.

2. For each choice in step 1, there are $\binom{n-n_1}{n_2}$ choices for the second group, since we have already taken n_1 items out of the larger group.

3. For each choice in step 1 and step 2, there are $\binom{n-n_1-n_2}{n_3}$ choices for the third group, since we already have taken $n_1 + n_2$ items out of the larger group.

4. Repeat this until we get to the kth group, where there are $\binom{n-n_1-\cdots-n_{k-1}}{n_k}$ choices for the kth group given all the previous choices

5. Multiply these together to get:

$$\begin{aligned}
\binom{n}{n_1\, n_2\, \cdots\, n_k} &= \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{k-1}}{n_k} \\
&= \frac{n!}{n_1!(n-n_1)!}\,\frac{(n-n_1)!}{n_2!(n-n_1-n_2!)}\,\frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3!)}\cdots\frac{n-n_1-\cdots-n_{k-1}!}{n_k!(n-n_1-\cdots-n_k)!} \\
&= \frac{n!}{n_1!\, n_2!\, n_3!\cdots n_k!(n-n_1-\cdots-n_k!)} \\
&= \frac{n!}{n_1!\, n_2!\, n_3!\cdots n_k!\, 0!} \\
&= \frac{n!}{n_1!\, n_2!\, n_3!\cdots n_k!}
\end{aligned}$$

where, after some very satisfying cancellation, we have used the fact that $n_1 + \cdots + n_k = n$ and $0! = 1$.

If you want to think about this another way, you can use the following reasoning. Imagine we have n objects and want to divide them into groups as above. Imagine we have a special box which contains the n objects in a line. Divide the box into sections using dividers, so that the first section contains n_1 objects, the second n_2 objects, all the way down to the kth section which contains k objects. For example, if we had $n = 10$ with four groups $n_1 = 2, n_2 = 4, n_3 = 1, n_4 = 3$, our box looks like:

$$**\,|\,****|\,*\,|\,***$$

where * represents a space to put an objects, and | is a section divider. There are $n!$ permutations of n objects, thus there are $n!$ distinct ways to place the n objects into the box. Since we are dividing our objects into groups, order does not matter within each group. The first section in the box represents the first group, so since order does not matter within that section, we divide by $n_1!$, the number of permutations of n_1 objects. Repeating this for each section, we divide by each $n_i!$ in turn to obtain our multinomial coefficient.

Example. A group of 15 international relations students will be divided into three groups. Each group consists of 5 students and is assigned a different country (Azerbaijan, Belarus, and Chechnya) on which to do a group final presentation. How many such divisions are possible:

Since we are are dividing 15 people into 3 distinct groups of 5, we can use the multinomial coefficient to determine the number or arrangements:

$$\binom{15}{5\,5\,5} = \frac{15!}{5!\,5!\,5!} = 756,756$$

Example. Now suppose we are dividing a group of 15 international relations students into three groups of 5 students each for a final project of the each group's choosing. How is this

problem different from the previous one? How many different arrangements are possible?

In the previous example, the three groups were distinct, so it matters which one a student is assigned to. In this case, the three groups are identical. Thus since the order of the three groups does not mater, we divide the previous result by 3!, which is the number of ways of ordering three items (in this case, the items are the three groups).

$$\frac{\binom{15}{5\,5\,5}}{3!} = \frac{15!}{5!\,5!\,5!\,3!} = 126,126$$

Example. How many different, distinct rearrangements are there of the word MISSISSIPPI?

We are looking for all distinct orderings of a string with four distinct symbols. If there were only two distinct symbols, we could use binomial coefficients, so we suspect multinomial coefficients may be in play here. Once again, consider a set with eleven elements in them, which we will label with the integers 1 - 11. We want to divide those elements into four subsets with 1, 2, 4, and 4 elements (respectively). We will label those subsets by the letters M, P, I, and S (no coincidence here!). Let's make a table, where each row is a subset, and letters M, I, S, and P in the row indicate which subset each element belongs to.

1	2	3	4	5	6	7	8	9	10	11
M	I	S	S	I	S	S	I	P	P	I
I	S	I	P	M	P	S	I	S	S	I

Note that there is a perfect correspondence between each appropriate collection of four subsets and each arrangement of MISSISSIPPI. Thus we can use the multinomial coefficient for this problem, and we get that the total number of distinct rearrangements is:

$$\binom{11}{1\,2\,4\,4} = \frac{11!}{1!\,2!\,4!\,4!} = 34,650$$

1.6 Conditional Probability

Sometimes the probability of an event will depend on whether or not another event has occurred. There are many real-life examples of this:

1. In weather prediction, the probability of rain tomorrow depends on whether or not it is raining today. This especially makes sense during the tropical storm season when it will often rain for days on end.

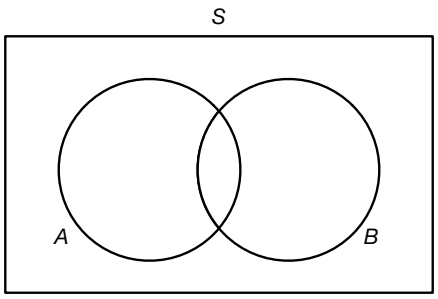
2. In the field of public health, probabilities of developing disease depend on many outside factors. The probability of developing lung cancer, for example, depends on factors such as smoking history, exposure to second-hand smoke, and occupational exposure to asbestos.

3. Poker players are constantly thinking about conditional probability. For example, what is the probability of an inside straight draw on the final two cards given the cards the player has already seen?

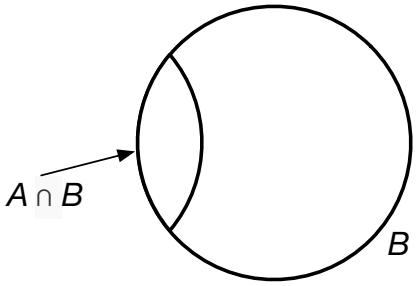
Let's do a simple die-tossing example, and then generalize.

Example. On a standard, fair six-sided die, the probability of rolling a 1 is 1/6. This is the *unconditional probability* of rolling a 1, since this probability does not depend on any additional information. Now suppose we know that our die roll is odd. What is the *conditional probability* of rolling a 1 given the die roll is odd? There are 3 possible odd rolls: 1, 3, and 5. We reduce our sample space to the three odd rolls since we know one of those occurred. Since we still have a discrete uniform distribution, each of the three odd rolls is equally likely. Thus the probability of rolling a 1 given an odd roll is 1/3.

Note that what we did above is reduce our sample space from all six possible rolls to the three odd rolls. We then kept the uniform distribution on the the new, smaller sample space. To generalize this to arbitrary events, let's draw a picture! Let $\mathcal{S}$ be our sample space, and consider two events A and B in $\mathcal{S}$.



We are interested in the probability of A given that B has occurred, which we will write $\mathbb{P}(A|B)$ ("the probability of A given B"). If $\mathbb{P}(B) = 0$, i.e. B cannot occur, then $\mathbb{P}(A|B)$ must also be 0 (does this make sense?) So we are only interested in the case where $\mathbb{P}(B) > 0$, i.e. B can actually happen. What we are going to do is *restrict* our sample space from $\mathcal{S}$ down to B since the only events we care about are ones where B has occurred. Imagine taking an eraser (or digital equivalent) to the picture above and removing everything that is outside of B. We are left with the following picture:



The piece of A which lies inside B is $A \cap B$, which is on the left of the picture above. If we think of probability intuitively as "area" in pictures like this, it makes sense that the conditional probability of A given B is the ratio of $\mathbb{P}(A \cap B)$ to $\mathbb{P}(B)$.

Let A and B be two events. The *conditional probability* of A given that B has occurred is given by:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

as long as $\mathbb{P}(B) = 0$. (If $\mathbb{P}(B) = 0$, then this is zero.)

As a sanity check, let's use this definition of conditional probability and redo our die roll example from above.

Example. What is the probability of rolling a 1 on a standard six-sided die given than an odd number was observed?

Let A be the event that a 1 is rolled, and B the event that an odd number is rolled. Since $A \subset B$, $A \cap B = A$ (does this make sense?) Using the discrete uniform distribution, $\mathbb{P}(A) = 1/6$ and $\mathbb{P}(B) = 1/2$ (since B is composed of three simple events, each with probability $1/6$). Thus we have:

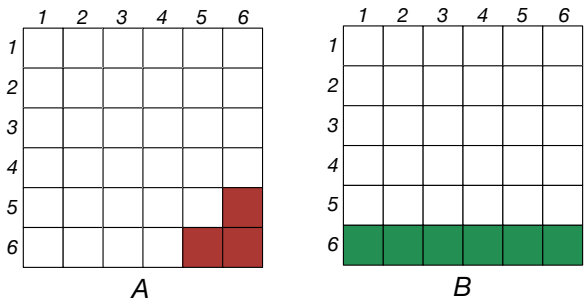
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{(1/6)}{(1/2)} = \frac{1}{3} \approx 0.339$$

This agrees with our intuitive computation above.

Let's do another example, this time using the roll of two dice.

Example. You roll two standard, fair six-sided dice. What is the the probability that the sum is greater than or equal to 11, given that the first roll is a 6?

Let A be the event that the sum is greater than or equal to 11, and B the event that the first roll is a 6. The events are shown graphically below.



First let's use the sample space reduction method. Given that B has occurred, i.e. the first roll is a 6, we reduce our sample space to B, i.e. to the bottom row of the picture. There are six squares in the bottom row, and since we still have the uniform distribution, each is equally likely. Of those 6, 2 of them (the rightmost two) have a sum of dice greater than or equal to 11. Thus the conditional probability is $\mathbb{P}(A|B) = 2/6 = 1/3$.

We can also use our definition of conditional probability. Using the uniform distribution, $\mathbb{P}(B) = 6/36$. Since $A \cap B$ is the two rightmost square on the bottom row, $\mathbb{P}(A \cap B) = 2/36$. By the definition of conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{2/36}{6/36} = \frac{1}{3}$$

Thus again we get the same answer using either method.

Sometimes using a reduced sample space is much easier that using the definition of conditional probability. Consider the following example.

Example. In the card game bridge (which is near and dear to my heart), 52 cards are dealt out equally to 4 players (called North, South, East and West). Given that North and South have a total of 8 spades between them (out of 13 total spades), what is the probability that East has 3 spades.

Here it is easiest to work with a reduced sample space. If North and South have 26 cards and 8 spades between them, then East and West much have the remaining 26 cards and 5 spades between them. The reduced sample space we will work with is the sample space of possible hands for East. Since East has 13 of the remaining 26 cards, there are $\binom{26}{13}$ possible hands for East. This is the size of the sample space. In how many of those hands does East have exactly 3 spades? To do this, we first choose 3 of the remaining 5 spades. There are $\binom{5}{3}$ ways to do this. Then we need to choose the non-spade cards. East has 10 non-spade cards, and there are $26 - 5 = 21$ non-spade cards to choose from, thus there are $\binom{21}{10}$ ways to choose these. Multiplying these together, there are $\binom{5}{3}\binom{21}{10}$ possible hands where East has exactly 3 of the remaining spades. Thus the conditional probability that East has 3 spades given that North and South have 8 spades between them is:

$$\frac{\binom{5}{3}\binom{21}{10}}{\binom{26}{13}} \approx 0.339$$

1.7 Independence

Two events A and B are independent if the whether or not one event occurs is unaffected by whether or not the other event occurs. We will make this mathematically precise in a moment, but here are some intuitive examples:

1. You are flipping coins repeatedly. Each coin flip is independent of all other coin flips, since there is no way for coin flips to affect each other. Now imagine you flip 10 heads in a row. It is tempting to say that the streak of heads must break, and that it is more likely than not that the 11th flip is a tail. This is known as the *gambler's fallacy*, and is false since the flips are independent, and each flip has a 1/2 probability of heads. A famous example of this occurred in the Monte Carlo casino in Monaco on August 18, 1913. In a game of roulette, the ball landed on black 26 times in a row, and gamblers lost a lot of money betting that the streak would break.

2. In blackjack, if a player is dealt an ace, the next deal is less likely to be an ace, since the cards are drawn without replacement. Successive deals of cards are thus not independent. This is the basis for counting cards in blackjack and other games.

3. In the field of public health, "smoking" and "contracting lung cancer" are not independent since a causal link has been shown between the two.

We are ready to formally definite the concept of independence.

Let A and B be two events. A and B are *independent* if any of the following hold:

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$
$$\mathbb{P}(B|A) = \mathbb{P}(B)$$
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

In plain language, two events are independent if the conditional probability is the same as the unconditional probability, or if the probability of both events occurring is the same as the product of the probabilities of the individual events.

Let's look at a few examples of independence.

Example. Let us roll a single fair, six-sided die, and consider the following events:

- A: we roll an odd number
- B: we roll an even number
- C: we roll a 1 or a 2

1. Are A and B independent?

   Because $A \cap B = \emptyset$, $\mathbb{P}(A \cap B) = 0$, thus $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B) = 0$. But $\mathbb{P}(A) = 1/2$. Since $\mathbb{P}(A|B) \neq \mathbb{P}(A)$, A and B are not independent. We did not expect them to be independent, since if A occurs, B cannot occur, and vice versa.

2. Are A and C independent?

   Using the sample space reduction method, we see that $\mathbb{P}(A|C) = 1/2$, since if we know we roll a 1 or a 2, half the time we will roll an odd number. Since $\mathbb{P}(A) = 1/2$ as well, $\mathbb{P}(A|C) = \mathbb{P}(A)$, thus A and C are independent.

For our next problem we revisit our familiar two-coin-flip example.

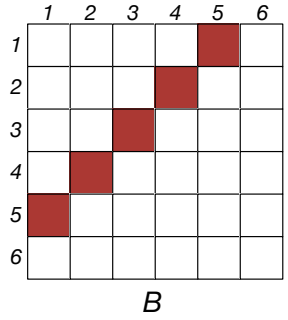
Example. Let us roll two fair, six-sided dice, and consider the following events:

- A: we roll a 4 on the first die

- B: the sum of the two dice is 6
- C: the sum of the two dice is 7

1. Are A and B independent?

   Here we use the "product of probabilities" test for independence. $\mathbb{P}(A \cap B) = 1/36$, since if we know the first die is 4 and the two dice sum to 6, the second die must be a 2; this corresponds to a single point in the sample space. $\mathbb{P}(A) = 1/6$, since a single die has a 1/6 chance of rolling any number. B comprises 5 simple events, as we can see graphically below; thus $\mathbb{P}(B) = 5/36$.



   Since $\mathbb{P}(A)\mathbb{P}(B) = (1/6)(5/36) = 5/216 \neq 1/36 = \mathbb{P}(A \cap B)$, these events are not independent. Why does this make sense? If we are interested in a sum of 6 on two dice, the outcome of the first throw matters; a sum of 6 is impossible if the first die is a 6, and is possible for all other rolls of the first die. Thus is makes sense that A and B should not be independent.

2. Are A and C independent?

   Here, we have again that $\mathbb{P}(A \cap C) = 1/36$, since if we know the first die is 4 and the two dice sum to 7, the second die must be a 3; $\mathbb{P}(A) = 1/6$ as above. We can see graphically that $\mathbb{P}(C) = 6/36 = 1/6$, since C comprises 6 simple events. Thus $\mathbb{P}(A)\mathbb{P}(C) = (1/6)(1/6) = 1/36 = \mathbb{P}(A \cap C)$, and so A and C are independent. Why does this make sense? No matter what we roll on the first die, there is an equal probability (1/6) of getting a sum of 7 when the second die is rolled.

1.8 Multiplicative and Additive Laws of Probability

The multiplicative and additive laws of probability give the probabilities of intersections and unions of events, and are important in constructing probabilities for more complicated events.

The multiplicative law of probability

Let A and B be two events. Then the probability of the intersection of the two events is:

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B|A) \\ &= \mathbb{P}(B)\mathbb{P}(A|B)\end{aligned}$$

If A and B are independent, then:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

This follows directly from the definition of conditional probability. We can extend this to find the probability of the intersection of multiple events. For example, for three events A, B, and C, we have:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B|A)\mathbb{P}(C|A \cap B)$$

We can extend this string of conditional probabilities to find the probability of the intersection of as many events as we want.

The additive law of probability gives the probability of the union of two events.

The additive law of probability

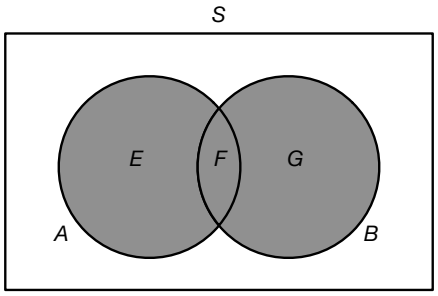
Let A and B be two events. Then the probability of the union of the two events is:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

If A and B are disjoint (mutually exclusive), i.e. $\mathbb{P}(A \cap B) = \emptyset$, then:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

To see this is true, consider the Venn diagram below.



1. $A \cup B$ (show in gray) is the disjoint union of E, F, and G, so by the additive property of probability, $\mathbb{P}(A \cup B) = \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G)$.
2. A is the disjoint union of E and F, so by $\mathbb{P}(A) = \mathbb{P}(E) + \mathbb{P}(F)$.
3. B is the disjoint union of F and G, so by $\mathbb{P}(B) = \mathbb{P}(F) + \mathbb{P}(G)$.
4. Addding together $\mathbb{P}(A)$ and $\mathbb{P}(B)$, and noting that $F = A \cap B$:

$$\begin{aligned}\mathbb{P}(A) + \mathbb{P}(B) &= \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(F) + \mathbb{P}(G) \\ &= \mathbb{P}(A \cup B) + \mathbb{P}(A \cap B)\end{aligned}$$

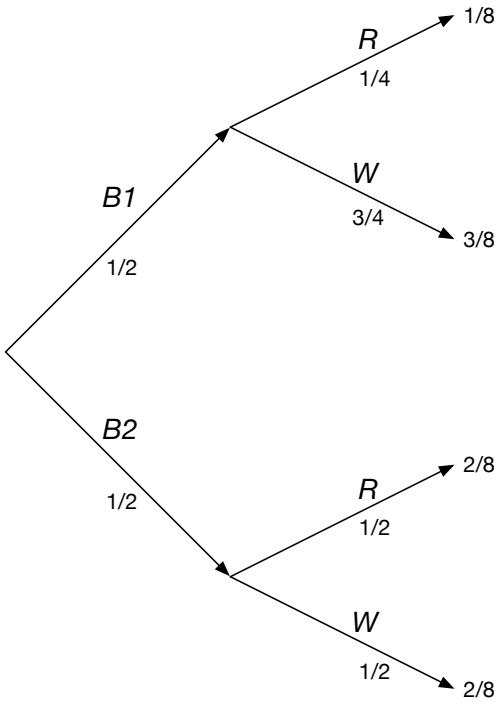
   Rearranging this, we get the additive rule.

Intuitively, if we add the probability of A and B, we are "double-counting" $A \cap B$, so we have to subtract it off to get $\mathbb{P}(A \cup B)$. This can be extended to unions of more than two sets, but the formulas are annoying and will not be useful to us in this course.
We can use the multiplication and addition rules in the following problem.

Example. You have two bags of balls. The first bag contains one red ball and three white balls. The second bag contains two red balls and two white balls. You choose a bag uniformly at random and then draw a ball from the chosen bag.

1. What is the probability that the ball you draw is red?

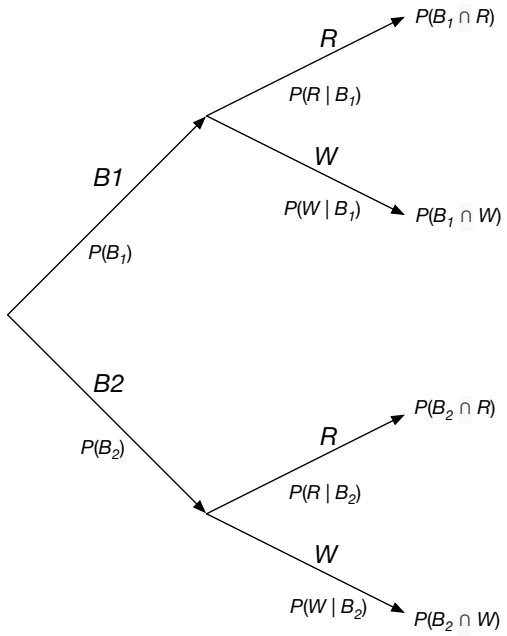
   We can do this directly using the multiplication and addition rules if we wish, but it is far easier to draw a *tree diagram*. Let B_1 be the event that we choose Bag 1, and B_2 the event that we choose Bag 2. Let R be the event that we draw a red ball, and W be the event that we draw a white ball. Then we have the following tree diagram.



Tree diagrams are really hard to explain in writing, but I will try. Starting at the node on the left, the first two edges of the tree are B_1 and B_2. Their respective probabilities are written below, and are both $1/2$, since the choice of bag is uniform. At the point, the tree branches again. Let's look at the top of the tree diagram, the part which follows B_1. The upper branch is labeled R, and indicates that the event R follows B_1. The probability listed there is the conditional probability of R given B_1, so $\mathbb{P}(R|B_1) = 1/4$. If we follow the arrows through B_1 and R, we reach the leaf of the tree which is labeled $1/8$. This leaf represents the event $B_1 \cap R$, since we passed through both of those events to get there. The probability $1/8$ listed there is the probability of $B_1 \cap R$. We get this by multiplying all the probabilities we passed through to get there. This follows the multiplication rule, since:

$$\mathbb{P}(B_1 \cap R) = \mathbb{P}(B_1)\mathbb{P}(R|B_1) = (1/2)(1/4) = 1/8$$

The interpretation of the other branches of the tree is similar. Here is a version of the tree diagram with all the probabilities (conditional and otherwise) labeled:



To find $\mathbb{P}(R)$, we just add up all the leaves that end passing through R. There are two such leaves, labeled with probabilities $1/8$ and $2/8$, thus $\mathbb{P}(R) = 1/8 + 2/8 = 3/8$.

In adding the probabilities above, we used the addition rule. To see how we used it, recall that the sample space is designated $\mathcal{S}$ and that B_1 and B_2 are disjoint with $B_1 \cup B_2 = \mathcal{S}$ (you can only choose one bag, and you must choose a bag):

$$\begin{aligned} R &= R \cap \mathcal{S} && \text{intersecting with } \mathcal{S} \text{ does nothing, since } A \subset \mathcal{S} \\ &= R \cap (B_1 \cup B_2) \\ &= (R \cap B_1) \cup (R \cap B_2) && \text{by the distributive law} \end{aligned}$$

Then since $R \cap B_1$ and $R \cap B_2$ are disjoint (why is this the case?), we can use the addition rule to get:

$$\mathbb{P}(R) = \mathbb{P}(R \cap B_1) + \mathbb{P}(R \cap B_2) = 1/8 + 2/8 = 3/8$$

2. What is the probability we chose Bag 1 in the first step given that we draw a red ball?

   Here we are looking for the conditional probability $\mathbb{P}(B_1|R)$. We computed $\mathbb{P}(R)$ in the first part, and we can read off $\mathbb{P}(B_1 \cap R) = 1/8$ from the tree. Thus by the definition of conditional expectation:

$$\mathbb{P}(B_1|R) = \frac{\mathbb{P}(B_1 \cap R)}{\mathbb{P}(R)} = \frac{1/8}{3/8} = \frac{1}{3}$$

1.9 Bayes' Rule and the Law of Total Probability

Given two events A and B, Bayes' rule is a mathematical formula for relating the two conditional probabilities $\mathbb{P}(A|B)$ and $P(B|A)$. In it's simplest form, Bayes' rule may be stated as follows:

Bayes' rule

Let A and B be two events. Then:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

where $\mathbb{P}(B) \neq 0$.

To see this is true, we can multiply both sides by $\mathbb{P}(B)$ to get:

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

Both sides of this equation are equal to $\mathbb{P}(A \cap B)$ by the definition of conditional probability, thus the statement of Bayes' rule is true. Let's return to the drawing-balls-from-bags example, and use Bayes' rule on it.

Example. You have two bags of balls. The first bag contains one red ball and three white balls. The second bag contains two red balls and two white balls. You choose a bag uniformly at random and then draw a ball from the chosen bag. What is the probability that you choose the Bag 1 in the first step given that the ball drawn is red?

Using the same events as above, we are looking for conditional probability $\mathbb{P}(B_1|R)$. For the opposite conditional probability, $\mathbb{P}(R|B_1) = 1/4$ since we know exactly what happens if we draw from Bag 1. $\mathbb{P}(B_1) = 1/2$, since we are choosing the bags uniformly at random. To use Bayes' rule, the only remaining probability we need is $\mathbb{P}(R)$, which is harder to compute.

Luckily we found it above using the tree diagram, so we know $\mathbb{P}(R) = 3/8$. (We will learn a formula later which gets around this pesky probability.) Thus we plug everything into Bayes' rule to get:

$$\mathbb{P}(B_1|R) = \frac{\mathbb{P}(R|B_1)\mathbb{P}(B_1)}{\mathbb{P}(R)} = \frac{(1/4)(1/2)}{3/8} = \frac{1}{3}$$

which agrees what we obtained above.

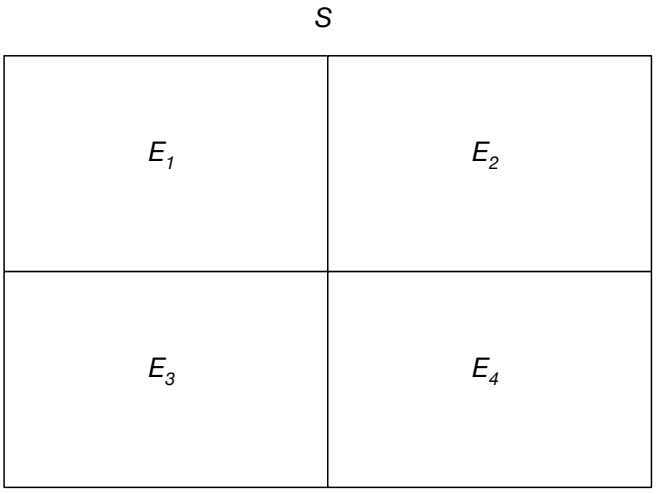
Bayes' rule is most often used in conjunction with the *Law of Total Probability*, which we state below. First we need to define a *partition* of a sample space.

Partition of a sample space

A *partition* of a sample space $\mathcal{S}$ is a finite collection of disjoint subsets of $\mathcal{S}$ whose union is $\mathcal{S}$. Intuitively, a partition of $\mathcal{S}$ is a "division of $\mathcal{S}$ into separate pieces". Mathematically, a partition of $\mathcal{S}$ is a collection $\{E_1, E_2, \ldots, E_k\}$ of subsets of $\mathcal{S}$ such that:

1. $\mathcal{S} = E_1 \cup E_2 \cup \cdots \cup E_k$
2. $E_i \cap E_j = \emptyset$ for $i \neq j$

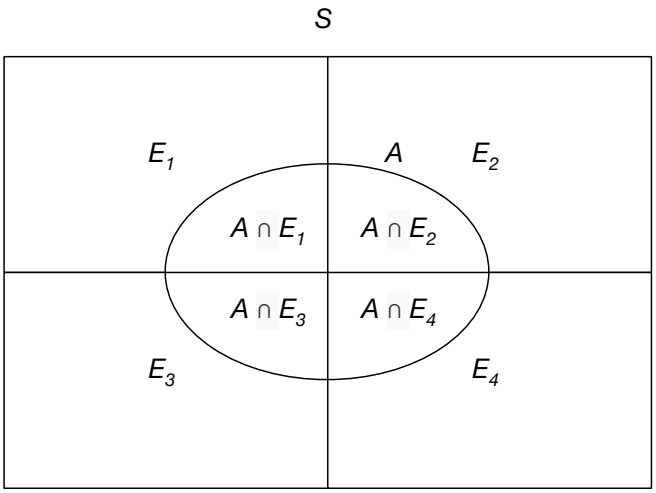
For any event A, the collection $\{A, A^c\}$ is always a partition of $\mathcal{S}$ (can you see why this is the case?) For a finite sample space, the collection of all simple events is also a partition. Here is a partition of $\mathcal{S}$ into 4 subsets illustrated graphically.



If A is any event, and $\{E_1, E_2, \ldots, E_k\}$ is a partition of $\mathcal{S}$, then we can decompose A according to the partition by:

$$A = (A \cap E_1) \cup (A \cap E_2) \cup \cdots \cup (A \cap E_k)$$

where this union consists of disjoint sets (since the partition consists of disjoint sets). For $k = 4$, this is illustrated graphically below:



We are now ready to state the Law of Total Probability.

Law of total probability

Let $\{E_1, E_2, \ldots, E_k\}$ be a partition of $\mathcal{S}$ such that $\mathbb{P}(E_i) > 0$ for all i. Then for any event A, we can decompose the probability of A according to our partition as:

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A|E_i)\mathbb{P}(E_i)$$

To see this is true, first we decompose our event A as we did above into disjoint sets:

$$A = (A \cap E_1) \cup (A \cap E_2) \cup \cdots \cup (A \cap E_k)$$

Then we can write:

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}((A \cap E_1) \cup (A \cap E_2) \cup \cdots \cup (A \cap E_k)) \\
&= \mathbb{P}(A \cap E_1) + \mathbb{P}(A \cap E_2) + \cdots + \mathbb{P}(A \cap E_k) \\
&= \mathbb{P}(A|E_1)\mathbb{P}(E_1) + \mathbb{P}(A|E_2)\mathbb{P}(E_2) + \cdots + \mathbb{P}(A|E_k)\mathbb{P}(E_k)
\end{aligned}$$

where we have used the fact that the decomposition of A is a union of disjoint sets, and have also used the definition of conditional probability.

We can now combine Bayes' rule and the Law of Total Probability to get another version of Bayes' rule.

Bayes' rule, total probability version

Let A and B be two events, and let $\{E_1, E_2, \ldots, E_k\}$ be a partition of $\mathcal{S}$ such that $\mathbb{P}(E_i) > 0$ for all i. Then:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\sum_{i=1}^{k} \mathbb{P}(B|E_i)\mathbb{P}(E_i)}$$

where $\mathbb{P}(B) \neq 0$.

To get this, take Bayes' rule and expand $\mathbb{P}(B)$ in the denominator using our partition and the Law of Total Probability.

Let's revisit the drawing-balls-from-bags example one final time, using the total probability version of Bayes' rule.

Example. You have two bags of balls. The first bag contains one red ball and three white balls. The second bag contains two red balls and two white balls. You choose a bag uniformly at random and then draw a ball from the chosen bag. What is the probability that you choose the Bag 1 in the first step given that the ball drawn is red?

Using the same events as in the above two versions, we are once again looking for conditional probability $\mathbb{P}(B_1|R)$. Our partition is $\{B_1, B_2\}$ since one of these two events must occur (so the union is $\mathcal{S}$), but both cannot occur (so they are disjoint). Writing out Bayes' rule, we get:

$$\mathbb{P}(B_1|R) = \frac{\mathbb{P}(R|B_1)\mathbb{P}(B_1)}{\mathbb{P}(R|B_1)\mathbb{P}(B_1) + \mathbb{P}(R|B_2)\mathbb{P}(B_2)}$$

The probabilities involved in this are straightforward:

1. $\mathbb{P}(B_1) = \mathbb{P}(B_2) = 1/2$
2. $\mathbb{P}(R|B_1) = 1/4$
3. $\mathbb{P}(R|B_2) = 1/2$

So we just plug these in to get:

$$\mathbb{P}(B_1|R) = \frac{(1/4)(1/2)}{(1/4)(1/2) + (1/2)(1/2)} = \frac{1/8}{3/8} = \frac{1}{3}$$

which agrees with both versions above!

Example. Consider the following card game. We have 3 cards that are identical in shape and size. The cards are colored as follows:

1. both sides are red

2. both sides are black

3. one side is red and the other side is black

The three cards are shuffled and one is randomly chosen and it put down on a table. If the face-up side of the chosen card is red, what is the probability that the other side is black?

Let RR, BB, and RB denote the events that the chosen card is all red, all black, or red-black (respectively). Let R be the event that the face-up side of the chosen card is red. Note that $\{RR, BB, RB\}$ is a partition of our sample space. We are interested in whether the red-black card has been chosen, i.e. $\mathbb{P}(RB|R)$. First we write out the total probability version of Bayes' rule, then we plug in the probabilities we know:

$$\begin{aligned}
\mathbb{P}(RB|R) &= \frac{\mathbb{P}(R|RB)\mathbb{P}(RB)}{\mathbb{P}(R|RR)\mathbb{P}(RR) + \mathbb{P}(R|BB)\mathbb{P}(BB) + \mathbb{P}(R|RB)\mathbb{P}(RB)} \\
&= \frac{(1/2)(1/3)}{(1)(1/3) + 0(1/3) + (1/2)(1/3)} \\
&= \frac{1/6}{1/2} \\
&= \frac{1}{3}
\end{aligned}$$

Thus the probability that the other side is black is only 1/3. It is tempting to guess that that the probability of the other side being black is 1/2 since there are two possible cards it could be (all red and red-black), and they should be equally likely. It turns out that this is not the case! The two cards are not equally likely. Here is one way to see this. Instead of considering three cards, let's think of them as six sides. We can label the six sides as:

1. R_1 and R_2: the two red sides of the all-red card

2. B_1 and B_2: the two black sides of the all-black card

3. R_3 and B_3: the red and black sides of the red-black card

If red is face up, it must be R_1, R_2, or R_3, and these are equally likely. For R_1 and R_2, the other side of the card is red. Only for R_3 is the other side black. Thus the probability of the other side of the card being black is 1/3, which agrees with the result from Bayes' theorem.

Here is another application of Bayes' theorem which concerns medical testing. This is a classic example in the field of public health.

Example. The OraQuick HIV test is a rapid test for HIV, which can give a result in 20 minutes. The test has the following properties:

1. Sensitivity (probability of a positive test given that you have the disease): 0.996 (99.6%)

2. Specificity (probability of a negative test given that you don't have the disease): 0.999 (99.9%)

What is the positive predictive value of the test, i.e. the probability that a patient has HIV given that the OraQuick test is positive[8].

Let T be the event that the test is positive, and H the event that the patient has HIV. We are interested in the positive predictive value, which is $\mathbb{P}(H|T)$. Let's write out Bayes' rule:

$$\mathbb{P}(H|T) = \frac{\mathbb{P}(T|H)\mathbb{P}(H)}{\mathbb{P}(T)}$$

We cannot possibly know the denominator $\mathbb{P}(T)$ (the probability of a positive test), so we will use the total probability of Bayes' theorem, with partition $\{H, H^c\}$.

$$\mathbb{P}(H|T) = \frac{\mathbb{P}(T|H)\mathbb{P}(H)}{\mathbb{P}(T|H)\mathbb{P}(H) + \mathbb{P}(T|H^c)\mathbb{P}(H^c)}$$

What do we know? $\mathbb{P}(T|H)$ is the sensitivity, so that is 0.996. We don't know $\mathbb{P}(T|H^c)$, but we do know the specificity, which is $\mathbb{P}(T^c|H^c) = 0.999$. Since $\mathbb{P}(T|H^c) + \mathbb{P}(T^c|H^c) = 1$ (either the test is positive or it's not), $\mathbb{P}(T|H^c) = 1 - \mathbb{P}(T^c|H^c) = 1 - 0.999 = 0.001$. All that remains is $\mathbb{P}(H)$. ($\mathbb{P}(H^c) = 1 - \mathbb{P}(H)$, so we are all set once we have $\mathbb{P}(H)$.) What is $\mathbb{P}(H)$? That is the probability that a randomly-selected patient has HIV, what is known in public health parlance and the disease prevalence. This can be a hard quantity to determine and is the crux of this entire problem.

The CDC estimates that 1.2 million people in the US have HIV[9], out of a total population of 319 million. That puts the prevalence of HIV at 0.00376. So we will take $\mathbb{P}(H) = 0.00376$ for now, which gives us $\mathbb{P}(H^c) = 1 - 0.00376$. Plugging everything into Bayes' theorem:

$$\mathbb{P}(H|T) = \frac{(0.996)(0.00376)}{(0.996)(0.00376) + (0.001)(1 - 0.00376)} = 0.79$$

So we get a positive test in 80% of the cases, which is decent but not great. Let's repeat this with a different value for the prevalence of HIV. In Fulton County, GA (which contains Atlanta) the prevalence of HIV is approximately 0.0123 (significantly higher than for the nation as a whole). This gives us a positive predictive value of:

$$\mathbb{P}(H|T) = \frac{(0.996)(0.0123)}{(0.996)(0.0123) + (0.001)(1 - 0.0123)} = 0.925$$

which is higher. The take-home message here for public health is the following. The positive predictive value of a test depends on the prevalence of the disease you are testing for; it increases as a disease becomes more common, i.e. increases with disease prevalence. Even the most sensitive and specific test does a poor job of detecting rare diseases.

[8]Sensitivity, specificity, and positive predictive value are public health terms; you don't have to know them. Data here is from the CDC website.

[9]This estimate is for the end of 2012, but is the most recent CDC data I can find.

2 Discrete Random Variables

2.1 Random Variables

A *random variable* is a real-valued function on a sample space[10]. A random variable generally is a quantity we wish to measure; the output of the random variable depends on which element of the sample space is chosen. In this section, we will look at discrete random variables.

> A *discrete* random variable Y is a random variable which can only take on a finite or countable set of distinct values.

Here are some examples of discrete random variables:

1. The number of voters in Providence, Rhode Island who prefer Jorge Elorza.
2. The number of defective light bulbs out of a shipment of 1000 light bulbs.
3. The number of times I play a slot machine in Las Vegas until I win.

We will use the following simple example to illustrate features of discrete random variables.

Example. Let $\mathcal{S}$ be the sample space representing the flip of two fair coins. Let Y be the number of heads flipped. Then Y is a discrete random variable, since it can only have the values 0, 1, or 2. We can illustrate it graphically below.

	H	*T*
H	2	1
T	1	0

Uppercase letters, such as Y, are used to designate random variables. We use lowercase letters, such as y, to designate a value that a random variable can take. The expression $(Y = y)$ is shorthand for the set of all points in our sample space $\mathcal{S}$ for which the random variable Y outputs the value y. Since $(Y = y)$ is a subset of $\mathcal{S}$, it is an event in our sample space. In the two-coin-toss problem, for example, the possible values of Y are 0, 1, and 2, so we have:

- $(Y = 0) = \{(T, T)\}$
- $(Y = 1) = \{(H, T), (T, H)\}$
- $(Y = 2) = \{(H, H)\}$

[10] It's not really a variable at all, but we are stuck with the terminology.

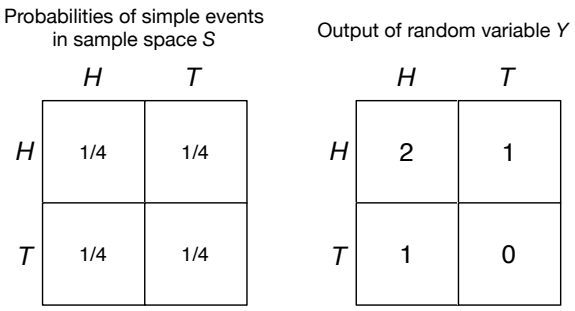
Since $(Y = y)$ is an event in our sample space, we can talk about its probability, i.e. $\mathbb{P}(Y = y)$. In fact, the point of random variables is to do just this!

Probability of a discrete random variable

The probability that a discrete random variable Y takes the value y, denoted $\mathbb{P}(Y = y)$, or $p(y)$ for short, is the probability of the event $(Y = y)$, the set of all points in the sample space $\mathcal{S}$ which output the value y.

$\mathbb{P}(Y = y)$ is the sum of the probabilities of all the simple events in $\mathcal{S}$ which are assigned the value y by the random variable Y.

Back to our two-coin-toss problem, let's look at the probabilities of the random variable Y. For convenience, here is a picture of the sample space probabilities next to a graphical representation of Y. Since we are using the discrete uniform distribution for coin tosses, each simple event in our sample space has probability 1/4.



Using the rule above, we can compute the following probabilities for Y by adding up the probabilities of the underlying simple events. In this case, since we are using the discrete uniform distribution, we could also just count the number of simple events which lead to each output of Y and divide by 4, the size of the sample space.

- $\mathbb{P}(Y = 0) = 1/4$
- $\mathbb{P}(Y = 1) = 1/4 + 1/4 = 1/2$
- $\mathbb{P}(Y = 2) = 1/4$

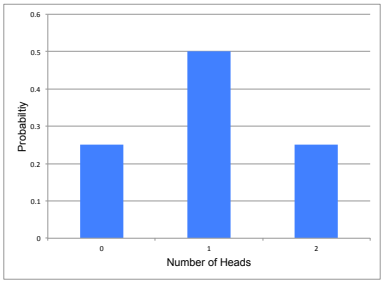
Probability mass function

The *probability mass function (pmf)* is a function which gives the probability that a discrete random variable Y is exactly equal to a value y. The pmf can be represented as a function, table, or graph which gives the values $p(y) = \mathbb{P}(Y = y)$ for all possible values y which Y can take.

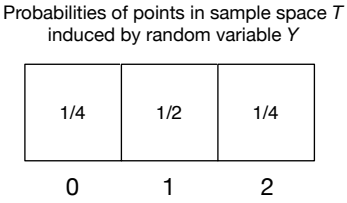
In the two-coin-flip example, we can represent the pmf of Y in a table:

y	$p(y)$
0	1/4
1	1/2
2	1/4

We can also represent the pmf graphically as a histogram[11].



A discrete random variable induces a probability distribution on the sample space of all possible values the random variable can take. This is a different sample space from the original sample space. Back to our two-coin-flip example, the random variable Y induces a probability distribution on a new sample space $\mathcal{T} = \{0, 1, 2\}$. The probabilities of the sample points in $\mathcal{T}$ are the probabilities $p(y)$ for $y = 0, 1, 2$. We can illustrate this new sample space in a picture.



Often (as we shall see), we care much more about the sample space induced by a random variable than the underlying sample space. Since a discrete random variable induces a probability distribution, the following must be true.

For any discrete random variable Y:

$$0 \leq p(y) \leq 1 \text{ for all } y$$

$$\sum_{\text{all } y} p(y) = 1$$

[11] I will undoubtedly lose some of my math street cred if I admit to using Microsoft Excel for these histograms, but in some cases it really is the easiest tool to use.

where $p(y) = \mathbb{P}(Y = y)$. Since we have a discrete sample space, the sum is finite or countable.

Let's look at two more examples, this time involving the rolls of two six-sided dice.

Example. Let $\mathcal{S}$ be the sample space representing the rolls of two six-sided dice. Consider the following two random variables:

1. X = the sum of the two dice
2. Y = the larger of the two die rolls (if they are the same, then it's just equal to both die rolls)

Let's look at these random variables graphically.

Two random variables on the sample space of two die rolls

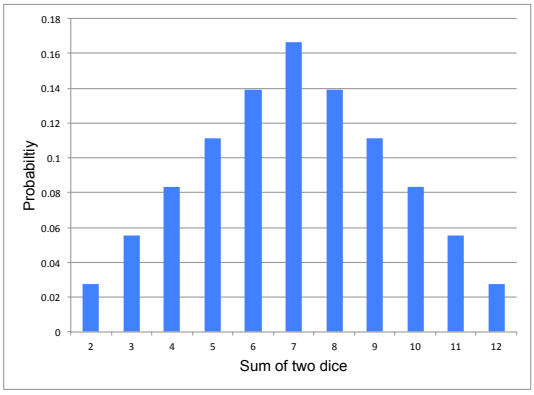
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

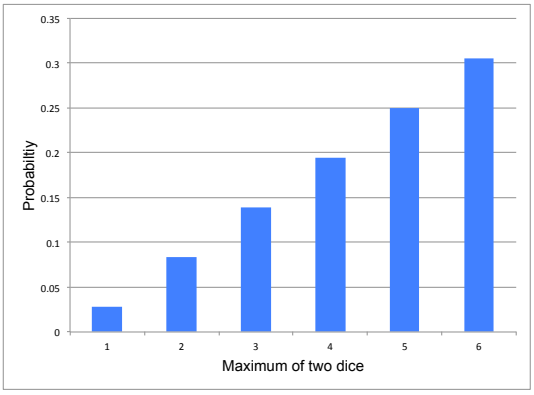
X

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

Y

The random variable X induces a probability distribution on the set of integers $\{2, 3, 4, \ldots, 12\}$, and the random variable Y induces a probability distribution on the set of integers $\{1, 2, 3, 4, 5, 6\}$. Let's look at the pmfs of both random variables using histograms.





Both of these distributions are nonuniform, even though the underlying distribution of the two dice is uniform. The first distribution, that of the random variable X, is a familiar one

to aficionados of board games such at Settlers of Catan and Monopoly. We can also write the pmfs in table form. For the random variable Y, we have:

y	$p(y)$
1	1/36
2	3/36
3	5/36
4	7/36
5	9/36
6	11/36

The pmf for X can be expressed similarly.

2.2 Expected Value

Given a discrete random variable, we can definite its mean, or expected value.

Expected value of a random variable

For a discrete random variable Y with probability function $p(y)$, the *expected value* or *mean* is defined to be

$$\mathbb{E}(Y) = \sum_{\text{all } y} y\, p(y)$$

where the sum is taken over all possible values y can take. We can think of the expected value as a weighted average of the values of Y with each possible output y weighted by its probability $p(y)$. The expected value is sometimes written as μ (for mean)

Here is one interpretation of the expected value of a random variable. Think of a random variable Y as an observation from an experiment. Suppose we perform the experiment n times, and observe n values of Y, which we shall designate $y_1, y_2, \ldots, y_n$. Then for large n,

$$\frac{y_1 + y_2 + \cdots + y_n}{n} \approx \mathbb{E}(Y)$$

where the approximation "gets better" as n gets larger, i.e. as we perform more experiments. The quantity on the left hand side is known as the *empirical mean* or *sample mean* and looks like what we likely learned in high school (add a bunch of stuff up and divide by the number of things). The expected value is, in a sense, the limit of the empirical mean as the sample size approaches infinity. We will make this more precise later in the course, but this is a good concept to keep in mind.

Example. Let X represent the roll of a standard, fair six-sided die. (In this case, the underlying sample space is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ with the discrete uniform distribution, and

the random variable X is the same as the sample space element selected.) Then the expected value of X is:

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=1}^{6} x\mathbb{P}(X=x) \\ &= \sum_{x=1}^{6} x\frac{1}{6} \\ &= \frac{1}{6}\sum_{x=1}^{6} x \\ &= \frac{21}{6} \\ &= 3.5\end{aligned}$$

where used the fact from the discrete uniform distribution that $\mathbb{P}(X=x) = 1/6$ for all x. Note that the expected value of 3.5 is not a possible value of X, i.e. we cannot roll a 3.5 on a single die. Given our "long term average" interpretation, this is saying that we expect the empirical average to approach 3.5 as the number of rolls increases, not that a 3.5 is the most likely die roll.

Example. Let Y be the random variable above representing the maximum of two dice. What is the expected value of Y.

To find the expected value, we do a weighted average using the probabilities in the table above.

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{y=1}^{6} y\mathbb{P}(Y=y) \\ &= 1\cdot\frac{1}{36} + 2\cdot\frac{3}{36} + 3\cdot\frac{5}{36} + 4\cdot\frac{7}{36} + 5\cdot\frac{9}{36} + 6\cdot\frac{11}{36} \\ &= \frac{1+6+15+28+45+66}{36} \\ &= \frac{161}{36} \approx 4.47\end{aligned}$$

2.3 Properties of Expectation

We will discuss several properties of the expected value. The first and and one of the most important is the *linearity of expectation*.

Linearity of expectation

Let X and Y be two random variables[12], and let a and b be constants. Then

$$\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

This is called *linearity* in reference to linear algebra, i.e. we can separate addition and pull out constants. This holds whether or not X and Y are independent.

As a corollary of this, if we have random variables $X_1, X_2, \ldots, X_n$, then

$$\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)$$

Linearity of expectation is a really nice property since it does not require the random variables to be independent. Let's do a problem to illustrate the usefulness of linearity of expectation.

Example. One evening, n customers dine at a restaurant. Each gives their hat to a hat-check person at a restaurant. (These are fashionable diners!) After dinner, the hat-check person gives the hats back to the customers in a random order, i.e. each customer receives one of the hats uniformly at random. What is the expected number of customers that get their own hat back?

Let X be the number of customers who get their own hat back. Then X is a discrete random variable taking values $0, 1, \ldots, n$. By the definition of expected value:

$$\mathbb{E}(X) = \sum_{i=1}^{n} i\, \mathbb{P}(X = i)$$

At this point, we have a mess! We have to compute $\mathbb{P}(X = i)$ for all i, which would involve a sophisticated combinatorial argument, as well as considerable time and mental energy. Luckily for us, there is another way.

We will use the method of indicator random variables, together with linearity of expectation, to solve this problem. An *indicator random variable* is a random variable I which only takes the values 0 and 1. It is used to indicate whether (or not) an event takes place: $I = 1$ if the event happens, and $I = 0$ if the event does not happen.

We will define some indicator random variables for this problem. First, let's number the customers $1, 2, \ldots, n$. For $i = 1, \ldots, n$, let X_i be the indicator random variable for the event that customer i gets their own hat back. In other words,

$$X_i = \begin{cases} 1 & \text{customer } i \text{ gets their own hat back} \\ 0 & \text{otherwise} \end{cases}$$

From the way we have constructed these indicator random variables, we see that

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

[12] So far we have only discussed what an expected value is in terms of discrete random variables, but this is true for all random variables.

Does this make sense? If we add up the indicator random variables, we are adding a 1 whenever a customer gets their own hat back, which gives us the total number of customers who get their hat back. Now we use linearity of expectation. The indicator variables X_i are not independent, since, for example, if customers $1, 2, \ldots, n-1$ all get their own hat back, then customer n must also get their own hat back. But that doesn't matter, since linearity of expectation does not require independence.

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(X_1 + X_2 + \cdots + X_n) \\ &= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)\end{aligned}$$

If we can compute the expected value of the indicator random variables, we are all set. By the definition of expected value:

$$\begin{aligned}\mathbb{E}(X_i) &= 0 \cdot \mathbb{P}(X_i = 0) + 1 \cdot \mathbb{P}(X_i = 1) \\ &= \mathbb{P}(X_i = 1)\end{aligned}$$

$\mathbb{P}(X_i = 1)$ is the probability that customer i gets their own hat back. Since the hats are distributed uniformly at random, and there are n hats to distribute, we must have $\mathbb{P}(X_i = 1) = 1/n$. Thus,

$$\mathbb{E}(X_i) = \frac{1}{n} \quad \text{for } i = 1, 2, \cdots, n$$

Note that this does *not* depend on i, i.e. this is exactly the same for all n customers. Substituting this above:

$$\begin{aligned}\mathbb{E}(X) &= \frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n} \qquad\qquad n \text{ terms in this sum} \\ &= 1\end{aligned}$$

The key to this method is that $\mathbb{E}(X_i)$ does not depend on i. Why does this make sense? I like to think of this in terms of symmetry. We number the customers for convenience, but mathematically there is no distinction between the n customers[13]. Imagine the customers lining up to leave the restaurant. The person at the front of the line is handed a hat uniformly at random, then the customer leaves. This is repeated until all customers have left. If we swap any two customers in line, nothing should change. The expected number of customers who receive their own hat should remain the same.

How do you know when to use the method of indicator random variables and linearity of expectation? Like everything in probability, it is difficult to come up with hard-and-fast rules for when to use a given tool. That being said, here are a few guidelines for when this method is useful:

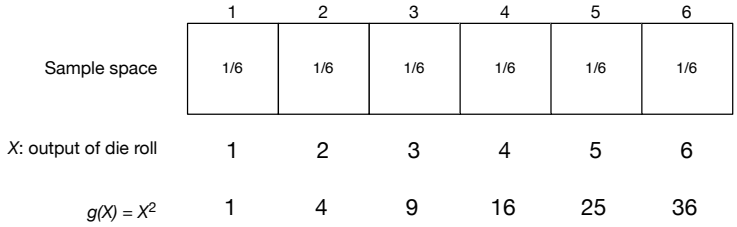
1. You are looking for an *expected value* involving a group of people or objects (not a probability).

2. You have a symmetric group of people or objects, i.e. you can swap them around without affecting the result.

[13] If you are the restaurant proprietor, don't tell them this!

We can also talk about functions of random variables. If Y is a random variable and $g(y)$ is a real-valued function, then $g(Y)$ is another random variable (since $g(Y)$ is also a function on our sample space). To evaluate $g(Y)$, just take the output of Y and run it through g.

Example. Let X be the output of a standard six-sided die, and let $g(x) = x^2$. Then $g(X)$ is also a discrete random variable. First, let's show X and $g(X)$ graphically:



There are multiple ways to write the pmf for $g(X)$ in a table. I think the most useful way is to take the pmf table for X and add an extra column for $g(X)$. We will see why this makes sense shortly.

x	$g(x)$	$p(x)$
1	1	1/6
2	4	1/6
3	9	1/6
4	16	1/6
5	25	1/6
6	36	1/6

The expected value of a function of discrete random variable is computed in a similar fashion to the expected value of a random variable.

Expected value of a function of a discrete random variable

Let Y be a discrete random variable with probability function $p(y)$, and let g be a real-valued function. Then the expected value of the random variable $g(Y)$ is given by:

$$\mathbb{E}[g(Y)] = \sum_{\text{all } y} g(y)\, p(y)$$

where the sum is taken over all possible values y can take. Note that the *only* difference from the expected value of Y is that we have $g(y)$ in place of y in the sum. The probabilities $p(y)$ are *unchanged*. This is again a weighted average, but this time we are taking a weighted average of the possible values of $g(Y)$.

Example. Let X be the output of a standard six-sided die, and let $g(x) = x^2$. Find $E[g(X)]$.

We can use the formula for the expected value of a function of a random variable, together with the pmf table above to get:

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_{\text{all } x} g(x)\, p(x) \\ &= 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 9 \cdot \frac{1}{6} + 16 \cdot \frac{1}{6} + 25 \cdot \frac{1}{6} + 36 \cdot \frac{1}{6} \\ &= \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) \\ &= \frac{91}{6} \approx 15.17\end{aligned}$$

Before we go on, we mention one more property of expected value: the expected value of a constant.

Expected value of a constant

Let c be a constant. Then $\mathbb{E}(c) = c$. In other words, the expected value of a constant is just the constant itself.

Combining this with the linearity of expectation, we get the following expression for the expected value of a shifted and scaled random variable.

Expected value of $aY + b$

If Y is a random variable, and a and b are constants, then

$$\mathbb{E}(aY + b) = a\mathbb{E}(Y) + b$$

2.4 Variance

So far we have seen one quantitative descriptor of the distribution of a random variable: the expected value. In this section, we discuss another descriptor: the variance. To begin, let's compare the pmfs of two discrete random variables.

Example. Let X be the number of heads observed when a fair coin is flipped 6 times, and let Y be the uniform distribution on the integers $\{0, 1, 2, 3, 4, 5, 6\}$. Since Y is the uniform distribution on a finite set with 7 elements, $\mathbb{P}(Y = y) = 1/7$ for all Y. X takes the same values as Y since the number of heads observed in 6 coin flips is an integer between 0 and 6. What is $\mathbb{P}(X = x)$? Since there are two possibilities for each flip, there are 2^6 possible

flip-sequences. The number of flip-sequences which give us x heads is given by the binomial coefficient $\binom{6}{x}$ (why is this true?). Thus we have:

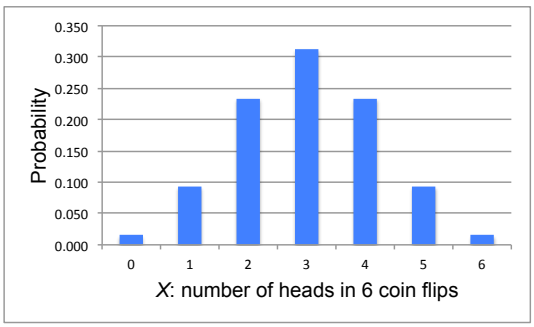
$$\mathbb{P}(X = x) = \frac{\binom{6}{x}}{2^6}$$

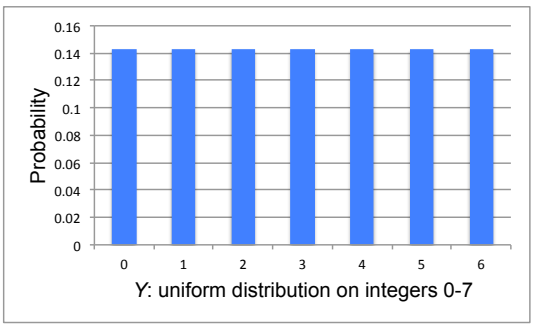
Putting both pmfs in tables we get:

x	$p(x)$
0	1/64
1	6/64
2	15/64
3	20/64
4	15/64
5	6/64
6	1/64

y	$p(y)$
0	1/7
1	1/7
2	1/7
3	1/7
4	1/7
5	1/7
6	1/7

Now let's look at the pmfs as histograms:





Both of these two random variables have an expected value (mean) of 3. (You can compute it yourself if you want, or reason that both distributions are symmetric, and the "middle" is 3.) Despite having the same mean, the two distributions are very different. X is relatively centered about the mean, while Y is all "spread out". We want a way to quantify this amount of "spread". There are many ways to do this. For example, we could use the

average distance from the mean. Statisticians have settled on a slightly different measure of "spread": the variance. In words, the variance of a random variable is the expected value of the squared-difference from the mean. Mathematically, we have the following definition:

Variance of a discrete random variable

Let Y be a discrete random variable with probability function $p(y)$, and let $\mu = \mathbb{E}(Y)$ be its expected value (mean). Then we define the *variance* of Y by

$$Var(Y) = \mathbb{E}[(Y-\mu)^2]$$

Using the formula for the expected value of a function of a random variable, this becomes

$$Var(Y) = \sum_{\text{all } y} (y-\mu)^2 \, p(y)$$

The variance is sometimes denoted by the symbol σ^2. The *standard deviation* of a random variable is the square root of the variance, and is denoted σ.

Let's compute the variance of the two random variables above. Since Y is more "spread out" than X, we expect it to have a higher variance.

Example. Let X be the number of heads observed when a fair coin is flipped 6 times, and let Y be the uniform distribution on the integers $\{0,1,2,3,4,5,6\}$. Compute the variances of both random variables.

The mean of both random variables is 3, so we have:

$$\begin{aligned} Var(X) &= \sum_{\text{all } x} (x-3)^2 \, p(x) \\ &= (0-3)^2\frac{1}{64} + (1-3)^2\frac{6}{64} + (2-3)^2\frac{15}{64} + (3-3)^2\frac{20}{64} + (4-3)^2\frac{15}{64} + (5-3)^2\frac{6}{64} + (6-3)^2\frac{1}{64} \\ &= \frac{(9)(1)}{64} + \frac{(4)(6)}{64} + \frac{(1)(15)}{64} + \frac{(0)(20)}{64} + \frac{(1)(15)}{64} + \frac{(4)(6)}{64} + \frac{(9)(1)}{64} \\ &= \frac{96}{64} = \frac{3}{2} \end{aligned}$$

$$\begin{aligned} Var(Y) &= \sum_{\text{all } y} (y-3)^2 \, p(y) \\ &= (0-3)^2\frac{1}{7} + (1-3)^2\frac{1}{7} + (2-3)^2\frac{1}{7} + (4-3)^2\frac{1}{7} + (4-3)^2\frac{1}{7} + (5-3)^2\frac{1}{7} + (6-3)^2\frac{1}{7} \\ &= \frac{1}{7}(9+4+1+0+1+4+9) \\ &= \frac{28}{7} = 4 \end{aligned}$$

As predicted from the histograms, Y has a much higher variance.

The variance is often not computed by using the definition above, but by using a formula affectionately known as the *Magic Variance Formula*.

Magic Variance Formula

Let Y be a random variable and let $\mu = \mathbb{E}(Y)$ be its expected value (mean). Then the variance of Y is given by

$$Var(Y) = \mathbb{E}[Y^2] - \mu^2$$

To verify this formula, we expand out the binomial in the definition of variance, and use linearity of expectation.

$$\begin{aligned} Var(Y) &= \mathbb{E}[(Y-\mu)^2] \\ &= \mathbb{E}(Y^2 - 2\mu Y + \mu^2) \\ &= \mathbb{E}(Y^2) - 2\mu\mathbb{E}(Y) + \mathbb{E}(\mu^2) \\ &= \mathbb{E}(Y^2) - 2\mu^2 + \mu^2 \\ &= \mathbb{E}(Y^2) - \mu^2 \end{aligned}$$

where we used the fact that $\mathbb{E}(Y) = \mu$, and the expected value of a constant is just the constant.

Variance behaves much differently from expectation. Unlike expectation, it is *not* linear. For a random variable Y, we have the following expression for the variance of $aY + b$.

Variance of $aY + b$

If Y is a random variable, and a and b are constants, then

$$Var(aY + b) = a^2 Var(Y)$$

To see this, we use the Magic Variance Formula. First we compute $\mathbb{E}[(aY+b)^2]$.

$$\begin{aligned} \mathbb{E}[(aY+b)^2] &= \mathbb{E}(a^2Y^2 + 2abY + b^2) \\ &= a^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(Y) + b^2 \end{aligned}$$

where we used the linearity of expectation and the fact that the expected value of a constant is the constant itself. Next, we compute $[\mathbb{E}(aY+b)]^2$. Once again using linearity of expectation, we get:

$$\begin{aligned} [\mathbb{E}(aY+b)]^2 &= (a\mathbb{E}(Y) + b)^2 \\ &= a^2\mathbb{E}(Y)^2 + 2b\mathbb{E}(Y) + b^2 \end{aligned}$$

Using the Magic Variance Formula, we subtract these two to get:

$$\begin{aligned} Var(aY+b) &= \mathbb{E}[(aY+b)^2] - [\mathbb{E}(aY+b)]^2 \\ &= \left(a^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(Y) + b^2\right) - \left(a^2\mathbb{E}(Y)^2 + 2b\mathbb{E}(Y) + b^2\right) \\ &= a^2\left(E(Y^2) - [E(Y)]^2\right) \\ &= a^2 Var(Y) \end{aligned}$$

where in the final line we have used the Magic Variance Formula one final time.

Note that the b disappears entirely. Why does this make sense? The variance represents the spread of a random variable around its mean. Shifting a random variable by a constant amount should not affect this spread.

We conclude this section with a result about the variance of a sum of *independent* random variables which we will need later. We have not formally defined independence for random variables, but the intuition is that the output of one does not affect the output of each other. This result is only true for independent random variables.

Variance of a sum of independent random variables

If $X_1, X_2, \ldots, X_n$ are *independent* random variables, then

$$Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$$

i.e. the variance of a sum is the sum of the variances. This is only true for independent random variables.

2.5 Bernoulli Trials

One of the simplest models in probability is that of a sequence *Bernoulli trials*. Essentially Bernoulli trials are the generalization of repeated coin tossing.

Bernoulli trials

A sequence of Bernoulli trials is a sequence of experiments satisfying the following assumptions:

1. Each trial has exactly two possible outcomes, designated *success* and *failure*.
2. The trials are independent, i.e. the outcome of one trial does not influence the outcome of any other trial.
3. On each trial, the probability of success is p and the probability of failure is $1-p$, where $0 \leq p \leq 1$. (Sometimes, for convenience, we define $q = 1-p$).

4. These probabilities are constant for all trials, i.e. p does not change as the experiment is repeated.

Since Bernoulli trials have only two possible outcomes, we can think of them as "yes or no" events. Here are some examples of Bernoulli trials:

1. Flipping a coin of a coin, where we designate a flip of heads as a success. If it is a fair coin, then $p = 1/2$. If it is an unfair coin, then $p \neq 1/2$.
2. Rolling two standard, six-sided dice, where we designate a roll of doubles as success (and every other roll is failure). This is how you roll to get out of jail in Monopoly.
3. Playing a slot machine in Las Vegas.
4. Calling a registered voter in Providence uniformly at random and asking them if they are voting for Jorge Elorza.
5. Repeated free throw attempts by your instructor, assuming that he never improves (sadly, this is realistic!)

Here are some examples of trials which are not Bernoulli trials. Why are these not Bernoulli trials?

1. Drawing cards one at a time from a standard deck of 52 playing cards, where we designate success as drawing an ace.
2. Calling a list of 100 registered voters one-at-a-time and asking them if they are voting for Jorge Elorza.
3. Repeated free throw attempts by your instructor, assuming his skill can improve (if only a little) with practice.

An individual Bernoulli trial modeled by a Bernoulli random variable.

Bernoulli random variable

A *Bernoulli random variable* Y with parameter p is a random variable which is 1 with probability p and 0 with probability $1 - p$. It models a single Bernoulli trial, where 1 indicates success and 0 indicates failure. We can write its pmf in the table below:

y	$p(y)$
0	$1 - p$
1	p

To indicate that Y is a Bernoulli random variable with parameter p, we will sometimes write $Y \sim \text{Bernoulli}(p)$.

What is the mean and variance of a Bernoulli random variable? Let $Y \sim \text{Bernoulli}(p)$. Then for the expected value:

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{\text{all } y} y\, p(y) \\ &= 0(1-p) + 1(p) \\ &= p\end{aligned}$$

For the variance, we will use the Magic Variance Formula. This requires computing $\mathbb{E}(Y^2)$.

$$\begin{aligned}\mathbb{E}(Y^2) &= \sum_{\text{all } y} y^2\, p(y) \\ &= 0^2(1-p) + 1^2(p) \\ &= p\end{aligned}$$

By the Magic Variance Formula,

$$\begin{aligned}Var(Y) &= \mathbb{E}(Y^2) - [E(Y)]^2 \\ &= p - p^2 \\ &= p(1-p)\end{aligned}$$

Summarizing, these results, we have:

Characteristics of a Bernoulli random variable

Let Y be a Bernoulli random variable with parameter p. Then:

$$\begin{aligned}\mathbb{E}(Y) &= p \\ Var(Y) &= p(1-p) = pq\end{aligned}$$

where $q = 1 - p$.

Bernoulli trials are useful in many situations. Here are two questions we might want to ask about a sequence of Bernoulli trials.

1. How many successes do we have out of a *fixed* number of trials?
2. How many trials does it take to get our first success?

The first question is answered by the binomial distribution, and the second by the geometric distribution.

2.6 Binomial distribution

The binomial distribution models the number of successes in a fixed number of Bernoulli trials. Let's start with an example.

Example. Suppose we are playing a slot machine in Las Vegas. Suppose p is the probability of winning on one play[14]. Suppose you play 10 times. What is the probability that you win 2 times?

First, this is an example of Bernoulli trials, since each play of the slot machine has two possible outcomes (win and loss), is independent, and the probability of winning is does not change as you continue to play.

What is the sample space for this problem? Let us use the symbol `W` for a win and the symbol `L` for a loss. Then any sequence of 10 plays can be represented as a sequence of 10 `W`s and `L`s. One sequence might be `LLWLLWLLLL`. Is each sequence equally likely? Unless $p = 1/2$, different sequences can have different probabilities. For example, if $p < 1/2$, `LLLLLLLLLL` is much more probable than `WWWWWWWWWW`. What is the probability of a given sequence? Since we are looking at the probability of 2 wins (and thus 8 losses), let's look at a typical sequence where this is the case: `LLWLLWLLLL`. Since these are Bernoulli trials and thus each trial is independent, we just multiply together the probability of winning or losing at each trial.

$$\mathbb{P}(\texttt{LLWLLWLLLL}) = (1-p)(1-p)p(1-p)(1-p)p(1-p)(1-p)(1-p)(1-p) = p^2(1-p)^8$$

We see from this that the probability of a sequence depends only on the total number of wins and losses, not on the order in which those wins and losses occur. Thus the probability of *any* sequence of 10 trials composed of 2 wins and 8 losses will be $p^2(1-p)^8$.

The only question remaining is how many events in the sample space correspond to 2 wins and 8 losses, i.e. how many strings of length 10 contain exactly 2`W`s and 8`L`s? Recalling our combinatorics, this is given by the binomial coefficient $\binom{10}{2}$. Since each such string is a simple event in our sample space, we add up the probabilities of all $\binom{10}{2}$ of them to get:

$$\mathbb{P}(\text{2 wins out of 10 trials}) = \binom{10}{2} p^2(1-p)^8$$

We now generalize this. Suppose we have a sequence of n Bernoulli trials, with probability of success p for each trial. What is the probability of obtaining y successes, where y is an integer between 0 and n? If we designate success by `S` and failure by `F`, we can take the sample space to be strings of n letters, where each letter is either `S` or `F`. Just as in the example above, the probability of any string depends only on the number of `S`s and `F`s, not on their order. If we have y successes, this corresponds to a string of length n with y `S`s and

[14]This can be quite different depending on the type of slot machine you are playing, although all slot machines are designed so that you lose money on average. There are numerous websites devoted to slot machine odds.

$(n-y)$ Fs. The probability of that event is $p^y(1-p)^{n-y}$. Using the binomial coefficient, the number of strings of length n composed of y Ss and $(n-y)$ Fs is $\binom{n}{y}$. Thus we have:

$$\mathbb{P}(y \text{ successes out of } n \text{ trials}) = \binom{n}{y} p^y (1-p)^{n-y}$$

This probability distribution, which represents the number of successes in a fixed number n Bernoulli trials, is called the *binomial probability distribution*. A random variable Y with this distribution is called a *binomial random variable*.

Binomial distribution

A discrete random variable Y has a *binomial distribution* with n trials and probability of success p if

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y} \qquad\qquad y = 0, 1, \ldots, n$$

Y is a *binomial random variable* with parameters n and p, written $Y \sim \text{Binomial}(n, p)$. This models the number of successes out of n Bernoulli trials, where the probability of a single success is p.

First let's check to make sure Y is a well-defined discrete random variable, i.e. the probabilities of all its possible outputs sum to 1. Using the binomial theorem, we see that

$$\sum_{y=0}^{n} p(y) = \sum_{y=0}^{n} \binom{n}{y} p^y (1-p)^{n-y} = [p + (1-p)]^n = 1^n = 1$$

What is the expected value and variance of a binomial random variable? There are many ways to compute this, but we will use a clever trick based on linearity. Let $Y \sim \text{Binomial}(n, p)$. For each trial $i = 1, 2, \ldots, n$ let:

$$Y_i = \begin{cases} 0 & \text{trial } i \text{ is a failure} \\ 1 & \text{trial } i \text{ is a success} \end{cases}$$

Since $\mathbb{P}(Y_i = 1) = p$, $Y_i \sim \text{Bernoulli}(p)$. Since each Bernoulli random variable is 1 only if that trial is a success and Y is the number of trials,

$$Y = Y_1 + Y_2 + \cdots + Y_n$$

Using the linearity of expectation,

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}(Y_1 + Y_2 + \cdots + Y_n) \\ &= \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \cdots + \mathbb{E}(Y_n) \\ &= p + p + \cdots + p = np \end{aligned}$$

Furthermore, since Bernoulli trials are independent, the Y_i are all independent. Thus we can do the same thing for the variance:

$$\begin{aligned} Var(Y) &= Var(Y_1 + Y_2 + \cdots + Y_n) \\ &= Var(Y_1) + Var(Y_2) + \cdots + Var(Y_n) \\ &= p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p) \end{aligned}$$

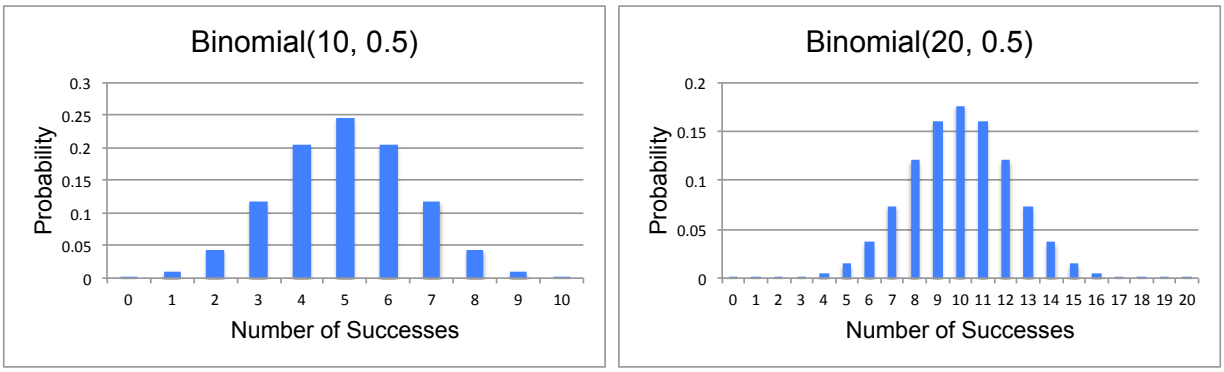
Summarizing these properties, we have:

Properties of binomial distribution

Let Y be a binomial random variable with parameters n and p. Then

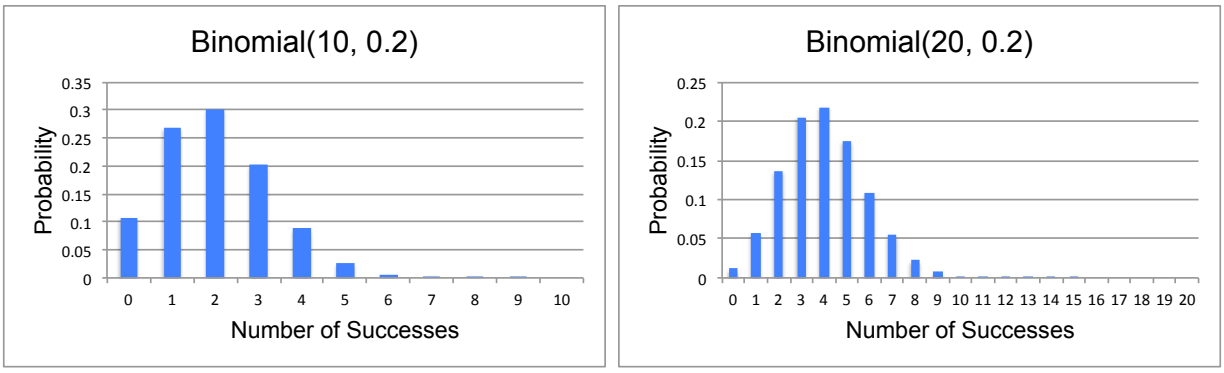
$$\begin{aligned} \mathbb{E}(Y) &= np \\ Var(Y) &= np(1-p) \end{aligned}$$

Let's look at histograms of the binomial distribution for a few choices of parameters. First here are histograms for $p = 1/2$ (fair coin):



These distributions are perfectly symmetric about the mean, which is expected for the case where $p = 1/2$. Note that as n increases, these look more and more like "bell curves". While we have not yet talked about the normal distribution, keep in mind how these histograms look. For large enough n, we will be able to approximate the (discrete) binomial distribution by the (continuous) normal distribution.

The histograms look a bit different for p significantly different from $1/2$. These histograms are for $p = 0.2$.



These are not symmetric. Since $p < 1/2$, the distributions are skewed to the left, which is what we expect since failure is more likely that success. Although not to the same extent as the case where $p = 1/2$, these also start to look like bell curves as n increases. We will also be able to approximate these by normal distributions for large n, but the farther p is from $1/2$, the larger n will need to be for this approximation to be reasonable.

Example. When a certain variety of pea plant with purple flowers is cross-fertilized, the offspring have purple flowers 3/4 of the time and white flowers 1/4 of the time[15].

1. You cross-fertilize 20 of these pea plants. What is the the probability that none of them are white?

   We can model this problem as a sequence of 20 Bernoulli trials, with "success" defined as a cross-fertilization producing a white flower. (Are all the assumptions of Bernoulli trials satisfied here?) Since we are looking for the number of successes in a *fixed* number of trials, this is a binomial distribution. Let $X \sim Binomial(20, 1/4)$. Then, using the binomial pmf:

   $$\mathbb{P}(X = 0) = \binom{20}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^{20} = \left(\frac{3}{4}\right)^{20} \approx 0.0032$$

   In other words, it is highly unlikely that we will have no white flowers.

2. What is the probability that we will have at least 2 white flowers?

   Let A be the event that we get at least two white flowers. It is much easier here to compute the probability of the opposite event A^c, the event that we get 0 or 1 white flower.

   $$\begin{aligned}\mathbb{P}(A^c) &= \mathbb{P}(X = 0) + \mathbb{P}(X = 1)\\ &= \binom{20}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^{20} + \binom{20}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{19}\\ &= \left(\frac{3}{4}\right)^{20} + 20\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^{19} \approx 0.024\end{aligned}$$

[15]This is one of the varieties Gregor Mendel tested. In this problem, the pea plants are heterozygous for flower color, i.e. genotype Pp.

Thus we have $\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 0.976$.

3. What is the expected number of white flowers?

   The expected value of a binomial random variable is np. For this case, we have $n = 20$ and $p = 0.25$, thus
   $$\mathbb{E}(X) = np = (20)(0.25) = 5$$

2.7 Geometric Distribution

The geometric distribution models the number of Bernoulli trials it takes to get the first success. Using the slot machine example from the previous section, if were willing to play a slot machine repeatedly until you won (i.e. you had infinite patience and money), the number of plays it took to get your first win would follow the geometric distribution.

Consider a sequence of Bernoulli trials with probability of success p. Here is our experiment: we will perform a series of trials until we get a success. The outcome of the experiment is the number of trials it takes to get our first success. Let Y be the random variable denoting the number of the trial on which the first success occurs. The smallest possible value for Y is 1, since we could succeed the first time[16]. However, there is no upper limit for Y; in principle, the trials could go on indefinitely. Y is a discrete random variable, but Y can take values in the countably infinite set $\mathbb{N} = 1, 2, 3, \ldots$. This is the first example we have seen of a discrete random variable with a countable range.

As in the binomial case, we can describe all possible outputs of our experiment by strings consisting of the letters `S` and `F`, where `S` indicates a success and `F` indicates a failure. This time, however, the strings are not of fixed length; however, the final element in each string is always `S`, and every element before that is `F`. The event $(Y = 1)$ corresponds to the string `S`, $(Y = 1)$ corresponds to `FS`, etc. Since the trials are independent, we have the following table for the pmf of Y.

y	event		$p(y)$
1	`S`	success on first trial	p
2	`FS`	failure on first trial, success on second trial	$(1-p)p$
3	`FFS`	first success on third trial	$(1-p)^2p$
4	`FFFS`	first success on fourth trial	$(1-p)^3p$
	$\vdots$		$\vdots$
k	$\underbrace{\texttt{FFF}\ldots\texttt{FS}}_{k-1 \text{ times}}$	first success on kth trial	$(1-p)^{k-1}p$
	$\vdots$		$\vdots$

We use this pmf to define a geometric random variable.

[16] Y cannot be 0, since you cannot succeed if you don't play.

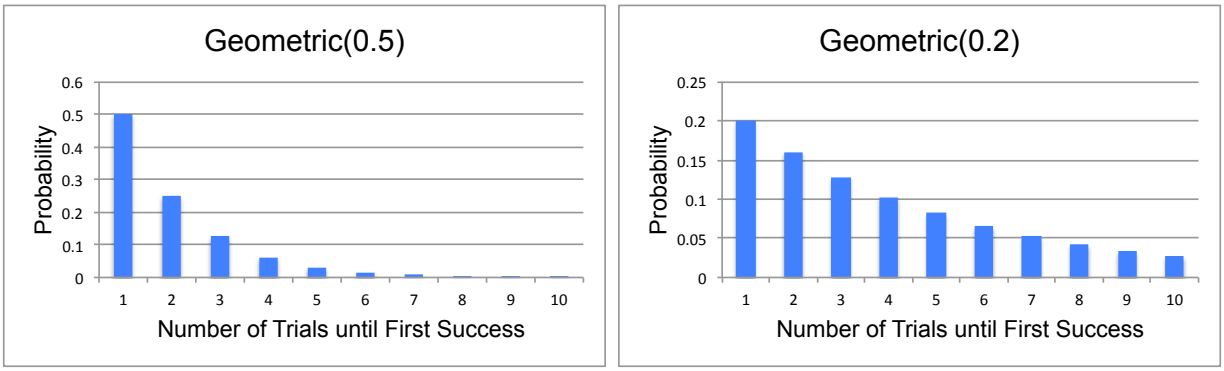
Geometric distribution

A discrete random variable Y has a *geometric distribution* with probability of success p if:

$$p(y) = (1-p)^{y-1}p \qquad\qquad y = 1, 2, 3, \ldots$$

Y is a *geometric random variable* with parameter p, written $Y \sim \text{Geometric}(p)$. This models the number of trials it takes to get the first success, where the probability of a single success is p.

Here are histograms for the geometric distribution for parameters 0.5 and 0.2.



Note that in all cases, the probability $p(y)$ decreases as y increases. Why does this make sense? The decrease is less prominent for the $p = 0.2$ case since failure is more likely; thus it should take a greater number of trials to get a single success.

Just as in the case for a binomial random variable, we will verify that a geometric random variable is a valid random variable by proving that the probabilities of all its possible outputs sum to 1. We will also find its expected value and variance.

Before we do that, we need a result about the sum of a geometric series. This may be familiar from precalculus or calculus. There are many ways to state this result. This is the one that I like.

Geometric series

A *geometric series* is a series whose successive terms are related by a common ratio. We can write a generic geometric series as:

$$a + ar + ar^2 + ar^3 + \cdots = \sum_{k=1}^{\infty} ar^{k-1}$$

In this case, the first term is a and the common ratio is r. If $|r| < 1$, then the infinite series converges, and:

$$\sum_{k=1}^{\infty} ar^{k-1} = \frac{a}{1-r}$$

Let's use this result to show that the probabilities of a geometric random variable sum to 1. Let Y be a geometric random variable with parameter p. Then

$$\begin{aligned}\sum_{y=1}^{\infty} p(y) &= \sum_{y=1}^{\infty} p(1-p)^{k-1} \\ &= \frac{p}{1-(1-p)} \\ &= \frac{p}{p} \\ &= 1\end{aligned}$$

where we used the formula above with the first term $a = p$ and the common ratio $r = p - 1$. For the expected value and variance of a geometric random variable we have the following.

Properties of geometric distribution

Let Y be a geometric random variable with parameter p. Then

$$\mathbb{E}(Y) = \frac{1}{p}$$
$$Var(Y) = \frac{1-p}{p^2}$$

The variance is tedious to prove, and since not much is gained by doing it, we will omit it. The proof of the expected value involves a nice calculus trick. We will assume that p is between 0 and 1, since if $p = 1$, you are guaranteed to get your first success on the first trial,

and if $p = 0$, you can never succeed.

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{y=1}^{\infty} yp(y) && \text{definition of expected value} \\
&= \sum_{y=1}^{\infty} y(1-p)^{y-1}p \\
&= p\sum_{y=1}^{\infty} yq^{y-1} && \text{writing } q = 1-p, \text{ for convenience} \\
&= p\sum_{y=1}^{\infty} \frac{d}{dq}q^y && \text{power rule of differentiation} \\
&= p\frac{d}{dq}\sum_{y=1}^{\infty} q^y && \text{swapping summation and differentiation}[17] \\
&= p\frac{d}{dq}\left(\frac{q}{1-q}\right) && \text{geometric series; first term } q, \text{ common ratio } q < 1 \\
&= p\frac{(1-q)(1)-(q)(-1)}{(1-q)^2} && \text{quotient rule of differentiation} \\
&= \frac{p}{p^2} \\
&= \frac{1}{p}
\end{aligned}$$

The geometric distribution is often used the model the distribution of how long you need to wait for a particular event to happen. In this case, the waiting time is given in discrete "chunks" of time such as hours or days.

Example. You are using a computer which runs Windows 10[18]. Suppose the probability of Windows 10 crashing in any given 1-hour period is 0.05.

1. What is the probability that the computer crashes within the first two hours after you start it up?

   Measuring time in one-hour intervals, let X be the number of one-hour intervals until your computer crashes. We will model X as a geometric random variable with parameter $p = 0.05$. (A "success" in this case is defined as your computer crashing!) We are looking for $\mathbb{P}(X \leq 2)$:

$$\begin{aligned}
\mathbb{P}(X \leq 2) &= \mathbb{P}(X = 1) + \mathbb{P}(X = 2) \\
&= p + (1-p)p \\
&= (0.05) + (0.95)(0.05) \\
&= 0.0975
\end{aligned}$$

[17] This can be justified; if you want to know more about this and find these types of mathematical subtleties interesting, I recommend taking a course in real analysis.

[18] This problem is not intended to be an endorsement of any particular operating system.

2. What is the probability that your computer will still be running two hours after you start it up?

   Here want the probability that the first crash occurs after Hour 2. There are two equivalent ways to look at this. First we can compute $\mathbb{P}(X > 2)$ directly using the result from the first part.

$$\begin{aligned}\mathbb{P}(X > 2) &= 1 - \mathbb{P}(X \leq 2) \\ &= 1 - 0.0975 \\ &= 0.9025\end{aligned}$$

   Alternatively, the probability that your computer is still running after two hours is equal the probability that it does not crash in Hour 1 and that it does not crash in Hour 2.

$$\begin{aligned}\mathbb{P}(X > 2) &= \mathbb{P}(\text{no crash in Hour 1} \cap \text{no crash in Hour 2}) \\ &= \mathbb{P}(\text{no crash in Hour 1})\,\mathbb{P}(\text{no crash in Hour 2}) \\ &= (1-p)^2 \\ &= 0.95^2 \\ &= 0.9025\end{aligned}$$

   There is a rougly 90% chance your computer is still running after two hours.

3. What is the expected number of hours your computer will run until it crashes?

   We are looking for $\mathbb{E}(X)$. Since X is a geometric random variable, its expected value is $1/p$, so:

$$\mathbb{E}(X) = \frac{1}{p} = \frac{1}{0.05} = 20$$

4. Given that your computer is still running at the 10-hour mark, what is the probability that it is still running after two more hours?

   Here, we want the probability that the computer does not crash in Hour 11 or Hour 12 give than it has made it to the 10-hour mark. Since the computer has made it to the 10-hour mark, we know that $X > 10$. We are interested in $\mathbb{P}(X > 12 | X > 10)$. By the definition of conditional probability,

$$\begin{aligned}\mathbb{P}(X > 12 | X > 10) &= \frac{\mathbb{P}(X > 12 \cap X > 10)}{\mathbb{P}(X > 10)} \\ &= \frac{\mathbb{P}(X > 12)}{\mathbb{P}(X > 10)}\end{aligned}$$

   since if $X > 12$, then it must also be true that $X > 10$. Using the same logic as in the second method of the previous part, if $X > 12$, then the computer must not have

crashed in Hours 1-12, so $\mathbb{P}(X > 12) = (1-p)^{12}$. Similarly, $\mathbb{P}(X > 10) = (1-p)^{10}$. Plugging these in above, we get

$$\mathbb{P}(X > 12 | X > 10) = \frac{(1-p)^{12}}{(1-p)^{10}} = (1-p)^2 = \mathbb{P}(X > 2) = 0.9025$$

Thus the probability that the computer is still running after two more hours given that it has already been running for 10 hours is the same as the probability that the computer is still running two hours after startup. This property of the geometric distribution is called *memorylessness*

Is the geometric distribution a good model for the previous problem? The assumption here is that the probability of crashing within any one-hour period is constant, no matter how long the computer is running for. Since you are the one modeling the problem, you need to decide how reasonable this assumption is. Do you expect this to be the case, or do you, for example, expect that a particular computer will be more likely to crash the longer it has been running? This can depend on many factors including the age of the computer and the OS it is running. Perhaps it is reasonable assumption if the computer has only been running for a short period of time.

The memoryless property, which we observed in the previous example, is a fundamental property of the geometric distribution. We can state it mathematically in the following way.

Memoryless property of the geometric distribution

Let Y be a geometric random variable with parameter p. Then for all m and n

$$\mathbb{P}(Y > m + n | Y > m) = \mathbb{P}(Y > n)$$

If the first success has *not* occurred by trial number m, the distribution of the remaining number of trials needed to get the first success is the same as if you started from scratch.

If we think of Bernoulli trial in terms of a casino game such as roulette, the memoryless property of the geometric distribution makes sense. The individual spins of a roulette wheel are independent, so no betting strategy based on past outcomes of the roulette wheel can work.

2.8 The Hypergeometric Distribution

In this section we will look briefly at the hypergeometric distribution, which models sampling without replacement. Consider the following example.

Example. You have a bag of 20 marbles, 12 of which are red and 8 of which are black.

1. You draw a single marble from the bag five times, replacing it before each new draw. What is the probability that 3 of the 5 marbles drawn are red?

This can be modeled with a binomial random variable. Since the marbles are replaced before each draw, the probability of drawing a red marble is constant at $12/20 = 0.6$. The draws are thus independent, so we have a sequence of 5 Bernoulli trials. Considering a draw of a red marble as a success, let $X \sim \text{Binomial}(5, 0.6)$. Then:

$$\mathbb{P}(X = 3) = \binom{5}{3} 0.6^3 0.4^2 \approx 0.346$$

2. You draw five marbles from the bag without replacement. What is the probability that 3 of the 5 marbles drawn are red?

   This time we cannot use the binomial distribution since the probability of drawing a red marble changes as marbles are drawn from the bag. Using combinatorics, there are $\binom{20}{5}$ possible draws. There are $\binom{12}{3}$ ways to choose 3 of the 12 red marbles, and $\binom{8}{2}$ ways to choose 2 of the 12 black marbles. Letting Y be the number of red marbles in our sample of 5, we have

$$\mathbb{P}(Y = 3) = \frac{\binom{12}{3}\binom{8}{2}}{\binom{20}{5}} \approx 0.397$$

The probabilities here are quite different (by about 5%). Since we are sampling 1/4 of the total marbles, it is unsurprising that there is a significant difference between sampling without replacement and sampling with replacement. What would happen if sampled a smaller fraction of the total marbles?

Example. Repeat the previous problem if we instead have a bag of 200 marbles, 120 of which are red and 80 of which are black.

If we sample with replacement, the probability will not change; if X is the number of red marbles drawn out of 5, we still have $X \sim \text{Binomial}(5, 0.6)$, so $\mathbb{P}(X = 3) = 0.346$. For sampling without replacement, let Y again be the number of red marbles out of a sample of 5. Then

$$\mathbb{P}(Y = 3) = \frac{\binom{120}{3}\binom{80}{2}}{\binom{200}{5}} \approx 0.350$$

In this case, the two probabilities differ by only about 0.5%, which is much smaller. The take-home message here is that if we sample a fraction of the total population, sampling without replacement is approximately equivalent to sampling with replacement, i.e. we can approximate sampling without replacement by a binomial distribution.

The distribution for sampling without replacement in this scenario is known as the *hypergeometric distribution*. It applies whenever we sample without replacement from a population which can be divided into two distinct groups. Examples of the hypergeometric distribution include:

1. Drawing a sample without replacement from a bag of marbles of two different colors.

2. Polling a sample of a population of eligible voters with a yes-no question or a choice between two candidates.

3. Drawing to complete a flush in Texas Hold'em poker. If, for example, the player has four clubs after the flop, then the probability of getting a club on the next two draws is hypergeometric since the draws are done without replacement.

The pmf for the hypergeometric distribution can be derived from combinatorics. I include it here for completeness and as an application of combinatorics. You do not need to memorize it, and there will be no homework or exam problems which require the direct use of the hypergeometric pmf. That being said, you do need to know the combinatorics behind sampling without replacement and problems involving calculating probabilities for sampling without replacement, like the examples above, are fair game.

We will write the hypergeometric pmf in terms of the marble problem. Suppose we have a bag of N marbles, r of which are red and the remaining $(N-r)$ of which are black. Suppose we take a sample of n marbles from the bag without replacement. Let Y be the number of marbles in our sample. Then the random variable Y has a hypergeometric distribution, and:

$$\mathbb{P}(Y = y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$$

where $y = 0, 1, 2, \ldots, n$. We have the additional restrictions that $y \leq r$ and $n - y \leq N - r$, which ensure we cannot draw more red or black marbles than are initially present in the bag.

To see this, note that there are a total of $\binom{N}{n}$ possible draws (order does not matter). The number of ways to choose y red marbles out of a total of r red marbles is $\binom{r}{y}$. The remaining $(n-y)$ marbles in the sample of n marbles must be black, and there are $\binom{N-r}{n-y}$ ways of choosing $(n-y)$ black marbles from a total of $(N-r)$ black marbles.

The final question to settle is when can we approximate a hypergeometric distribution (sampling without replacement) by a binomial distribution (sampling with replacement)? There is no hard-and-fast rule, but a good guideline is that if the sample size is less than 1/20 of the population size, the binomial distribution is a reasonable approximation.

2.9 Poisson Distribution

The Poisson distribution is the final discrete probability distribution we will discuss in this course. It it used to model the number of events which occur during a fixed interval of time (or space) under the following two assumptions:

1. The average rate of occurrence of the events is constant.

2. The events occur independently from each other.

Often the event in question is relatively rare. Examples of situations in which a Poisson distribution is a good model include:

1. The number of phone calls received per hour at a call center.
2. The number of pieces of non-junk mail received per day.
3. The number of traffic accidents occurring at a particular intersection per week.
4. The number of customers who enter a restaurant during a 15-minute period (although you might argue here that the average rate of arrival changes depending on the time of day.)
5. The number of decays per second of a radioactive isotope.
6. The number of errors per page in a manuscript.

Let's take the first example. Let Y be the number of calls received per hour at a call center, and suppose the average number of calls per hours is λ. We will do the following:

1. Split our hour up into n subintervals, where each subinterval is so small that at most one phone call can occur per subinterval. Let p be the probability that a phone call occurs in a given subinterval.
2. We can then model Y as $Y \sim \text{Binomial}(n, p)$, since the total number of calls in one hour is the total number of subintervals which contain one call.
3. The mean of Y is np, since it is a binomial random variable. Since the average number of calls per hour is λ, we will let $\lambda = np$ so $p = \lambda/n$
4. According to the binomial distribution, the probability $p(y) = \mathbb{P}(Y = y)$ is:

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y} = \frac{n!}{y!(n-y)!} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y}$$

5. Finally, we will let $n \to \infty$.

$$\begin{aligned}
\lim_{n\to\infty} p(y) &= \lim_{n\to\infty} \frac{n!}{y!(n-y)!} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \\
&= \lim_{n\to\infty} \frac{n(n-1)(n-2)\cdots(n-y+1)}{n^y} \frac{\lambda^y}{y!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-y} \\
&= \frac{\lambda^y}{y!} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n \frac{n}{n} \frac{(n-1)}{n} \frac{(n-2)}{n} \cdots \frac{(n-y+1)}{n} \left(1 - \frac{\lambda}{n}\right)^{-y} \\
&= \frac{\lambda^y}{y!} \lim_{n\to\infty} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\text{this has limit of } e^{-\lambda}} \underbrace{\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{y-1}{n}\right)\left(1 - \frac{\lambda}{n}\right)^{-y}}_{\text{these all have limit of 1}} \\
&= e^{-\lambda} \frac{\lambda^y}{y!}
\end{aligned}$$

The limiting probability $p(y)$ is the pmf for the *Poisson distribution*.

Poisson distribution

A discrete random variable Y has a *Poisson distribution* with parameter $\lambda > 0$ if

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

Y is called a *Poisson random variable* with parameter λ, which is often denoted $Y \sim \text{Poisson}(\lambda)$.

Note that the Poisson distribution can output a value of 0, which corresponds to no events happening in the fixed span of time.

As with the other discrete distributions, we will check that this distribution is well defined by verifying that the probabilities sum to 1.

$$\begin{aligned}\sum_{y=0}^{\infty} p(y) &= \sum_{y=0}^{\infty} e^{-\lambda}\frac{\lambda^y}{y!} \\ &= e^{-\lambda} \underbrace{\sum_{y=0}^{\infty}\frac{\lambda^y}{y!}}_{\text{Taylor series for } e^{\lambda}} \\ &= e^{-\lambda}e^{\lambda} = 1\end{aligned}$$

The mean of the Poisson distribution is λ, which makes sense since that is the average rate of occurrence of the events which we used when we constructed the Poisson distribution as a limit of the binomial distribution. Not only is the mean of a Poisson distribution λ, but the variance is also λ. Thus we have the following properties.

Properties of the Poisson distribution

Let Y be a Poisson random variable with parameter λ. Then

$$\mathbb{E}(Y) = \lambda$$
$$Var(Y) = \lambda$$

We will prove that the mean is λ. Verifying the variance is also λ is more tedious and will be omitted. For a Poisson random variable Y,

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{y=0}^{\infty} yp(y) \\
&= \sum_{y=0}^{\infty} ye^{-\lambda}\frac{\lambda^y}{y!} \\
&= \sum_{y=1}^{\infty} e^{-\lambda}\lambda^y\frac{y}{y(y-1)!} && \text{first term of the sum is 0} \\
&= \sum_{y=1}^{\infty} e^{-\lambda}\frac{\lambda^y}{(y-1)!} && \text{first term of the sum is 0}
\end{aligned}$$

Now we let $z = y - 1$. Substituting this in, we get

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{y=1}^{\infty} e^{-\lambda}\frac{\lambda^{z+1}}{z!} \\
&= \lambda\sum_{z=0}^{\infty} e^{-\lambda}\frac{\lambda^z}{z!} \\
&= \lambda
\end{aligned}$$

where the sum in the second-to-last line is 1 since we are summing over the entire pmf of a Poisson random variable.

Here is an example of where we can use the Poisson distribution.

Example. Customers arrive at your restaurant for dinner at an average rate of 5 customers per 15 minutes. Modeling this with a Poisson distribution:

1. What is the probability that no one arrives in a 15-minute period?

   Let X be the number customers who arrive in a 15-minute period. Then $X \sim$ Poisson(5). If no customers arrive in that time, then $X = 0$, so:

   $$\mathbb{P}(X = 0) = \frac{e^{-5}5^0}{0!} = e^{-5} \approx 0.0067$$

   So there is less than 1% chance of this occurring.

2. What is the probability that at least 3 customers arrive in a 15-minute period?

   Here we want $\mathbb{P}(X \geq 3)$. It is easier to compute the complement probability, i.e. the probability that 2 or fewer customers arrive in a 15-minute period.

   $$\begin{aligned}
   \mathbb{P}(X \geq 3) &= 1 - \mathbb{P}(X \leq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) \\
   &= 1 - \frac{e^{-5}5^0}{0!} - \frac{e^{-5}5^1}{1!} - \frac{e^{-5}5^2}{2!} \\
   &\approx 1 - 0.0067 - 0.0337 - 0.0842 = 0.08754
   \end{aligned}$$

3. What is the probability that exactly 5 customers arrive in a 15-minute period?

$$\mathbb{P}(X = 5) = \frac{e^{-5}5^5}{5!} \approx 0.175$$

3 Continuous Random Variables

3.1 Introduction

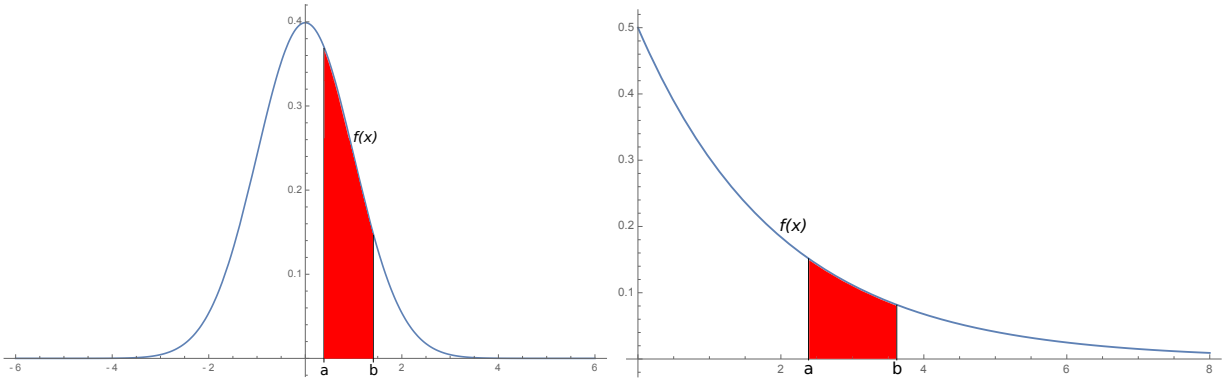
Many quantities of interest in the real world are not discrete is nature. Examples include the following:

1. The amount of rainfall in one day
2. The weight of an adult chimpanzee
3. The functional lifetime of an O-ring in a jet engine

A random variable which can take on any value within a range is called a *continuous random variable*. These are fundamentally different from discrete random variables in the following way. Recall that to specify a discrete random variable, all we have to do is assign probabilities between 0 and 1 to every possible output of the random variable in such a way that all the probabilities add up to 1. This is not possible for a random variable which can take values in an interval on the real number line. We must, therefore, use a different technique to describe a continuous random variable. Here will we use calculus for (essentially) the first time in the course.

3.2 Probability Density Functions

For a continuous random variable, rather than talk about the probability that a random variable equals a particular value, we talk about the probability that a random value falls within a particular range. (In fact, as we will see, the probability that a continuous random variable equals a specific value is 0.) A continuous random variable is described by a *probability density function (pdf)*. Here are examples of pdfs:



A pdf is a nonnegative function $f(x)$ where the total area under the curve is 1 (this is analogous to the discrete case where the probabilities of all the outputs sum to 1). The probability that a random variable falls into an interval $[a, b]$ is the area under the density curve between a and b. This is illustrated in red in the above pdfs. Since we are talking about areas under curves, we need to use calculus. In particular, we need to integrate! Here is the formal definition of a pdf:

Probability Density Function (pdf)

The function $f(y)$ is a *probability density function* (pdf) for a continuous random variable Y if $f(y) \geq 0$ for all y, and

$$\int_{-\infty}^{\infty} f(y)dy = 1$$

The probability that Y falls into the interval $[a, b]$ is the area under the density curve between a and b, i.e.

$$\mathbb{P}(a \leq Y \leq b) = \int_a^b f(y)dy$$

Before we continue, let's mention one way in which continuous and discrete distributions are very different. For a continuous distribution, the probability of any single point is 0. We can see that from the density function since for a continuous random variable Y with density y,

$$\mathbb{P}(Y = a) = \int_a^a f(y)dy = 0$$

Thus all the following probabilities are the same:

$$\mathbb{P}(a \leq Y \leq b) = \mathbb{P}(a \leq Y < b) = \mathbb{P}(a < Y \leq b) = \mathbb{P}(a < Y < b) = \int_a^b f(y)dy$$

In other words, it does not matter which inequality sign (¡ or $\leq$) we use for the endpoints of the interval since the probability of hitting the endpoints in 0. This is not at all the case for the discrete case, where individual points have positive probability. If X is the number of flips of heads in 20 tosses of a fair coin, then $\mathbb{P}(8 \leq X \leq 12)$ and $\mathbb{P}(8 < X \leq 12)$ are very different!

Example. Consider the function

$$f(y) = \begin{cases} cy^2 & 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the value of c for which $f(y)$ is a valid density function.

   First note that $f(y)$ is nonnegative, so there is nothing to worry about there. Next we need to make sure that the density function integrates to 1.

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(y)dy \\ &= \int_0^2 cy^2 dy && \text{since } f(y) \text{ is 0 outside } [0, 2] \\ &= c\frac{y^3}{3}\Big|_0^2 \\ &= \frac{8}{3}c \end{aligned}$$

   Thus, choosing $c = 3/8$, we have a valid density function.

2. What is $P(1 \leq Y \leq 2)$?

   Integrating the density function from 1 to 2, we get:

$$\begin{aligned}\mathbb{P}(1 \leq Y \leq 2) &= \int_1^2 f(y)dy \\ &= \int_1^2 \frac{3}{8}y^2 dy \\ &= \frac{3}{8}\frac{y^3}{3}\Big|_1^2 \\ &= \frac{7}{8}\end{aligned}$$

3.3 Cumulative Distribution Functions

Another way to describe a continuous random variable is with its *cumulative distribution function* (cdf)[19].

Cumulative Distribution Function (cdf)

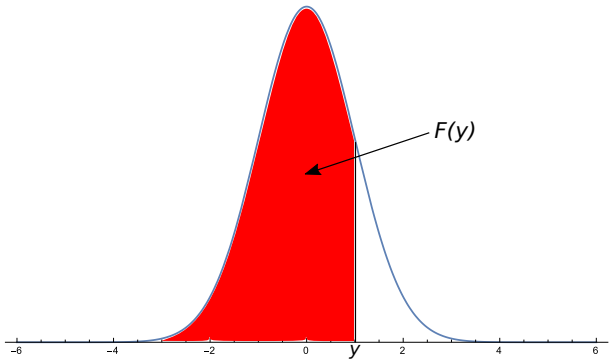
Let Y be a random variable. The the *cumulative distribution function* cdf of Y, denoted $F(y)$, is defined by

$$F(y) = \mathbb{P}(Y \leq y)$$

If Y has density function $f(y)$, then

$$F(y) = \int_{-\infty}^{y} f(y)dy$$

The cdf gives the probability that our random variable Y is less than or equal to a certain value. For a continuous random variable, $F(y)$ can be visualized graphically as the area under the density curve to the left of y.



[19]Sometimes you will see this called simply a *distribution function*. I will always use the term cdf.

Note that traditionally the cdf is written with an uppercase F, while the density is written with a lowercase f. The cdf is defined for discrete as well as continuous random variables, although we will never use it in the discrete case[20]. The cdf has the following properties.

Properties of cdfs

Let Y be a random variable with cdf $F(y)$. Then:

1. $F(y)$ is a nondecreasing function.
2. $\lim_{y\to-\infty} F(y) = 0$
3. $\lim_{y\to\infty} F(y) = 1$
4. For a continuous random variable Y, the cdf $F(y)$ is a continuous function.

For a continuous random variable, the cdf and pdf are related via the fundamental theorem of calculus.

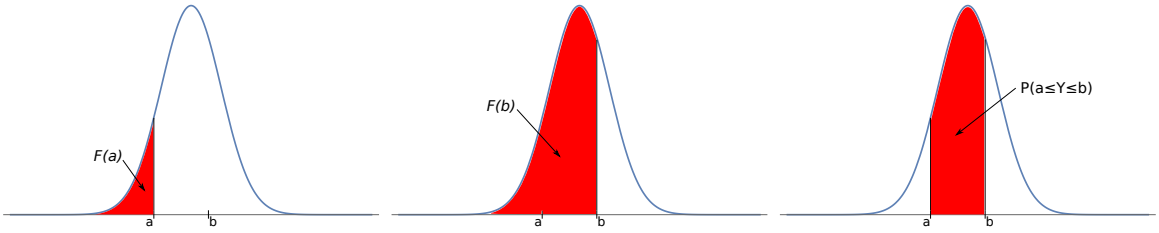
Relationship between cdf and pdf

Let Y be a continuous random variable with cdf $F(y)$ and density $f(y)$. Then we have the following relationships:

1. $$F(y) = \int_{\infty}^{y} f(y)dy$$
2. $$f(y) = \frac{DF(y)}{dy} = F'(y)$$
3. $$\mathbb{P}(a \leq Y \leq b) = \int_{a}^{b} f(y)dy = F(b) - F(a)$$

We can illustrate the third relationship above, $\mathbb{P}(a \leq Y \leq b) = F(b) - F(a)$, using the graphs below. You can see that subtracting the first area from the second area yields the third area.

[20]The cdf for a discrete random variable is always a step function, since the cdf only increases on the finite or countable set of points which have positive probabilities.



You may ask why we bother at all with cdfs since it seems easier to work with densities. There are three good reasons to consider cdfs. From a mathematical standpoint, the cdf is a more fundamental object than the pdf; every continuous random variable has a cdf, while there are continuous random variables which have no density[21]. From a practical standpoint, the normal distribution, perhaps the most important distribution is all of probability, has a density which is unwieldy, thus we will use the cdf to compute probabilities involving that distribution. Finally, the cdf is used to define the *median* and *quartiles* of a probability distribution, which are important descriptors of a probability distribution.

3.4 Median and Quartiles

The *median* is the "middle" of a probability distribution. It is the value which separates the upper half of the distribution from the lower half of the distribution. The median is more robust to outlier values than the mean, and so in some cases may be a better descriptor of a typical outcome than the mean. Mathematically, if Y is a random variable, the we can define the median of Y as the value m such that $\mathbb{P}(Y \leq m) = \mathbb{P}(Y \geq m) = 1/2$. For a continuous random variable Y with cdf $F(y)$, we can define the median using the cdf as the value m for which $F(m) = 1/2$. Similarly we can define the 1st and 3rd quartiles (the median is sometimes called the 2nd quartile).

Median and quartiles

Let Y be a continuous random variable with cdf $F(y)$. Then we define the *median* m, *first quartile* Q_1, and *third quartile* Q_3 by the following relationships:

$$F(Q_1) = \mathbb{P}(Y \leq Q_1) = 1/4$$
$$F(m) = \mathbb{P}(Y \leq m) = 1/2$$
$$F(Q_3) = \mathbb{P}(Y \leq Q_3) = 3/4$$

Let's revisit our example from the previous section.

Example. Let Y be a continuous random variable with density $f(y)$ defined by

$$f(y) = \begin{cases} \frac{3}{8}y^2 & 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

[21] All continuous random variables we will encounter in this course will have a density

Find the median of Y.

Let m be the median of Y. Since $F(m) = 1/2$, where $F(y)$ is the cdf of Y, we need to first find the cdf. For $y < 0$, $F(y) = 0$ and for $y > 2$, $F(y) = 1$ (do you see why this is the case?) For $0 \leq y \leq 2$, which is the region we care about, we integrate the density to get the cdf.

$$\begin{aligned} F(y) &= \int_0^y \frac{3}{8} t^2 dt \\ &= \frac{3}{8} \frac{t^3}{3} \Big|_0^y \\ &= \frac{y^3}{8} \end{aligned}$$

To find the median, we solve $F(m) = 1/2$ for m.

$$\begin{aligned} \frac{m^3}{8} &= \frac{1}{2} \\ m^3 &= 4 \\ m &= 2^{2/3} \approx 1.59 \end{aligned}$$

3.5 Expectation and Variance

Just as with discrete random variables, we can talk about the expected value and variance of a continuous random variable. They have the exact same interpretations as in the discrete case. Expectation and variance work almost exactly the same way in the continuous case as in the discrete case. In fact, we can use the exact same formulas, if we make two key changes:

1. The probability mass function $p(y)$ is replaced by the probability density function $f(y)$.
2. Summation is replaced by integration.

Making these changes, we have the following definitions for the expected value of a continuous random variable.

Expected value of a continuous random variable

Let Y be a continuous random variable with density $f(y)$. Then we define the expected value by

$$\mathbb{E}(Y) = \int_{\infty}^{\infty} y f(y) dy$$

If $g(y)$ is a real-valued function, then the expected value of $G(Y)$ is given by

$$\mathbb{E}[g(Y)] = \int_{\infty}^{\infty} g(y) f(y) dy$$

The variance of a continuous random variable is defined the same way as in the discrete case.

Variance of a continuous random variable

Let Y be a continuous random variable with density $f(y)$, and let $\mu = \mathbb{E}(Y)$. Then the variance of Y is defined byL

$$Var(Y) = \mathbb{E}[(Y-\mu)^2]$$

This is usually computed using the Magic Variance Formula, which holds for continuous random variables as well:

$$Var(Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2$$

Let's compute the expected value and variance of the continuous random variable we used in the section on probability densities.

Example. Let Y be the continuous random variable defined by the pdf

$$f(y) = \begin{cases} \frac{3}{8}y^2 & 0 \le y \le 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value and variance of Y?

Using the formula for expected value,

$$\begin{aligned} \mathbb{E}(Y) = \int_{-\infty}^{\infty} yf(y)dy & \\ &= \int_0^2 y\frac{3}{8}y^2 dy \\ &= \frac{3}{8}\int_0^2 y^3 dy \\ &= \frac{3}{8}\frac{y^4}{4}\Big|_0^2 \\ &= \frac{3}{8}\frac{16}{4} = 1.5 \end{aligned}$$

For the variance, we will compute $\mathbb{E}(Y^2)$ and use the Magic Variance Formula.

$$\begin{aligned}\mathbb{E}(Y^2) = \int_{-\infty}^{\infty} y^2 f(y)dy & \\ &= \int_0^2 y^2 \frac{3}{8} y^2 dy \\ &= \frac{3}{8}\int_0^2 y^4 dy \\ &= \frac{3}{8}\frac{y^5}{5}\Big|_0^2 \\ &= \frac{3}{8}\frac{32}{5} = 2.4\end{aligned}$$

Thus by the Magic Variance Formula we have:

$$Var(Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = 2.4 - (1.5)^2 = 0.15$$

We will now look at three specific continuous distributions.

3.6 Continuous Uniform Distribution

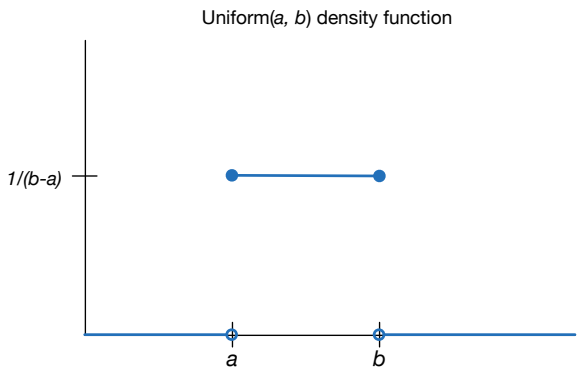
The *continuous uniform distribution*, which we shall generally just call the *uniform distribution*, describes the probability distribution on a finite interval $[a, b]$ which has the property that all subintervals of equal length are equally probable. The uniform distribution must be specified on a finite interval and is not defined for interval of infinite length. Here are some examples where we can use the uniform distribution to model a problem.

1. A RIPTA bus is scheduled to arrive at tunnel on Thayer St. at 8:00 am, but experience has shown that its arrival times vary between 8:00 am and 8:15 am. We could model this as a uniform distribution on the time interval $[0, 15]$, representing the number of minutes the bus is behind schedule.

2. Strokkur, one of the most famous geysers in Iceland, erupts approximately once every 10 minutes[22]. For a given 10-minute interval, we can model the probability that Strokkur will erupt by a uniform distribution on the interval $[0, 10]$

A uniform distribution is specified in terms of parameters a and b, which are the endpoints of the interval $[a, b]$ on which the uniform distribution is defined. What is the density function for the uniform distribution? Look at the picture below:

[22]Its neighbor Geysir, from which we get the word "geyser", hardly ever erupts these days.



The uniform density is a horizontal line between a and b and is 0 otherwise. (Does this make sense?) What is the height of the horizontal line? Since the integral of a probability density must integrate to 1, and since the uniform density is nothing more than a box, the area of the box must be 1. For that to be the case, the height of the box must be $1/(b-a)$. This is summarized below.

Uniform random variable

A continuous random variable Y has a *uniform distribution* on the interval $[a, b]$ if the pdf of Y is given by:

$$f(y) = \begin{cases} \frac{1}{b-a} & a \leq y \leq b \\ 0 & \text{otherwise} \end{cases}$$

Y a *uniform random variable*, which we can write as $Y \sim \text{Uniform}(a, b)$.

Let's do an example.

Example. A circle has a radius which is uniformly distributed on the interval $[0, 1]$. What is the expected value of the area of the circle?

Let $Y \sim \text{Uniform}(0, 1)$. Then Y has density

$$f(y) = \begin{cases} 1 & 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The area of a circle of radius y is given by $g(y) = \pi y^2$. Then the expected value of the area

is:

$$\begin{aligned}\mathbb{E}[g(Y)] &= \int_{-\infty}^{\infty} g(y)f(y)dy \\ &= \int_0^1 \pi y^2 \, 1dy \\ &= \pi\frac{y^3}{3}\Big|_0^1 \\ &= \frac{\pi}{3}\end{aligned}$$

As with every other distribution, we are interested in the mean and the variance of the uniform distribution. These are given below.

Properties of the uniform distribution

Let Y have the *uniform distribution* on the interval $[a, b]$. Then

$$\begin{aligned}\mathbb{E}(Y) &= \frac{a+b}{2} \\ Var(Y) &= \frac{(b-a)^2}{12}\end{aligned}$$

It makes sense that the mean of the uniform distribution is halfway between the endpoints. To verify this, for $Y \sim \text{Uniform}(a, b)$ with density $f(y)$ as given above:

$$\begin{aligned}E(Y) &= \int_{-\infty}^{\infty} yf(y)dy \\ &= \int_a^b y\frac{1}{b-a}dy \\ &= \frac{1}{b-a}\frac{y^2}{2}\Big|_a^b \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{(b+a)(b-a)}{2(b-a)} \\ &= \frac{a+b}{2}\end{aligned}$$

The variance can be found similarly by computing $\mathbb{E}(Y^2)$ and using the Magic Variance Formula.

3.7 Normal Distribution

The most important and most useful continuous probability distribution is the normal distribution, also known as the Gaussian distribution or the "bell curve" for its familiar bell-like

shape. Many observations from nature are approximately normally distributed, and we will see later that in the appropriate limit, just about everything has a normal distribution. In particular, we will see that for large enough n, we can approximate a binomial random variable with a normal distribution; this is especially useful since computations with the normal distribution are often easier than dealing with the pesky factorials in the binomial pmf.

Examples of the normal distribution from the natural (or artificial) world include:

1. Weights of chimpanzees (and just about any other species) follow the normal distribution.
2. Errors in scientific measurement (due to imperfect instrument and imperfect scientists) are normally distributed.
3. Scores on the SAT (and other standardized tests) are normally distributed (in this case by design).

We define a normal distribution by its density function.

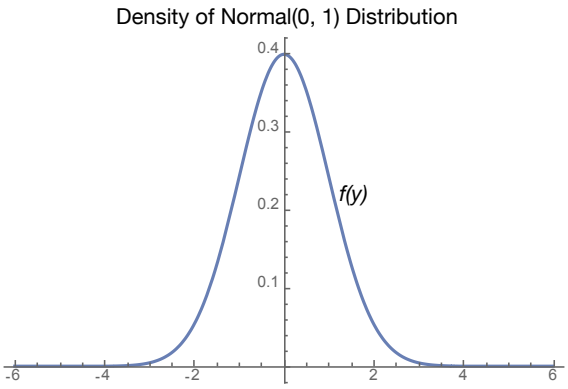
Normal distribution

A continuous random variable Y has a *normal distribution* with parameters μ and σ if its density $f(y)$ is given by

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Y is a *normal random variable*, denoted $Y \sim \text{Normal}(\mu, \sigma)$. Sometimes σ^2 is given as the parameter instead of σ.

A graph of the pdf for the normal distribution with parameters $\mu = 0$ and $\sigma = 0$ is shown below.



The parameters of the normal distribution, μ and σ, are the mean and standard deviation of the normal distribution.

Properties of the normal distribution

Let Y be a normal random variable with parameters μ and σ. Then

$$\mathbb{E}(Y) = \mu$$
$$Var(Y) = \sigma^2$$

σ is the standard deviation of Y.

The fact that this is a valid pdf, i.e. that it integrates to 1, is a standard exercise in multivariable calculus[23]. The normal distribution with $\mu = 0$ and $\sigma = 1$ is so prevalent that it is called the *standard normal distribution*. It is represented by the letter Z.

Standard normal distribution

The *standard normal distribution* is a normal distribution with parameters $\mu = 0$ and $\sigma = 1$ and is designated by the letter Z. The pdf of Z is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Let Z be a standard normal random variable, and suppose we wish to calculate $\mathbb{P}(-1 \leq Z \leq 1)$. Using the density of the standard normal:

$$\mathbb{P}(-1 \leq Z \leq 1) = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Unfortunately, there is no nice antiderivative for the integrand, so we cannot compute the integral using the fundamental theorem of calculus. One option is to use numerical integration techniques, which can be done using software packages such as Matlab or Mathematica. Another option is to use tables for the cdf of the standard normal distribution. Although this is perhaps a bit "old-school", it is important you know how to use these tables. There are many versions of the Z-table. The one I will provide (and the one which is on the course website) is the actual cdf for Z, i.e. is gives values for $F(z) = \mathbb{P}(Z \leq z)$. Since the standard normal is symmetric about 0, some tables only provide values on one side of the mean, since the others can be computed using symmetry. The table in the textbook by Wackerly et al (2008), for example, provides $\mathbb{P}(Z \geq z)$ for $z \geq 0$.

How do we compute this using a Z-table? Letting $F(z)$ be the cdf for the standard normal distribution, recall that for any continuous probability distribution we have:

$$\mathbb{P}(a \leq Z \leq b) = F(b) - F(a)$$

[23]The usual method is to square the integral, combine the integrands into a double integral, and change from cartesian to polar coordinates.

Then in this case, $\mathbb{P}(-1 \leq Z \leq 1) = F(1) - F(-1)$. Looking at the Z-table[24], we see that $F(1) = 0.8413$ and $F(-1) = 0.1587$. Subtracting, we obtain $\mathbb{P}(-1 \leq Z \leq 1) = 0.6826$.

Let's do an example so we can practice using a Z-table.

Example. Let Z be a standard normal random variable, and let $F(z)$ be its cdf.

1. Find $\mathbb{P}(Z > 2)$.

   Since $\mathbb{P}(Z > 2) = 1 - \mathbb{P}(Z \leq 2) = F(2)$, we just need to find $F(2)$. Looking this up in the Z-table, we see that $F(2) = 0.9772$, thus $\mathbb{P}(Z > 2) = 1 - 0.9772 = 0.0228$.

2. Find $\mathbb{P}(-2 \leq Z \leq 2)$

   $\mathbb{P}(-2 \leq Z \leq 2) = F(2) - F(-2)$. From the Z-table, $F(2) = 0.9772$ and $F(-2) = 0.0228$. Thus we have $\mathbb{P}(-2 \leq Z \leq 2) = 0.9772 - 0.0228 = 0.9544$.

3. Find $\mathbb{P}(0 \leq Z \leq 1.73)$.

   $\mathbb{P}(0 \leq Z \leq 1.73) = F(1.73) - F(0)$. From the Z-table, $F(1.73) = 0.9582$ (go to the row labeled 1.7, then across to the column corresponding to 0.03). We could use the table for $F(0)$, but since the distribution is symmetric about 0, we know $F(0) = 0.5$. Subtracting, we get $\mathbb{P}(0 \leq Z \leq 1.73) = 0.9582 - 0.5 = 0.4582$.

Before we go on, let us comment on two of the probabilities we already calculated. Recall the the standard deviation of the standard normal random variable is 1. Then $\mathbb{P}(-1 \leq Z \leq 1)$ is the probability of falling within one standard deviation of the mean and $\mathbb{P}(-2 \leq Z \leq 2)$ is the probability of falling within to standard deviations of the mean. These are useful numbers to remember, since they are good guidelines for interpreting the normal distribution.

Guidelines for normal probabilities

Let Y be a random variable with a normal distribution. Then:

1. The probability of falling within 1 standard deviation of the mean is about 0.68
2. The probability of falling within 2 standard deviations of the mean is about 0.95
3. The probability of falling within 3 standard deviations of the mean is about 0.997

This is known as the *68-95-99.7 rule*

What do we do when we don't have a normal random variable. We can transform *any* random variable $Y \sim \text{Normal}(\mu, \sigma)$ into a standard normal random variable Z using the following formula:

$$Z = \frac{Y - \mu}{\sigma}$$

[24]The value of $F(z)$ is denoted by z in our table. The column on the left gives us the first decimal place and tells us what row to use. Then we read across the row to match the second decimal place

In other words, we subtract the mean and divide by the standard deviation.

Example. A machine produces ball bearings which diameters which are normally distributed with mean 3.0005 cm and standard deviation 0.0010 cm. Specifications require ball bearing diameters which lie in the interval 3.0000 ± 0.0020 cm. What fraction of the total production meets those specifications.

Let Y be the diameter of a ball bearing produced by the machine. Then $Y \sim \text{Normal}(3.0005, 0.0010)$. We want the probability that Y is within the required range, i.e. $\mathbb{P}(2.9980 \leq Y \leq 3.0020)$. To do this, we convert this to an interval involving the standard normal random variable Z.

$$y = 2.9980 \qquad z = \frac{2.9980 - 3.0005}{0.0010} = -2.5$$

$$y = 3.0020 \qquad z = \frac{3.0020 - 3.0005}{0.0010} = 1.5$$

So $\mathbb{P}(2.9980 \leq Y \leq 3.0020) = \mathbb{P}(-2.5 \leq Z \leq 1.5)$. Looking at the Z-table, we find that $F(-2.5) = 0.0062$ and $F(1.5) = 0.9332$. Thus:

$$\mathbb{P}(2.9980 \leq Y \leq 3.0020) = \mathbb{P}(-2.5 \leq Z \leq 1.5) = F(1.5) - F(-2.5) = 0.9332 - 0.0062 = 0.9270$$

So approximately 92.7% of the ball bearings meet the required specifications.

3.8 Exponential Distribution

The final continuous distribution we will discuss is the exponential distribution. The exponential distribution belongs to the family of gamma distributions, and is perhaps the most useful member of that family; it is the only gamma distribution we will consider in this class.

The exponential distribution is used to model the length of time between events which occur independently and at a constant average rate. For example, it could be used the model the length of time between customer arrivals at a restaurant or phone calls at a call center. If this reminds you of a Poisson distribution, that is good! If we have a sequence of events which occur independently from each other and at a constant average rate, then:

1. The Poisson distribution is a discrete distribution which measures the number of events which occur in a fixed span of time.

2. The exponential distribution is a continuous distribution which measures the amount of time between two subsequent events.

Thus, the Poisson distribution and the exponential distribution ask different questions about the same problem.

The exponential distribution is also used to model the lifetime of electronic and mechanical components. Recall that we used the geometric distribution in a similar way; we did an example where the number of hours until a computer crashed was modeled as a geometric random variable. If we want to model the lifetime as a continuous random variable instead

of a discrete one, then we use the exponential distribution. The exponential distribution has another similarity to the geometric distribution; as we shall see, it is also memoryless. In fact, the geometric and exponential distributions are the only two probability distributions with this property.

Let's define the exponential distribution, then use it to model some problems.

Exponential distribution

A continuous random variable Y has an exponential distribution with parameter $\lambda > 0$ if its density function is:

$$f(y) = \begin{cases} \lambda e^{-\lambda y} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

Y is an exponential random variable with parameter λ, which is denoted $Y \sim$ Exponential(λ).

You will sometimes see the exponential density written as $(1/\beta)e^{-y/\beta}$. I prefer the above version, since the parameter λ is the same as that in the Poisson distribution, i.e. the average rate at which the events occur.

As always, we first verify that the exponential distribution is a valid probability density.

$$\begin{aligned} \int_{-\infty}^{\infty} f(y)dy &= \int_0^{\infty} \lambda e^{-\lambda y} dy \\ &= \lambda \lim_{t\to\infty} \int_0^t e^{-\lambda y} dy \\ &= \lambda \left(-\frac{1}{\lambda}\right) \lim_{t\to\infty} e^{-\lambda y}\Big|_0^t \\ &= -\left(\lim_{t\to\infty} e^{-\lambda t} - 1\right) \\ &= 1 \end{aligned}$$

since the limit of the exponential in the second-to-last line above is 0.

Next, we find the expected value and variance of the exponential distribution.

Properties of the exponential distribution

Let Y be an exponential random variable with parameter $\lambda > 0$. Then

$$\mathbb{E}(Y) = \frac{1}{\lambda}$$
$$Var(Y) = \frac{1}{\lambda^2}$$

For the expected value, if $Y \sim \text{Exponential}(\lambda)$ has density $f(y)$, then

$$\begin{aligned}
\mathbb{E}(Y) &= \int_{-\infty}^{\infty} y f(y) dy \\
&= \int_0^{\infty} \lambda y e^{-\lambda y} dy \\
&= \lambda \lim_{t\to\infty} \int_0^t \lambda y e^{-\lambda y} \\
&= \lambda \lim_{t\to\infty} \left(-\frac{1}{\lambda} y e^{-\lambda y}\Big|_0^t + \frac{1}{\lambda}\int_0^t e^{-\lambda y} dy \right) && \text{integration by parts} \\
&= -\lim_{t\to\infty} t e^{-\lambda t} + 0 + \frac{1}{\lambda}\int_0^{\infty} e^{-\lambda y} dy \\
&= \frac{1}{\lambda}
\end{aligned}$$

where in the second-to-last line the limit is 0 (exponentials grow faster than any power) and the integral is 1 (since it's the integral of the exponential density). Similarly, using the Magic Variance Formula and integrating by parts twice, we can prove the variance of an exponential random variable.

Example. While procrastinating studying for APMA 1650, you decide to create a website for your cat. Suppose your cat's website gets an average of 5 unique visits per hour. What is the probability that your cat will go more than 30 minutes without a visitor?

We can model the number of visits to your cat's website per hour as a Poisson distribution with parameter $\lambda = 5$ and the time between visits as an exponential distribution with the same parameter $\lambda = 5$. In this case, we are interested in the time between visits. Let $X \sim \text{Exponential}(5)$. Then X has density $f(x) = 5e^{-5x}$ Then since we are working in units of hours, the probability we want is $\mathbb{P}(X \geq 1/2)$.

$$\begin{aligned}
\mathbb{P}(X \geq 1/2) &= \int_{1/2}^{\infty} 5e^{-5x} dx \\
&= -\frac{5}{5} e^{-5x}\Big|_{1/2}^{\infty} \\
&= e^{-5/2} \approx 0.082
\end{aligned}$$

Just as the geometric distribution is memoryless, the exponential distribution is memoryless. If we think of this in terms of the lifetime of, say, a light bulb, this means that the probability of the bulb burning out in the next 30 minutes is the same whether we just turned the light bulb on or whether it has been on for 5 days. To prove the memoryless property of the exponential distribution, it will be useful to have an expression for the cdf of the exponential distribution.

Cumulative distribution function (cdf) for the exponential distribution

---

Let Y be an exponential random variable with parameter $\lambda > 0$. Then its cdf $F(y)$ is given by:

$$F(y) = \begin{cases} 1 - e^{-\lambda y} & y > 0 \\ 0 & y \leq 0 \end{cases}$$

To see this, we just integrate the density function. Let $Y \sim \text{Exponential}(\lambda)$, and let $f(y)$ be its density. For $Y \leq 0$, the cdf $F(y) = 0$. (Why is this the case?) For $y > 0$, we use the definition of the cdf to get:

$$\begin{aligned} F(y) &= \mathbb{P}(Y \leq y) \\ &= \int_{-\infty}^{y} f(t)dt \\ &= \int_{0,y} \lambda e^{-\lambda t} dt \\ &= -e^{-\lambda t}\Big|_{0}^{y} \\ &= 1 - e^{-\lambda y} \end{aligned}$$

We can use this to show the memoryless property of the exponential distribution.

Memoryless property of the exponential distribution

Let Y be an exponential random variable with parameter λ. Then for all a and b

$$\mathbb{P}(Y > a + b | Y > a) = \mathbb{P}(Y > b)$$

To see this, we first use the exponential cdf $F(y)$ to see that

$$\mathbb{P}(Y > y) = 1 - \mathbb{P}(Y \leq y) = 1 - F(y) = 1 - (1 - e^{-\lambda y}) = e^{-\lambda y}$$

Using this and the definition of conditional probability:

$$\begin{aligned}
\mathbb{P}(Y > a+b | Y > a) &= \frac{\mathbb{P}(Y > a+b \cap Y > a)}{\mathbb{P}(Y > a)} \\
&= \frac{\mathbb{P}(Y > a+b)}{\mathbb{P}(Y > a)} \\
&= \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} \\
&= e^{-\lambda b} \\
&= \mathbb{P}(Y > b)
\end{aligned}$$

We can also use the cdf of the exponential distribution to find the median. To do this we solve $F(m) = 1/2$, where $F(y)$ is the exponential cdf.

$$\begin{aligned}
F(m) &= \frac{1}{2} \\
1 - e^{-\lambda m} &= \frac{1}{2} \\
e^{-\lambda m} &= \frac{1}{2} \\
-\lambda m &= \log\left(\frac{1}{2}\right) = -\log 2 \\
m &= \frac{\log 2}{\lambda}
\end{aligned}$$

The median of the exponential distribution is thus a little smaller than the mean (by a factor of $\log 2 \approx 0.693$).

3.9 Bounds on Probabilities

Sometimes what we are interested in is the probability of an outlier event, i.e. the probability that a random variable deviates from the mean by more than a certain amount. Of course, if we know the exact probability distribution, this is not hard to calculate. However, even if we don't know the probability distribution, we can still obtain upper bounds on the probability of outlier events. In this section we will look at two inequalities which allow us to do exactly this. Markov's Inequality can be used if only the mean is known, but it gives us the weakest bound. Chebyshev's Inequality is a better bound, but required knowledge of both the mean and the variance. The moral of the story is: the more information we have about a probability distribution, the better estimates we can get.

3.9.1 Markov's Inequality

We first look at Markov's Inequality, which gives the probability that a nonnegative random variable exceeds a certain threshold in terms of the expected value of the random variable. To use Markov's Inequality, we only need to know the expected value.

Markov's Inequality

Let Y be a nonnegative random variable with expected value $E(Y)$. If $a > 0$, then

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}(Y)}{a}$$

To see that this is true, let I_a be the indicator random variable for the event $(Y \geq a)$, i.e. define I_a by

$$I_a = \begin{cases} 1 & Y \geq a \\ 0 & \text{otherwise, i.e. if } Y < a \end{cases}$$

Next note that $aI_a \leq Y$. To see this is true, if $Y < a$, then $aI_a = 0 \leq Y$ since Y is nonnegative (this is why we require Y to be nonnegative). If $Y \geq a$, then $aI_a = a \leq Y$. Thus we have:

$$\begin{aligned} \mathbb{E}(aI_a) &\leq E(Y) \\ a\left[1 \cdot \mathbb{P}(Y \geq a) + 0 \cdot \mathbb{P}(Y < a)\right] &\leq E(Y) \\ a\mathbb{P}(Y \geq a) &\leq E(Y) \\ \mathbb{P}(Y \geq a) &\leq E(Y) \end{aligned}$$

where in the second line we used the definition of the expected value of a discrete random variable.

Example. Suppose we randomly select an article from a journal where the mean article length is known to be 1000 words. Find an upper bound on the probability that an article exceeds 1400 words in length.

Let Y be the length of an article in this journal. We know nothing about Y except that its mean is 1000. Certainly Y is nonnegative, thus we can find an upper bound for $\mathbb{P}(Y \geq 1400)$ using Markov's inequality:

$$\mathbb{P}(Y \geq 1400) \leq \frac{E(y)}{1400} = \frac{1000}{1400} \approx 0.71$$

This is not a great upper bound, but it's better than nothing!

3.9.2 Chebyshev's Inequality

The second inequality we will look at is Chebyshev's Inequality, which gives the probability that a random variable deviates from its mean by more than a certain amount. This does not require a nonnegative random variable, but it does require knowledge of both the mean and the variance of a random variable.

Chebyshev's Inequality

Let Y be a nonnegative random variable with expected value $\mathbb{E}(Y)$ and variance $Var(Y)$. If $a > 0$, then

$$\mathbb{P}(|Y - E(Y)| \geq a) \leq \frac{Var(Y)}{a^2}$$

Note the absolute value sign inside the probability for Chebyshev's Inequality. This means that Chebyshev's Inequality gives a bound for the probability of deviating from the mean *in either direction*. Contrast this to Markov's Inequality, which is a bound on the probability of *exceeding* a certain value.

To see this is true, take the nonnegative random variable $(Y - \mathbb{E}(Y))^2$ and substitute it into Markov's Inequality along with the constant a^2. This gives us:

$$\mathbb{P}([Y - \mathbb{E}(Y)]^2 \geq a^2) \leq \frac{E[(Y - E(Y)]^2}{a^2} = \frac{Var(Y)}{a^2}$$

where the last equality uses the definition of variance. Then since $|x| \geq a$ if and only if $x^2 \geq a^2$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) = \mathbb{P}([X - \mathbb{E}(X)]^2 \geq a^2) \leq \frac{Var(Y)}{a^2}$$

which is Chebyshev's Inequality.

Sometimes we like to measure deviation from the mean in terms of "numbers of standard deviations". For the normal distribution, this is encapsulated in the 68-95-99.7 rule. We can state Chebyshev's Inequality in these terms if we like. Let Y be a random variable with mean μ and variance σ^2. Recall that the standard deviation is the square root of the variance, so the standard deviation of Y is σ. Then Chebyshev's Inequality states that the probability of deviating at least k standard deviations from the mean is bounded by:

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

To see this is true, just take $a = k\sigma$ in Chebyshev's Inequality above.

Example. Suppose we randomly select an article from a journal article length is distributed with a mean 1000 words and a standard deviation of 150 words. Find an upper bound on the probability that an article is outside of the range 600 - 1400 words?

Let Y be the length of an article in this journal. Here we know both the mean and the standard deviation of Y, we can can use Chebyshev's Inequality. We are looking for the probability that Y deviates from the mean by at least 400. Recalling that the variance is the square of the standard deviation:

$$\mathbb{P}(Y \geq 1400 \cup Y \leq 600) = \mathbb{P}(|Y - 1000| \geq 400) \leq \frac{Var(Y)}{400^2} = \frac{150^2}{400^2} \approx 0.14$$

Let's compare this to the bound we got in the previous example using Markov's Inequality. Since $(Y \geq 1400) \subset (Y \geq 1400 \cup Y \leq 600)$, using the properties of probability:

$$\mathbb{P}(Y \geq 1400) \leq \mathbb{P}(Y \geq 1400 \cup Y \leq 600) \leq 0.14$$

This is a much better bound than we got from Markov's Inequality. If we have reason to suspect that the distribution of Y is symmetric about the mean, we can divide $\mathbb{P}(Y \geq 1400 \cup Y \leq 600)$ by 2 to get:

$$\mathbb{P}(Y \geq 1400) \leq \frac{\mathbb{P}(Y \geq 1400 \cup Y \leq 600)}{2} \leq \frac{0.14}{2} = 0.07$$

which is even better! It is important that we can only do this is we have (or suspect we have) a symmetric probability distribution. This is not true in general.

Just for comparison purposes, let's see how much better this bound is if we know the exact distribution of Y. Suppose Y is normally distributed with the same parameters, i.e. $Y \sim \text{Normal}(1000, 150)$. Then, standardizing to the standard normal random variable Z with cdf $F(y)$:

$$\begin{aligned}
\mathbb{P}(600 \leq Y \leq 1400) &= \mathbb{P}\left(\frac{600 - 1000}{150} \leq Z \leq \frac{1400 - 1000}{150}\right) \\
&= \mathbb{P}(-2.67 \leq Z \leq 2.67) \\
&= F(2.67) - F(-2.67) \\
&= 0.9962 - 0.0038 = 0.9924
\end{aligned}$$

Thus we have:

$$\mathbb{P}(Y \geq 1400 \cup Y \leq 600) = 1 - \mathbb{P}(600 \leq Y \leq 1400) \leq 0.0076$$

Although Chebyshev's Inequality gives a decent bound on the probability of outliers, there is no substitute for knowing the actual probability distribution!

4 Multivariate Distributions

4.1 Introduction

Experimenters will often measure more than one quantity, and are often interested in the distribution of all observed quantities. As as example, a naturalist measures the height and weight of chimpanzees. They might be interested in the distribution of height-weight pairs. Since the distribution involves two quantities, we call it a *bivariate distribution*. One question might be whether or not these two quantities are independent. (We suspect they are not, since a reasonable assumption is that taller chimpanzees tend to weigh more.)

Another important application of multivariate distributions is statistical sampling. Suppose $Y_1, Y_2, \ldots, Y_n$ are n successive trials of an experiment. Statisticians are interested in the distribution of $(Y_1, Y_2, \ldots, Y_n)$, and can use information about this distribution to infer characteristics of the experiment or the population from which the experiment sampled.

In this section, we will primarily be interested in bivariate distributions, the probability distribution of two random variables. As before, we start with the discrete case and then consider the continuous case.

4.2 Distribution of Two Discrete Random Variables

First, let's define the joint probability distribution for a pair of discrete random variables.

Joint probability distribution, discrete case

Let Y_1 and Y_2 be two discrete random variables. Then the *joint distribution* of Y_1 and Y_2 is given by the function of two variables:

$$p(y_1, y_2) = \mathbb{P}(Y_1 = y_1, Y_2 = y_2) \qquad \text{for all possible pairs } (y_1, y_2)$$

Sometimes this is called the *joint probability mass function* (joint pmf). Note that $\mathbb{P}(Y_1 = y_1, Y_2 = y_2)$ means $\mathbb{P}(Y_1 = y_1 \cap Y_2 = y_2)$. This is standard notation for expressing the joint probability of two random variables.

We have already encountered one example of a bivariate distribution. Recall the distribution of the rolls of two standard, six-sided dice which we discussed several times in the section on discrete random variables. Let X_1 be the roll of the first die and X_2 the roll of the second die. Then since we have a discrete uniform distribution, the joint distribution of X_1 and X_2 is given by:

$$p(x_1, x_2) = \frac{1}{36} \qquad x_1, x_2 = 1, 2, 3, 4, 5, 6$$

Just as in the case for a single discrete random variable, for a joint distribution of two discrete random variable, all the possible probabilities are nonnegative and they sum to 1.

Let Y_1 and Y_2 be discrete random variables with joint probability distribution $p(y_1, y_2)$. Then

$$0 \leq p(y_1, y_2) \leq 1 \text{ for all } y_1, y_2$$
$$\sum_{\text{all } (y_1, y_2} p(y_1, y_2) = 1$$

where the sum is taken over all possible pairs (y_1, y_2).

Just as in the case of a single discrete random variable, we can construct a valid joint probability distribution of two discrete random variables by assigning probabilities that add up to 1. Let Y_1 be a discrete random variable with m possible output values, and Y_2 a discrete random variable with n possible output values. Then there are mn possible joint outputs of the pair of random variables. Each of the outputs is an ordered pair of the form (y_1, y_2). If we make an mxn table and assign probabilities to each of the mn possible joint outputs so they add up to 1, we have constructed a joint probability distribution for the two discrete random variables Y_1 and Y_2.

Let's consider an example.

Example. Imagine we surveyed Brown undergraduates and asked them two questions.

1. Do you have an exam this week?
2. How many cups of coffee have you drunk today?

Let X_1 be the discrete random variable with values {`yes`, `no`} indicating whether or not a student has an exam this week. Let X_2 be the number of cups of coffee a student has drunk today. For simplicity, we will let X_2 take only the values {0, 1, 2} (Whether or not this is a realistic simplification is beyond the scope of this course!)

We can display the joint probability distribution for the pair (X_1, X_2) in a 2 x 3 table. There are 6 possible values for the pair (x_1, x_2). We can choose any probabilities for the six pairs as long as they sum to 1. One possible choice is shown in the table below.

		X_2		
		0	1	2
X_1	`yes`	2/20	3/20	3/20
	`no`	6/20	4/20	2/20

You can verify that the six probabilities do indeed sum to 1.

4.2.1 Marginal distribution

Consider again a joint distribution (Y_1, Y_2) of two discrete random variables with pmf $p(y_1, y_2)$. (You can think of the exam-coffee example above). Y_1 and Y_2 are themselves

discrete random variables. What are their distributions?

Suppose we wish to find the distribution for Y_1 by itself. We call this the *marginal distribution* of Y_1. Essentially what we want to do is take Y_2 out of the picture entirely. How do we do that? All we have to do is sum over all the possible values of Y_2! The probability that $Y_1 = y_1$ is the sum of $p(y_1, y_2)$ over all possible values y_2 that Y_2 can take; this is the marginal distribution of Y_1, and is written $p_1(y_1)$. Similarly, we can sum over all possible values Y_1 to get $p_2(y_2)$, the marginal distribution of Y_2. This is summarized below.

Marginal distribution, discrete random variables

Let Y_1 and Y_2 be discrete random variables with joint probability distribution $p(y_1, y_2)$. Then the marginal distribution of Y_1 is given by:

$$p_1(y_1) = \sum_{\text{all } y_2} p(y_1, y_2)$$

and the marginal distribution of Y_2 is given by.

$$p_2(y_2) = \sum_{\text{all } y_1} p(y_1, y_2)$$

In both cases, we just sum the joint distribution over all the possibilities of the other random variable.

Let's return to our example above.

Example. In the exam-coffee example above, compute the marginal distributions for X_1 and X_2.

To find the marginal distributions for each variable, we sum over all the possibilities of the other variable. If the joint distribution is presented in a two-dimensional table, this is easy. To find the marginal distribution of X_2, we sum the values in each column. The bottom row, which we will label "total", is the marginal distribution of X_2. Similarly, we can find the marginal distribution for X_1 by summing each row. The rightmost column, also labeled "total", is the marginal distribution for X_2. In fact, the marginal distribution is called "marginal" because its values lie in the margins of the joint distribution table.

		X_2			
		0	1	2	total
X_1	`yes`	2/20	3/20	3/20	8/20
	`no`	6/20	4/20	2/20	12/20
	total	8/20	7/20	5/20	

You can check that the two marginal distributions sum to 1 and are thus valid probability distributions for discrete random variables.

4.2.2 Conditional distribution

Suppose again we have a joint distribution (Y_1, Y_2) of two discrete random variables with joint pmf $p(y_1, y_2)$. Another question we might ask is what is the distribution of Y_1 given that $Y_2 = y_2$. In other words, what is the conditional distribution of Y_1 given that Y_2 takes a specific value.

Let's look once more a the exam-coffee example to see how we can do this.

Example. In the exam-coffee example above,what is the distribution of the number of cups of coffee drunk today (X_2) given that a student has a midterm this week ($X_1 = \texttt{yes}$)?

To do this, we look at the first row of the table, which corresponds to $X_1 = \texttt{yes}$. This is not a valid probability mass function, because the elements do not sum to 1. But we can fix that! All we have to do is divide by the marginal probability $p_1(\texttt{yes}) = \mathbb{P}(X_1 = \texttt{yes})$, which is conveniently located just to the right in the "total" column. If we do that, we get the conditional probability for X_2 given $X_1 = \texttt{yes}$, which we can write as $p(y_2|\texttt{yes})$ or $p(y_2|Y_1 = \texttt{yes})$:

y	$p(y\|\text{midterm})$
0	2/8
1	3/8
2	3/8

Now that we have seen an example, we will give the formal definition of the conditional distribution of two discrete random variables.

Conditional distribution, discrete random variables

Let Y_1 and Y_2 be discrete random variables with joint probability distribution $p(y_1, y_2)$. Let $p_2(y_2)$ be the marginal distribution of Y_2. Then the conditional distribution of Y_1 given $Y_2 = y_2$ is:

$$p(y_1|y_2) = \mathbb{P}(Y_1 = y_1|Y_2 = y_2) = \frac{\mathbb{P}(Y_1 = y_1, Y_2 = y_2)}{\mathbb{P}(Y_2 = y_2)} = \frac{p(y_1, y_2)}{p_2(y_2)}$$

where $p_2(y_2) > 0$. In words, the conditional distribution is the joint distribution divided by the marginal distribution. We can similarly define the conditional distribution of Y_2 given $Y_1 = y_1$.

4.2.3 Independence

The final question to settle is independence. Roughly speaking, two random variables are independent of if the probabilities of each one are not affected by the value of the other one. The following will serve as our definition for independence of two discrete random variables.

Independence of discrete random variables

Let Y_1 and Y_2 be discrete random variables with joint probability distribution $p(y_1, y_2)$. Let $p_1(y_1)$ and $p_2(y_2)$ be the marginal distributions of Y_1 and Y_2. Then Y_1 and Y_2 are independent if

$$p(y_1, y_2) = p_1(y_1)p_2(y_2) \qquad \text{for all } y_1, y_2$$

In other words, two random variables are independent if their joint distribution is the product of the two marginal distributions.

In the exam-coffee example above, using just about any pair of y_1 and y_2, we can show that Y_1 and Y_2 and not independent. Did we really expect them to be independent?

4.3 Distribution of Two Discrete Continuous Variables

We will essentially repeat the same discussion for a pair of continuous random variables. Since working with continuous random variables requires integration, this will require integration in two dimensions, i.e. multivariable calculus. Since it is likely that many of you have not taken multivariable calculus, all multivariable techniques will be taught as they are needed.

4.3.1 Joint Probability Density

Recall that in the discrete case, a probability distribution was described by a probability density function (pdf). For the joint distribution of two continuous random variable, we have a joint density function, which is the continuous analogue of the joint distribution function in the discrete case.

Joint probability density, continuous case

Let Y_1 and Y_2 be two continuous random variables. Then the *joint density* of Y_1 and Y_2 is a function of two variables $f(y_1, y_2)$ with the properties that:

1. $f(y_1, y_2) \geq 1$ for all y_1, y_2.
2. $$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 dy_2 = 1$$

In other words, $f(y_1, y_2)$ is nonnegative and integrates to 1.

In the bivariate case, instead of finding the probability that a single variable lies in an interval $[a, b]$, we find the probability that a pair of random variables lies within a region of the plane. To do this, we integrate the joint density over that region.

Probability of an event, continuous bivariate case

Let Y_1 and Y_2 be two continuous random variables with joint density $f(y_1, y_2)$. Let A be a region of the plane. Then

$$\mathbb{P}((Y_1, Y_2) \in A) = \int \int_A f(y_1, y_2) dy_1 dy_2$$

This notation may not be precise. Don't worry about it for now, we will do plenty of examples. Just remember the key idea: to find the probability that a pair of random variables lie in a region, integrate the joint density over that region.

The extra complication here is the double integral. Whereas a single integral is defined on a closed interval, a double integral is defined on a two-dimensional region of the plane. The key to success for any double integral problem is to draw the region of integration before doing anything else. Since this is so important, I will repeat it: always draw the region of integration!

We will learn how to handle this through a series of examples. We will keep coming back to this first example throughout this section.

Example. Let X and Y be random variables with joint distribution function $f(x, y)$ given by:

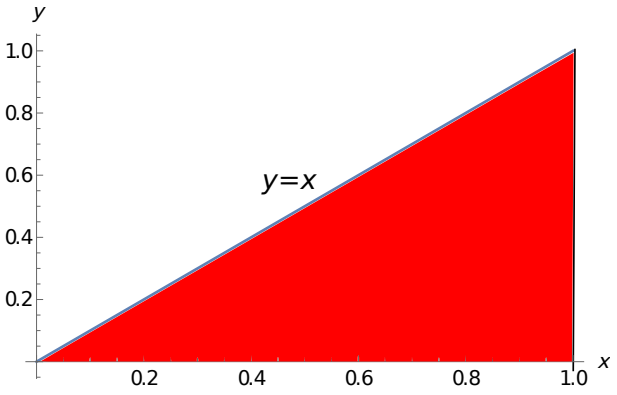
$$f(x, y) = \begin{cases} kxy & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the value of k such that $f(x, y)$ is a valid joint probability density function.

   As defined above (and similar to the one-dimensional case), for a joint probability density function to be valid, its integral must be 1 over the region of integration, i.e.

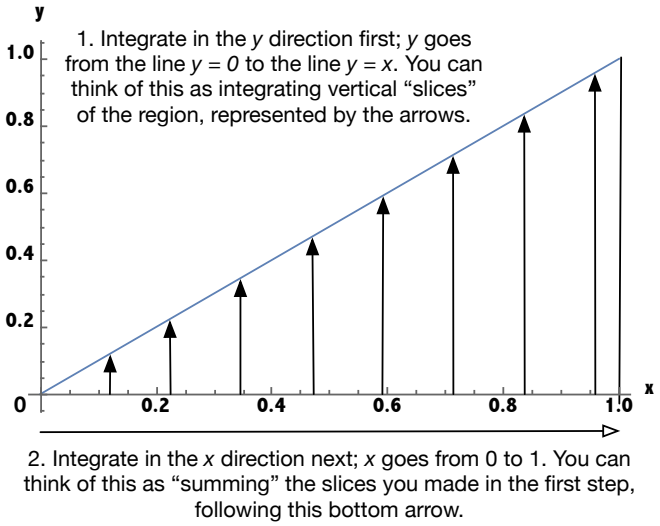
   $$\int \int f(x, y) dx\, dy = 1$$

   The bounds of the density function are the following: $y \geq 0, y \leq x$, and $x \leq 1$. This describes the triangular region illustrated below.



Whenever we have a double integral, we have two choices when we do our integration. We can integrate in x direction first, or we can integrate in the y direction first. Both ways give the same answer, but sometimes one is easier than the other. We will show both of them here.

Let's start by integrating in the y direction first.



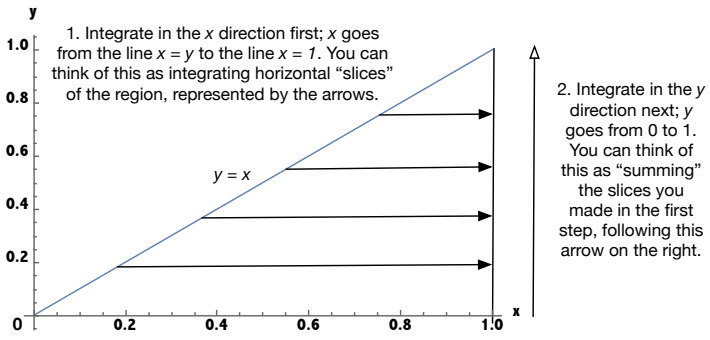
Following the directions on the picture above, we integrate y first; y goes from 0 to the diagonal line $y = x$, so those are the limits for the integral with respect to y (inner integral). Note that the upper limit is a function of x. Each integral in y is a vertical "slice" of our region. Then we integrate with respect to x. Imagining this as stacking our vertical slices side-by-side in the horizontal direction, x goes from 0 to 1, so those are the limits for the integral with respect to x (outer integral). Putting this together,

we have:

$$\begin{aligned} 1 &= \int_0^1 \int_0^x kxy \, dydx \\ &= k \int_0^1 x \frac{y^2}{2} \Big|_0^x dx \\ &= \frac{k}{2} \int_0^1 x^3 dx \\ &= \frac{k}{2} \frac{x^4}{4} \Big|_0^1 = \frac{k}{8} \end{aligned}$$

Multiplying by 8 gives us $k = 8$.

Let's do the integral the other way and verify to see if we get the same result. This time we integrate in the x direction first.



Once again following the directions on the picture above, we integrate x first; x starts at the diagonal line $y = x$ and goes to $x = 1$. The diagonal line has the equation $y = x$, which we solve for x to get $x = y$. Thus the lower limit is the function $x = y$. The upper limit is 1, so the limits of the integral with respect to x (inner integral) are y and 1. Each integral in x is a horizontal "slice" of our region. Then we integrate with respect to y. Imagining this as stacking our horizontal slices one on top of the other in the vertical direction, y goes from 0 to 1, so those are the limits for the integral with

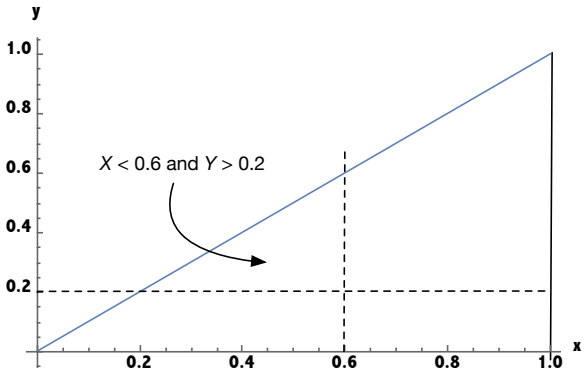
respect to y (outer integral). Putting this together, we have:

$$\begin{aligned}
1 &= \int_0^1 \int_y^1 kxy \, dxdy \\
&= k \int_0^1 y \frac{x^2}{2} \Big|_y^1 dy \\
&= \frac{k}{2} \int_0^1 y(1-y^2) dy \\
&= \frac{k}{2} \int_0^1 (y-y^3) dy \\
&= \frac{k}{2} \left(\frac{y^2}{2} - \frac{y^4}{4} \right) \Big|_0^1 = \frac{k}{8}
\end{aligned}$$

We get the same answer! Which one was easier?

2. Find $\mathbb{P}(X < 0.6 \cap Y > 0.2)$.

Here we are finding the probability that the pair (X, Y) falls in a specific region of the plane. The first step (as always) is to draw the region.



From this picture, we get the limits of integration. Let's integrate in the y direction first. Recall that we found that $k = 8$above. This gives us the integral:

$$\mathbb{P}(X < 0.6 \cap Y > 0.2) = \int_{0.2}^{0.6} \int_{0.2}^{x} 8xy \, dydx$$

This integral can be evaluated like the one above. The most important thing to know is how to set the problem up with the correct limits, but for completeness we will do

the computation below.

$$\begin{aligned}
\mathbb{P}(X < 0.6 \cap Y > 0.2) &= \int_{0.2}^{0.6} \int_{0.2}^{x} 8xy \, dydx \\
&= 8 \int_{0.2}^{0.6} x \frac{y^2}{2} \Big|_{0.2}^{x} dx \\
&= 4 \int_{0.2}^{0.6} (x^3 - 0.04x) dx \\
&= 4 \left(\frac{x^4}{4} - 0.04 \frac{x^2}{2} \right) \Big|_{0.2}^{0.6} \\
&= \left(0.6^4 - 0.2^4\right) - 0.08 \left(0.6^2 - 0.2^2\right) \\
&= 0.1024
\end{aligned}$$

We could also have integrated in the x direction first.

4.3.2 Marginal Distribution

Consider a joint distribution (Y_1, Y_2) of two continuous random variables with joint density $f(y_1, y_2)$. Y_1 and Y_2 are themselves continuous random variables, and their distributions are called the *marginal distributions* of Y_1 and Y_2. The probability densities of Y_1 and Y_2 are the *marginal densities* of Y_1 and Y_2.

How do we find the marginal densities? Recall that for the discrete case, we summed over the other variable, e.g. to find the marginal density of Y_1, we summed the joint distribution over all the values Y_2 can take. Here we do the exact same thing, except we replace summation with integration. To get the marginal density of Y_1, we integrate the joint density $f(y_1, y_2)$ over y_2. The marginal density of Y_1 is written $f_1(y_1)$, and is a function of y_1 alone. (If after the integration you still have terms involving y_2, something went wrong.) Similarly, we can find the marginal density of Y_2.

Marginal density, continuous random variables

Let Y_1 and Y_2 be continuous random variables with joint density $f(y_1, y_2)$. Then the marginal density of Y_1 is given by:

$$f_1(y_1) = \int f(y_1, y_2) dy_2$$

and the marginal density of Y_2 is given by.

$$f_2(y_2) = \int f(y_1, y_2) dy_1$$

In both cases, we integrate the joint density over the other random variable to remove it from the picture entirely. Sometimes we call this "integrating out" the other random variable.

Let's return to the first example from the joint density section.

Example. Let X and Y be random variables with joint distribution function $f(x, y)$ given by:

$$f(x,y) = \begin{cases} 8xy & 0 \le y \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the marginal densities for X and Y.

   Since our random variable are X and Y, we will denote the two marginal densities by $f_X(x)$ and $f_Y(y)$. First, we find the marginal density for X by integrating over y. Since we are integrating only in a single variable, there is only one choice for limits. Refer back to the picture of the region above. To integrate in y, we start at $y = 0$ and integrate until we reach the line $y = x$. Thus the limits of integration are 0 and x.

$$\begin{aligned} f_X(x) &= \int_0^x 8xy \, dy \\ &= 8x \frac{y^2}{2}\Big|_0^x \\ &= 4x^3 \end{aligned}$$

   With Y out of the picture, the random variable X is free to take values from 0 to 1. (The marginal distribution is one-dimensional, so there are no pesky regions to deal with!) The above expression for the marginal density is only valid for $0 \le x \le 1$. Outside that range, the marginal density is 0. Thus we write the marginal density $f_X(x)$ as:

$$f_X(x) = \begin{cases} 4x^3 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

   It is important that we write the marginal density in this way so that we know that the X is 0 outside $[0, 1]$. Since the marginal density is a valid probability density, you can check that it does in fact integrate to 1. Also note that the marginal density for X is a function of x alone; y does not appear anywhere since we integrated it out!

   Now we find the marginal density for Y by integrating over x. The limits for x are $x = 0$ and $x = y$. (See the picture of the region above if this is not clear.)

$$\begin{aligned} f_Y(y) &= \int_y^1 8xy \, dx \\ &= 8y \frac{x^2}{2}\Big|_y^1 \\ &= 4y(1 - y^2) \end{aligned}$$

With X out of the picture, the random variable Y can take values from 0 to 1, so we write the marginal density of Y as

$$f_Y(y) = \begin{cases} 4y(1-y^2) & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

2. Find the expected values for X and Y.

How do we find the expected values of X and Y. We use the marginal densities and do the same thing we do every night, Pinky - try to take over the world! The marginal densities are just standard probability densities for continuous random variables; thus we find the expected values of X and Y using the marginal densities and the formula for the expected value of a continuous random variable.

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} f_X(x)dx$$
$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} f_Y(y)dy$$

First we find the expected value for X. Notice that the limits of integration become 0 and 1, since that is the region where the marginal density is nonzero.

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x f_X(x)dx \\ &= \int_0^1 x f_X(x)dx \\ &= \int_0^1 x\, 4x^3 dx \\ &= 4\int_0^1 x^4 dx \\ &= 4/5 \end{aligned}$$

Similarly, we find the expected value of Y.

$$\begin{aligned} \mathbb{E}(Y) &= \int_{-\infty}^{\infty} y f_Y(y)dy \\ &= \int_0^1 y f_Y(y)\, dy \\ &= \int_0^1 y\, 4y(1-y^2)dx \\ &= 4\int_0^1 (y^2 - y^4)dx \\ &= 8/15 \end{aligned}$$

4.3.3 Conditional Distribution

Just as in the discrete case, we can talk about conditional distributions. In the continuous case, we will have a continuous density function. Suppose we have a joint distribution (Y_1, Y_2) of two continuous random variables with conditional density $f(y_1, y_2)$. We might be interested in the distribution of Y_1 given that $Y_2 = y_2$. Since Y_1 is a continuous random variable, we will express this as a conditional probability density.

Conditional distribution, continuous random variables

Let Y_1 and Y_2 be continuous random variables with joint probability density $f(y_1, y_2)$. Let $f_2(y_2)$ be the marginal density of Y_2. Then the conditional density of Y_1 given $Y_2 = y_2$ is:

$$f(y_1|y_2) = \frac{f(y_1, y_2)}{f_2(y_2)}$$

where $f_2(y_2) > 0$. As in the discrete case, the conditional density is the joint density divided by the marginal density. We can similarly define the conditional density of Y_2 given $Y_1 = y_1$.

Note that the conditional density can (and often will) depend on the value of the random variable we are conditioning on. For example, $f(y_1|y_2)$ may depend on the value y_2. Furthermore, the range of values y_1 can take may also depend on y_2 (we will see this in the example below). Contrast this to the marginal density, where we eliminate the other random variable entirely.

Once again, let's return to our example.

Example. Let X and Y be random variables with joint distribution function $f(x, y)$ given by:

$$f(x, y) = \begin{cases} 8xy & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the conditional density of X given $Y = y$.

   To find the conditional density $f(x|y)$, we divide the joint density $f(x, y)$ by the marginal density $f_Y(y)$. We found the marginal density above, so we have:

   $$f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{8xy}{4y(1 - y^2)} = \frac{2x}{1 - y^2}$$

   Note that this density does in fact depend on the value of y we are conditioning on. However, we are not done. The bounds on the conditional density are important and must be specified. Referring back to the picture of the region, we note that if $Y = y$,

then X can only range from the diagonal line $y = x$ to 1, i.e. X must be between y and 1. Thus the conditional density of X given $Y = y$ is given by:

$$f(x|y) = \begin{cases} \frac{2x}{1-y^2} & y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

2. Find the conditional density of Y given $X = x$.

   Again, we divide the joint density by the marginal density, where we computed the marginal density $f_X(x)$ above.

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{8xy}{4x^3} = \frac{2y}{x^2}$$

   As with the other conditional density, this depends on x. The bounds of the conditional density will also depend on x. Looking at the picture of the region, if $X = x$, then y can only range from 0 to the diagonal line $y = x$, i.e. Y must be between 0 and x. Thus the conditional density is:

$$f(y|x) = \begin{cases} \frac{2y}{x^2} & 0 \leq y \leq x \\ 0 & \text{otherwise} \end{cases}$$

3. Find the expected value of X given $Y = y$.

   A conditional density is just a probability density of a continuous random variable, so we can find its expected value using the standard formula for the expected value of a continuous random variable.

$$\begin{aligned} E[X|Y = y] &= \int_{\infty}^{\infty} x f(x|y) dx \\ &= \int_{y}^{1} x \frac{2x}{1-y^2} dx \\ &= \frac{2}{1-y^2} \int_{y}^{1} x^2 \\ &= \frac{2}{1-y^2} \frac{x^3}{3} \Big|_{y}^{1} \\ &= \frac{2(1-y^3)}{3(1-y^2)} \end{aligned}$$

   Note that we used the bounds on the conditional density in the second line above. Unsurprisingly, this depends on y.

4.3.4 Independence

We have a similar definition for independence in the case of continuous random variables.

Independence of continuous random variables

Let Y_1 and Y_2 be continuous random variables with joint density $f(y_1, y_2)$. Let $f_1(y_1)$ and $f_2(y_2)$ be the marginal densities of Y_1 and Y_2. Then Y_1 and Y_2 are independent if

$$f(y_1, y_2) = f_1(y_1)f_2(y_2) \qquad \text{for all } y_1, y_2$$

In other words, two continuous random variables are independent if their joint density is the product of the two marginal densities.

In the example we have been looking at for the past three sections, the joint density is not the product of the marginal densities, so X and Y are not independent.

4.3.5 Another Example

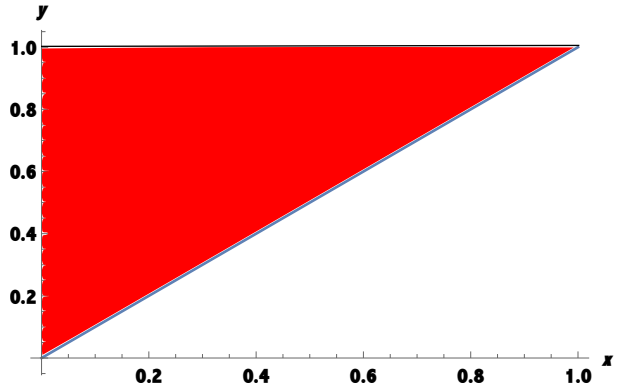
Bivariate densities are important enough that we will provide more examples problems involving them.

Example. Let X and Y be random variables with joint distribution function $f(x, y)$ given by:

$$f(x, y) = \begin{cases} cx & 0 \le x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the value of c such that $f(x, y)$ is a valid joint probability density function.

   The first step is always to draw the region.



To find the value of c, we integrate the joint density over the region and set the result equal to 1. We will integrate in the x direction first since there's already an x in the integrand and integrating in the x direction first allows both lower limits to be 0 (the other choice would work fine too):

$$\begin{aligned} 1 &= \int_0^1 \int_0^y cx dx dy \\ &= c \int_0^1 \frac{x^2}{2}\Big|_0^y dy \\ &= c \int_0^1 \frac{y^2}{2} dy \\ &= c \frac{y^3}{6}\Big|_0^1 \\ &= \frac{c}{6} \end{aligned}$$

Thus we find that $c = 6$ for this to be a valid joint probability density function.

2. Find the marginal densities for X and Y.

To do this, we integrate out each random variable in turn. Be careful to get the correct limits of integration (refer back to the picture of the region early and often.) For the marginal density of X we first integrate out y.

$$\begin{aligned} f_X(x) = \int_x^1 6x dy & \\ &= 6xy\Big|_x^1 \\ &= 6x(1-x) \end{aligned}$$

With Y removed from the picture, X can freely range from 0 to 1, so the marginal density of X is:

$$f_X(x) = \begin{cases} 6x(1-x) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

For the marginal density of Y, we first integrate out x:

$$\begin{aligned} f_Y(y) &= \int_0^y 6x dx \\ &= 3x^2\Big|_0^y \\ &= 3y^2 \end{aligned}$$

Then, with X removed from the picture, Y can freely range from 0 to 1, so the marginal density of Y is:

$$f_Y(y) = \begin{cases} 3y^2 & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

It is important that you give the correct bounds for the marginal densities.

3. What is the expected value of X?

   To find $\mathbb{E}(X)$, we use the formula for expected value of a continuous random variable with the marginal density for X.

$$\begin{aligned} \mathbb{E}(X) &= \int_{\infty}^{\infty} f_X(x)dx \\ &= \int_0^1 x6x(1-x)dx \\ &= 6\int_0^1 (x^2 - x^3)dx \\ &= 6\left(\frac{x^3}{3} - \frac{x^4}{4}\right)\Big|_0^1 \\ &= 6\left(\frac{1}{3} - \frac{1}{4}\right) \\ &= \frac{1}{2} \end{aligned}$$

   Similarly we can find the expected value of Y using the marginal density for Y.

4. What is the conditional density for X given $Y = y$?

   To get the conditional density $f(x|y)$, we first divide the joint density by the marginal density of Y:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{6x}{3y^2} = \frac{2x}{y^2}$$

   What are the bounds on the conditional density? Refer back to the picture of the region above. Given that $Y = y$, X can range from 0 to y, thus the conditional density is:

$$f(x|y) = \begin{cases} \dfrac{2x}{y^2} & 0 \le x \le y \\ 0 & \text{otherwise} \end{cases}$$

5. What is the expected value of X given $Y = y$?

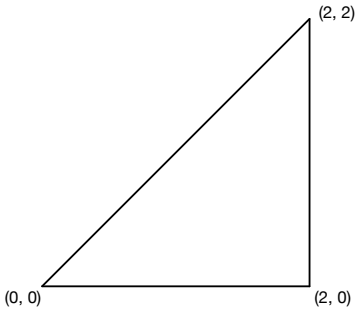
Here, we use the formula for the expected value of a continuous random variable with the conditional density $f(x|y)$.

$$\begin{aligned}\mathbb{E}(X|Y=y) &= \int_{-\infty}^{\infty} f(x|y)dx \\ &= \int_0^y x\frac{2x}{y^2}dx \\ &= \int_0^y \frac{2x^2}{y^2}dx \\ &= \frac{2x^3}{3y^2}\Big|_0^y \\ &= \frac{2y}{3}\end{aligned}$$

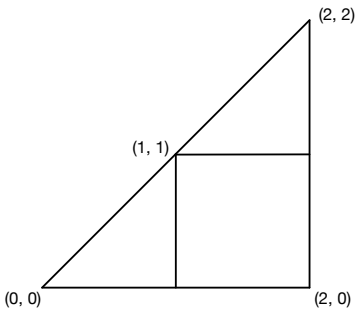
4.3.6 Joint Uniform Distribution

Here we will consider the bivariate uniform distribution. Recall the for the continuous uniform distribution, the probability of an interval is proportional to its length. For the two-dimensional uniform distribution, the probability of a region is proportional to its area. Let's look at an example.

Example. Let X and Y have a joint uniform distribution on the equilateral right triangle with sides of length 2.



1. What is the probability that the pair (X, Y) lies within the small square below (corners (1, 1) and (2, 0))?



The small square is half the area of the right triangle, so since this is the uniform distribution, the probability that (X, Y) lies within the small square is $1/2$.

2. What is the joint probability density of (X, Y)?

   Recall that in the one-dimensional case, the uniform probability density is a constant. If $X \sim \text{Uniform}(a, b)$ then X has density $f(x) = \frac{1}{b-a}$. This is just 1 divided by the length of the interval. This extends to the two-dimensional case. For a joint uniform distribution over a region, the joint uniform density is given by $1/A$, where A is the area of the region. Just as in the one dimensional case, the joint uniform density is only defined on a region with *finite* area.

   Since the area of the triangle is 2, the joint uniform density is $f(x, y) = 1/2$. But we are not done! We need to specify the bounds of the uniform density, which will depend on the geometry of the region. Outside the region, the uniform density must be 0. Looking at the picture of the region above, we see that $y \geq 0$, $y \leq x$, and $x \leq 2$. Thus the joint density is:

$$f(x, y) = \begin{cases} \frac{1}{2} & 0 \leq y \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

3. What are the marginal densities of X and Y?

   To find these, we take the joint density and integrate out each variable in turn. As always, we refer to the picture of the region to get the correct limits of integration. For the marginal density of X we integrate over y:

$$\begin{aligned} f_X(x) &= \int_0^x \frac{1}{2} dy \\ &= \frac{1}{2} y \Big|_0^x \\ &= \frac{x}{2} \end{aligned}$$

   With the correct bounds, the marginal density of X is:

$$f_X(x) = \begin{cases} \frac{x}{2} & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

   Why are these the correct bounds for the marginal density? For the marginal density of Y we integrate over x:

$$\begin{aligned} f_Y(y) &= \int_y^2 \frac{1}{2} dx \\ &= \frac{1}{2} x \Big|_y^1 \\ &= \frac{2 - y}{2} \end{aligned}$$

With the correct bounds, the marginal density of Y is:

$$f_Y(y) = \begin{cases} \frac{2-y}{2} & 0 \le y \le 2 \\ 0 & \text{otherwise} \end{cases}$$

4.4 Expected value of a function of two random variables

Recall that in the one-dimensional case, we defined the expected value of a function $g(y)$ of a random variable Y to be:

$$\mathbb{E}[g(Y)] = \sum_{\text{all } y} g(y)\, p(y) \qquad Y \text{ is a discrete random variable with pmf } p(y)$$

$$\mathbb{E}[g(Y)] = \int_{\infty}^{\infty} g(y) f(y) dy \qquad Y \text{ is a continuous random variable with density } f(y)$$

The expected value of a function of two random variables is defined in the same way, except we use the joint pmf or joint density.

Expected value of a function random variables

Let Y_1 and Y_2 be two discrete random variables with joint pmf $p(y_1, y_2)$. Let $g(y_1, y_2)$ be a real-valued function of y_1 and y_2. Then the expected value of $g(Y_1, Y_2)$ is given by:

$$\mathbb{E}[g(Y_1, Y_2)] = \sum_{\text{all } y_1} \sum_{\text{all } y_2} g(y_1, y_2) p(y_1, y_2)$$

Let Y_1 and Y_2 be two continuous random variables with joint density $p(y_1, y_2)$. Let $g(y_1, y_2)$ be a real-valued function of y_1 and y_2. Then the expected value of $g(Y_1, Y_2)$ is given by:

$$\mathbb{E}[g(Y_1, Y_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y_1, y_2) f(y_1, y_2) dy_1 dy_2$$

A very useful function of two random variables is the product of the two random variables, i.e. $g(Y_1, Y_2) = Y_1 Y_2$. We will meet this function again when we discuss covariance, so let's do an example of this using the same random variables X and Y we have used in the sections on joint density, marginal, and conditional distributions.

Example. Let X and Y be random variables with joint distribution function $f(x, y)$ given by:

$$f(x, y) = \begin{cases} 8xy & 0 \le y \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $\mathbb{E}(XY)$.

We are looking for the expected value of the function $g(XY) = XY$. Using the density $f(x, y)$ and the definition of the expected value of a function of two random variables, we have:

$$\mathbb{E}(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dxdy$$

As above, we need to choose an order of integration. We will choose here to integrate in the y direction first. Looking at the picture of region above to get the correct limits, we have:

$$\begin{aligned}
\mathbb{E}(XY) &= \int_0^1 \int_0^x xy(8xy)dydx \\
&= 8\int_0^1 \int_0^x x^2y^2dydx \\
&= 8\int_0^1 x^2\frac{y^3}{3}\Big|_0^x dx \\
&= 8\int_0^1 x^2\frac{x^3}{3}dx \\
&= \frac{8}{3}\int_0^1 x^5dx \\
&= \frac{8}{3}\frac{x^6}{6}\Big|_0^1 \\
&= \frac{8}{18} = \frac{4}{9}
\end{aligned}$$

We can use the definition of the expected value of a function of two random variables to show that if two random variables are independent, the expected value of the product is the product of the expected values.

Expected value of the product of independent random variables

---

Let Y_1 and Y_2 be two independent random variables. Then

$$\mathbb{E}(Y_1Y_2) = \mathbb{E}(Y_1)\mathbb{E}(Y_2)$$

An analagous result holds for the product of any number of independent random variables.

4.4.1 Covariance and Correlation

We have talked about independence of random variables several times. Heuristically, two random variables are independent if their outcomes do not affect each other. Suppose we have two random variables X and Y. There are two extreme cases to consider.

1. X and Y are independent, so they don't affect each other at all. The classic example of this is if X and Y are the result of two sequential coin flips.

2. X and Y are completely dependent, i.e. the output of one random variable completely determines the output of the other random variable. In this case, the two random variables are functions of each other. An example would be if $Y = 4X$. In this case, knowledge of output of either random variable gives you knowledge of the output of the other random variable.

There is an entire spectrum between these two extremes. The *covariance* of two random variables is a quantitative measure of the degree of dependence of two random variables. It is defined as follows.

Covariance of two random variables

Let Y_1 and Y_2 be two random variables with expected values $\mathbb{E}(Y_1)$ and $\mathbb{E}(Y_2)$. Then the *covariance* of Y_1 and Y_2 is defined by

$$Cov(Y_1, Y_2) = \mathbb{E}[(Y_1 - \mathbb{E}(Y_1))(Y_2 - \mathbb{E}(Y_2))]$$

A larger absolute value of the covariance indicates a greater dependence between Y_1 and Y_2. Unfortunately, covariance is difficult to interpret since the value of the covariance depends on the scale used to measure the random variables. To solve this problem, we standardize to get what we call the *correlation coefficient*. When you read a scientific study which posits a relationship between two variables, the strength of that relationship is usually given in terms of the correlation coefficient.

Correlation coefficient

Let Y_1 and Y_2 be two random variables, and let their standard deviations be σ_1 and σ_2. Then the *correlation coefficient* is denoted ρ, and is defined by:

$$\rho = \frac{Cov(Y_1, Y_2)}{\sigma_1 \sigma_2}$$

The correlation coefficient is always between -1 and 1, i.e. $-1 \leq \rho \leq 1$.

How do we interpret the correlation coefficient?

1. If $\rho > 0$, then the two random variables are positively correlated, i.e. Y_2 increases as Y_1 increases. If $\rho = 1$, we have perfect correlation; Y_2 and Y_1 are completely dependent, and all points (Y_1, Y_2) fall on a straight line with positive slope.

2. If $\rho < 0$, then the two random variables are negatively correlated, i.e. Y_2 decreases as Y_1 increases. If $\rho = 1$, we again have perfect correlation; Y_2 and Y_1 are completely dependent, and all points (Y_1, Y_2) fall on a straight line with negative slope.

3. If $\rho = 0$, then there is no correlation between the two random variables. It is important to note that this does not necessarily imply that the two random variables are independent.

Just like with the variance, the covariance is not generally computed directly. Instead, we use the Magic Covariance Formula, which is analogous to the Magic Variance Formula.

Magic Covariance Formula

Let Y_1 and Y_2 be two random variables with expected values $\mathbb{E}(Y_1)$ and $\mathbb{E}(Y_2)$. Then the *covariance* of Y_1 and Y_2 is given by

$$Cov(Y_1, Y_2) = \mathbb{E}(Y_1 Y_2) - \mathbb{E}(Y_1)\mathbb{E}(Y_2)$$

To verify this formula, we expand the product in the definition of covariance and use linearity of expectation. Letting $\mathbb{E}(Y_1) = \mu_1$ and $\mathbb{E}(Y_2) = \mu_2$,

$$\begin{aligned} Cov(Y_1, Y_2) &= \mathbb{E}[(Y_1 - \mu_1)(Y_2 - \mu_2)] \\ &= \mathbb{E}(Y_1 Y_2 - \mu_1 Y_2 - \mu_2 Y_1 + \mu_1 \mu_2) \\ &= \mathbb{E}(Y_1 Y_2) - \mu_1 \mathbb{E}(Y_2) - \mu_2 \mathbb{E}(Y_1) + \mu_1 \mu_2 \\ &= \mathbb{E}(Y_1 Y_2) - \mu_1 \mu_2 - \mu_2 \mu_1 + \mu_1 \mu_2 \\ &= \mathbb{E}(Y_1 Y_2) - \mu_1 \mu_2 \end{aligned}$$

Let's use the Magic Covariance Formula to compute the covariance of the random variables X and Y which we used as an example above for joint, marginal, and conditional densities.

Example. Let X and Y be random variables with joint distribution function $f(x, y)$ given by:

$$f(x, y) = \begin{cases} 8xy & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the covariance of X and Y.

In the section on marginal distributions of continuous random variables, we found that $\mathbb{E}(X) = 4/5$ and $\mathbb{E}(Y) = 8/15$. In the section on the expected value of a function of two random variables, we found that $\mathbb{E}(XY) = 4/9$. Using the Magic Covariance Formula,

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 4/9 - (4/5)(8/15) = 4/225$$

What happens if two random variables are independent?

Covariance of independent random variables

Let Y_1 and Y_2 be two independent random variables. Then

$$Cov(Y_1, Y_2) = 0$$

To see this, we use the Magic Covariance Formula together with the result from above that the expected value of a product of two independent random variables is the product of the expected values.

$$\begin{aligned} Cov(Y_1, Y_2) &= \mathbb{E}(Y_1 Y_2) - \mathbb{E}(Y_1)\mathbb{E}(Y_2) \\ &= \mathbb{E}(Y_1)\mathbb{E}(Y_2) - \mathbb{E}(Y_1)\mathbb{E}(Y_2) = 0 \end{aligned}$$

The converse, however, is not generally true (except in a few very special circumstances). In other words, if the covariance of two random variables is 0, we *cannot* conclude that the random variables are independent. You will have a homework problem about this.

The final result in this section concerns the variance of the sum of two random variables. Part of result was stated earlier and was used to prove the variance of the binomial distribution. The full result is below.

Variance of the sum of two random variables

Let Y_1 and Y_2 be two random variables. Then

$$Var(Y_1 + Y_2) = Var(Y_1) + Var(Y_2) + 2Cov(Y_1, Y_2)$$

If Y_1 and Y_2 are independent, then

$$Var(Y_1 + Y_2) = Var(Y_1) + Var(Y_2)$$

We can extend the first result to the case of a sum of more than two random variables, but the result is cumbersome. In the second case, however, the result extends easily. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables, then

$$Var(Y_1 + Y_2 + \cdots + Y_n) = Var(Y_1) + Var(Y_2) + \cdots + Var(Y_n)$$

To see this, we use the Magic Variance Formula and linearity of expectation:

$$\begin{aligned} Var(Y_1 + Y_2) &= \mathbb{E}[(Y_1 + Y_2)^2] - [\mathbb{E}(Y_1 + Y_2)]^2 \\ &= \mathbb{E}(Y_1^2 + 2Y_1Y_2 + Y_2^2) - [\mathbb{E}(Y_1) + \mathbb{E}(Y_2)]^2 \\ &= \mathbb{E}(Y_1^2) + 2\mathbb{E}(Y_1Y_2) + \mathbb{E}(Y_2^2) - [\mathbb{E}(Y_1)]^2 - 2\mathbb{E}(Y_1)\mathbb{E}(Y_2) - [\mathbb{E}(Y_2)]^2 \\ &= (\mathbb{E}(Y_1^2) - [\mathbb{E}(Y_1)]^2) + (\mathbb{E}(Y_2^2) - \mathbb{E}(Y_2)]^2) + 2[\mathbb{E}(Y_1Y_2) - \mathbb{E}(Y_1)\mathbb{E}(Y_2)] \\ &= Var(Y_1) + Var(Y_2) + 2Cov(Y_1, Y_2) \end{aligned}$$

where in the last line we used both the Magic Variance Formula and the Magic Covariance Formula. If Y_1 and Y_2 are independent, the covariance is 0, so we get $Var(Y_1 + Y_2) = Var(Y_1) + Var(Y_2)$.

5 Sampling Distributions

5.1 Introduction

In this section, we transition from the world of probability to the world of statistics. The field of probability is concerned with making predictions about probability distributions which are completely known. The contents of a deck of cards, for example, is known, and thus we can derive probabilistic statements about the likelihood of certain cards being flipped in a game of Texas Hold'em poker. The field of statistics addresses the converse problem. It allows us to make inferences about the distribution of a population based on a small sample drawn from that population. We then use the tools from probability to make probabilistic statements about the accuracy of our inferences. For example, we can model the distribution of Rhode Island voter preferences in the gubernatorial election with a binomial distribution. The parameters of the distribution are $n = 754,224$ (the number of registered voters as of election day, 2014) and p, the proportion of voters who prefer Gina Raimondo. The parameter p is unknown, and unless we accurately survey every single registered voter, we will not know its exact value. Such a survey is logistically and financially unfeasible, thus we turn to statistics. We poll a smaller sample of voters (say, 1000) and find the proportion of voters in that smaller sample who prefer Raimondo. This sample proportion is designated $\hat{p}$, where the "hat" indicates that it is an estimator for the true value p based off of a sample drawn from the larger population. Using the tools of statistics, we will be able to quantitatively evaluate how close our estimate $\hat{p}$ is to the true value p.

5.2 Statistics

We will use the following setup for our discussion.

1. We have a large population and are interested in studying a particular quantitative feature of that population. For example, the population could be the registered voters in Rhode Island, and we are interested in the yes/no question "Are you going to vote for Raimondo?". Another population could be the ball bearings produced by the factory, and we are interested in their diameter.

2. The population feature can be characterized by a pmf $p(y)$ (if it is discrete, as in the case of the number of voters who prefer Raimondo) or a density function $f(y)$ (if it is continuous, as in the case of the ball bearing diameters).

3. The population feature will have certain parameters. All populations have a mean and a variance, which will be designated by μ and σ^2. We will refer to these as the population mean and the population variance. Populations may also have other parameters. As an example, if we model a population with a uniform distribution on an unknown interval $[a, b]$, the endpoints a and b are parameters of the population. Much of statistics is concerned with estimating parameters of populations.

4. We take a small sample from the population. Let n be the sample size, and let the random variables $Y_1, \ldots, Y_n$ be the samples we take from the population.

5. We will assume that the sample size is small enough relative to the population size that the samples $Y_1, \ldots, Y_n$ are independent and have the same distribution $p(y)$ or $f(y)$ as the population. As an example, recall that we mentioned before that sampling without replacement (as in political polling) was approximately binomial if the sample size was less than 1/20 of the population size.

6. A *statistic* is a function of our samples $Y_1, \ldots, Y_n$. The function can only involve the samples themselves and known constants such as the sample size n. Since a statistic is a function of random variables, it is itself a random variable, thus we can characterize its distribution using the tools of probability.

In this first section we will be concerned with the probability distribution of various statistics. Before we continue, we will give a table of the most common statistics we will encounter. Let $Y_1, \ldots, Y_n$ be independent and identically distributed (iid) samples drawn from a population which is distributed according to a pmf $p(y)$ or a density $f(y)$. Let μ and σ^2 be the population mean and variance.

Statistic	Definition
sample mean	$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
sample variance	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
smallest order statistic (sample minimum)	$Y_1 = \min(Y_1, \ldots, Y_n)$
largest order statistic (sample maximum)	$Y_n = \max(Y_1, \ldots, Y_n)$
sample range	$R = Y_n - Y_1$

Note that you have to compute $\bar{Y}$ before you can compute S^2. The $n-1$ in the denominator of the sample variance may seem a bit mysterious, but we will see in a few classes why that makes sense. For now, we will look at the sample mean. To get the sample mean, we add together the samples and divide by the number of samples, so this is the empirical mean we have discussed before. The sample mean is a random variable, so we can find its mean and variance. Using linearity of expectation and the fact that all the Y_i have expected value of μ,

$$\begin{aligned}
\mathbb{E}(\bar{Y}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n\mu \\
&= \mu
\end{aligned}$$

Thus the expected value of the sample mean is the population mean μ.

For the variance, we recall that $Var(aY) = a^2 Var(Y)$ and that for independent random variables, $Var(Y_1 + \cdots + Y_n) = Var(Y_1) + \cdots + Var(Y_n)$. Using this plus the fact the all the Y_i have variance of σ^2,

$$\begin{aligned} Var(\bar{Y}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{1}{n^2} n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

The variance of the sample mean is thus the population variance divided by n. As n gets larger, this variance gets smaller, which makes intuitive sense since there should be less spread if we average more samples together.

Note that while we have characterized the mean and variance of the sample mean, we have said nothing about its distribution. In general, there is not much additional we can say. In the special case where the population has a normal distribution, however, we can say a lot more.

5.3 Sampling Distributions for Normally Distributed Populations

5.3.1 Distribution of Sample Mean

Suppose we have a population which is normally distributed. Then the sample mean is also normally distributed.

> Let $Y_1, \ldots, Y_n$ be a sample of size n drawn from a population which has a normal distribution with mean μ and variance σ^2. Then the sample mean
>
> $$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
>
> has a normal distribution with mean $\mathbb{E}(\bar{Y}) = \mu$ and variance $Var(\bar{Y}) = \sigma^2/n$.

The mean and variance of the sample mean was shown above. To show the sample mean has a normal distribution, we can use the method of moment generating functions or characteristic functions, which will not be done in this course for sake of time.

Since the sample mean $\bar{Y}$ is normally distributed with mean μ and variance σ^2/n,

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right)$$

is a standard normal random variable. (Recall that to get a standard normal, we subtract the mean and divide by the standard deviation, which is the square root of the variance.)

Let's use this in an example.

Example. A ball bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation σ mm. We unfortunately have lost the manual for the machine, so we do not know the value of μ. We call the company to get more information, but all they can tell us is that $\sigma = 0.1$.

1. We take a sample of 16 ball bearings from the machine and compute the sample mean $\bar{Y}$. Find the probability that $\bar{Y}$ is within 0.02 mm of the true mean μ.

   By what we learned above, the sample mean $\bar{Y}$ has a normal distribution with mean μ and variance $\sigma^2/n = 0.1^2/16$. The standard deviation of $\bar{Y}$ is $\sigma = 0.1/\sqrt{16}$. Thus we can find the probability by converting to the standard normal distribution.

$$\begin{aligned}
\mathbb{P}(|\bar{Y} - \mu| \leq 0.02) &= \mathbb{P}(-0.02 \leq (\bar{Y} - \mu) \leq 0.02) \\
&= \mathbb{P}\left(\frac{-0.02}{\sigma/\sqrt{n}} \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.02}{\sigma/\sqrt{n}}\right) \\
&= \mathbb{P}\left(\frac{-0.02}{0.1/\sqrt{16}} \leq Z \leq \frac{0.02}{0.1/\sqrt{16}}\right) \\
&= \mathbb{P}(-0.8 \leq Z \leq 0.8) \\
&= 0.7881 - 0.2119 \\
&= 0.5762
\end{aligned}$$

2. How many ball bearings should we sample if we wish the sample mean to be within 0.02 mm of μ with probability 0.95?

   We can use the 68-95-99.7 rule here. We want the sample mean to be with 0.02 mm of the mean with probability 0.95. Another way of writing this is that we want the population mean μ to lie within two sample standard deviations of the sample mean. Thus we want the sample standard deviation to be $0.02/2 = 0.01$. Since the sample standard deviation is $\sigma/\sqrt{n} = 0.1/\sqrt{n}$, all we have to is set these equal to each other and solve for n.

$$\begin{aligned}
\frac{0.1}{\sqrt{n}} &= 0.01 \\
\sqrt{n} &= \frac{0.1}{0.01} = 10 \\
n &= 100
\end{aligned}$$

3. How many ball bearings should we sample if we wish the sample mean to be within 0.02 mm of μ with probability 0.99?

   Here we can't rely on the 68-95-99.7 rule and have to actually do the math. We want:

$$\mathbb{P}(|\bar{Y} - \mu| \leq 0.02) = \mathbb{P}(-0.02 \leq (\bar{Y} - \mu) \leq 0.02) = 0.99$$

   Dividing this by the sample standard deviation $\sigma/\sqrt{n}$ to convert to the standard normal, we get

$$\begin{aligned} 0.99 &= \mathbb{P}\left(\frac{-0.02}{\sigma/\sqrt{n}} \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.02}{\sigma/\sqrt{n}}\right) \\ &= \mathbb{P}\left(\frac{-0.02}{0.1/\sqrt{n}} \leq Z \leq \frac{0.02}{0.1/\sqrt{n}}\right) \\ &= \mathbb{P}(-0.2\sqrt{n} \leq Z \leq 0.2\sqrt{n}) \end{aligned}$$

   At this point, we take complements and use the symmetry of the normal distribution.

$$\begin{aligned} \mathbb{P}(Z \leq -0.2\sqrt{n} \text{ or } Z \geq 0.2\sqrt{n}) &= 1 - 0.99 = 0.01 \\ \mathbb{P}(Z \leq -0.2\sqrt{n}) &= 0.005 \end{aligned}$$

   Consulting a Z table, we find that $\mathbb{P}(Z \leq -2.57) = 0.0051$ and $\mathbb{P}(Z \leq -2.58) = 0.0049$. Choosing the first one (you can choose either one) and setting that value of z equal to $-0.2\sqrt{n}$, we get:

$$\begin{aligned} -0.2\sqrt{n} &= -2.57 \\ \sqrt{n} &= 12.85 \\ n &= 165.123 \end{aligned}$$

   Rounding up to the nearest integer, we need a sample size of 166 or more to have 99% confidence that our sample mean is within 0.02 of the true population mean.

5.4 Other distributions

We will now briefly discuss two other useful distributions, the chi-square and t distribution. Since computations with these distributions is done via tables and not by using the densities directly, we will not give the densities here. If you are curious what these densities look like, you can consult Wikipedia or a statistics textbook.

5.4.1 Chi-square distribution

The chi-square distribution is the distribution of the sum of the squares of independent, standard normal random variables. Since the variance of a random variable is computed by first finding the expected value of the square of the random variable, the chi-square distribution is used when we make inferences about the variance of a population.

Let $Y_1, \ldots, Y_n$ be a sample of size n drawn from a population which has a normal distribution with mean μ and variance σ^2. Then the random variables $Z_i = (Y_i - \mu)/\sigma$ are independent, standard normal random variables, and

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

has a χ^2 (chi-square) distribution with n degrees of freedom (df).

To use the chi-square distribution, we use a chi-square table, which is provided on the course website. Chi-square tables are a little weird. Here's how we use them. Suppose Y has a chi-square distribution with 10 df. We will be looking at the row labeled 10 df. Now look across the top row. You see $\chi^2_{.995}, \chi^2_{.990}, \chi^2_{.975}$, etc. The decimals there are probabilities. What do they represent? Let's look at the row for 10 df. The first entry in that row, in the column labeled $\chi^2_{.995}$, is the value of y for which $\mathbb{P}(Y > y) = 0.995$. The entry in that box is 2.156, so $\mathbb{P}(Y > 2.156) = 0.995$. If we want the probability that Y is less than or equal to 2.156, then we just subtract from 1 to get $\mathbb{P}(Y \leq 2.156) = 1 - 0.995 = 0.005$. Similarly, looking at the 6th column, we have $\mathbb{P}(Y > 15.987) = 0.100$, so $\mathbb{P}(Y \leq 15.987) = 1 - 0.100 = 0.90$. Note that the probabilities across the top are "extreme" probabilities, i.e. very high or very low. In general, those are the probabilities we will be interested in. For other values, you can use a statistical package such as R.

First, let's practice using the chi-square table, then we will use the chi-square distribution in an example.

Example. Let $Z_1, \ldots, Z_6$ be independent samples from a standard normal distribution. Find a value a such that

$$\mathbb{P}\left(\sum_{i=1}^{6} Z_i^2 \leq a \right) = 0.95$$

By the above, $\sum_{i=1}^{6} Z_i^2$ has a chi-square distribution with 6 df. We want the probability that this is less than a to be 0.95, which is equivalent to

$$\mathbb{P}\left(\sum_{i=1}^{6} Z_i^2 > a \right) = 1 - 0.95 = 0.05$$

Thus we look at the chi-square table in the 6 df row and find the entry in the column labeled $\chi^2_{.050}$ to get 12.592. Thus we conclude that $a = 12.592$.

How do we use this practically. Recall from above that the sample variance is given by:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Then we can say the following about the distribution of S^2:

Let $Y_1, \ldots, Y_n$ be a sample of size n drawn from a population which has a normal distribution with mean μ and variance σ^2. Let S^2 be the sample variance as given above. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

has a χ^2 (chi-square) distribution with $n-1$ degrees of freedom (df).

Again, the proof of this is omitted for sake of time. Let's go back to our ball bearing example.

Example. A ball bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation σ mm. Once again, all we know is that $\sigma = 0.1$. Suppose we select 10 samples and compute the sample variance S^2. Find an interval $[a, b]$ such that the probability that S^2 lies in the interval is 0.90, i.e.

$$\mathbb{P}(a \leq S^2 \leq b) = 0.90$$

Multiplying by $(n-1)$ and dividing by σ^2, we get:

$$0.90 = \mathbb{P}(a \leq S^2 \leq b) = \mathbb{P}\left(\frac{(n-1)a}{\sigma^2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)b}{\sigma^2}\right)$$

From above, we know that $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $(n-1) = 10-1 = 9$ df. Let's call this X. Then we have:

$$0.90 = \mathbb{P}(a \leq S^2 \leq b) = \mathbb{P}\left(\frac{(n-1)a}{\sigma^2} \leq X \leq \frac{(n-1)b}{\sigma^2}\right)$$

where X has a chi-square distribution with 9 df. There are many ways we can choose a and b to make this happen, but the easiest way is to "split the difference". Taking complements, we can write this as:

$$\mathbb{P}\left(X \leq \frac{(n-1)a}{\sigma^2} \text{ or } X \geq \frac{(n-1)b}{\sigma^2}\right) = 1 - 0.90 = 0.10$$

Splitting the probability of 0.10 in half, we will find a and b such that:

$$\mathbb{P}\left(X \leq \frac{(n-1)a}{\sigma^2}\right) = 0.05 \text{ and } \mathbb{P}\left(X \geq \frac{(n-1)b}{\sigma^2}\right) = 0.05$$

Since the chi-square table gives us probabilities of being greater than a value, this is equivalent to:

$$\mathbb{P}\left(X > \frac{(n-1)a}{\sigma^2}\right) = 0.95 \text{ and } \mathbb{P}\left(X > \frac{(n-1)b}{\sigma^2}\right) = 0.05$$

(these are continuous random variables, so the strictness of the inequalities does not matter). Now we look at the chi-square table. We need the values for 0.95 and 0.05 in the 9 df row,

which are 3.325 and 16.919. Then we can solve for a and b, since we know $n = 10$ and $\sigma = 0.1$.

$$3.325 = \frac{(n-1)a}{\sigma^2} = \frac{9a}{0.1^2}$$
$$a = \frac{3.325 \cdot 0.1^2}{9} = 0.0037$$

$$16.919 = \frac{(n-1)b}{\sigma^2} = \frac{9b}{0.1^2}$$
$$a = \frac{16.919 \cdot 0.1^2}{9} = 0.0187$$

Thus our interval $[a, b]$ is [0.0037, 0.0187], which has 90% probability of including our sample variance S^2. Note that the true population variance $\sigma^2 = 0.1^2 = 0.01$ lies in that interval. If our sample variance is not in that interval, perhaps we should be suspicious of either our sampling technique or that something is going wrong with the machine.

5.4.2 t distribution

What happens when we do not know the population standard deviation. Recall that if we have normal population whose mean and standard deviation is known, the quantity

$$\sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right)$$

is a standard normal random variable (essentially a normalized version of the sample mean). If the population standard deviation is not known, then we can replace σ by our sample statistic S (the square root of the population variance S^2) to get

$$\sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right)$$

The probability distribution of this quantity is known as the t-distribution.

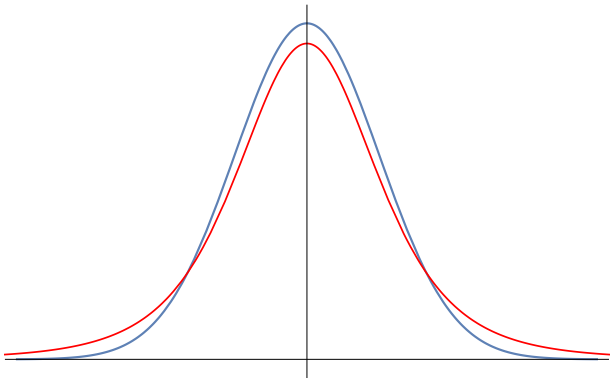
Let $Y_1, \ldots, Y_n$ be a sample of size n drawn from a population which has a normal distribution with mean μ and unknown variance. Let S^2 be the sample variance as given above. Then

$$T = \sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right)$$

has a t-distribution with $n - 1$ degrees of freedom (df).

The t distribution, like the standard normal Z distribution, is bell-shaped, has a mean of 0, and is symmetric about its mean. However, it has thicker tails than the Z distribution, so

outlier values are more common. The following plot shows the standard normal distribution in blue and the t distribution with 4 df in red.



As the number of df gets larger (i.e. approaches infinity), the t distribution approaches the standard normal distribution. Like the Z distribution and chi-square distribution, we use tables to work with the t-distribution. Let's do an example of this.

Example. Once again, we have a bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation σ mm. We not only have lost the manual for the machine, but the company has gone bankrupt so they cannot tell us anything about the machine. We take a sample of 16 ball bearings from the machine and compute the sample mean $\bar{Y}$. We also compute the sample standard deviation S and find that $S = 0.1$. Find a range $[a, b]$ such that the probability that the difference between the sample mean and the population mean μ falls within the interval with a probability of 0.90.

We are interested in finding a and b such that $\mathbb{P}(a \leq (\bar{Y} - \mu) \leq b) = 0.90$. Multiplying by $\sqrt{n}$ and dividing by S, we get:

$$\mathbb{P}\left(\frac{\sqrt{n}a}{S} \leq \sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right) \leq \frac{\sqrt{n}b}{S}\right) = 0.90$$

$$\mathbb{P}\left(\frac{\sqrt{16}a}{0.1} \leq T \leq \frac{\sqrt{16}b}{0.1}\right) = 0.90$$

$$\mathbb{P}\left(40a \leq T \leq 40b\right) = 0.90$$

Now we look at our t-table. Note that the t-table gives probabilities for the upper tail of the t-distribution and that the t-distribution is symmetric about the mean. We have a sample of 16, so we want 15 df. We will take a symmetric interval about the mean, so we want $\mathbb{P}(T \leq 40a) = 0.05$ and $\mathbb{P}(T \leq 40b) = 0.05$. For the upper tail, we look at the t-table in the row for 15 df and go across until we get to 0.05, which gives us a t-value of 1.753. By symmetry, the lower t-value is -1.753. Thus we have $40b = 1.753$, which implies $b = 0.0438$. By symmetry, $a = -0.0438$. Thus we are 90% confident that the difference between the sample and population means falls within the interval $[-0.0438, 0.0438]$.

What if we actually know the population standard deviation, as in the example above. Then we can use the standard normal distribution instead of the t-distribution. How does the

interval we get compare to that where the population standard deviation is not known. We expect to get a narrower interval in this case, since we have more information. Let's see if that is indeed the case.

$$\mathbb{P}\left(\frac{\sqrt{n}a}{\sigma} \leq \sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) \leq \frac{\sqrt{n}a}{\sigma}\right) = 0.90$$

$$\mathbb{P}\left(\frac{\sqrt{16}a}{0.1} \leq Z \leq \frac{\sqrt{16}a}{0.1}\right) = 0.90$$

$$\mathbb{P}\left(40a \leq Z \leq 40b\right) = 0.90$$

So far, this is exactly the same except we have Z in place of T. Again, we will look for a symmetric interval about the mean, so we want $\mathbb{P}(Z \leq 40a) = 0.05$ and $\mathbb{P}(Z \leq 40b) = 0.05$. For the lower tail, we look at the Z-table and can choose either $z = -1.65$ (corresponding to a probability of 0.0495) or $z = -1.64$ (corresponding to a probability of 0.0505). Choosing $z = -1.64$ we get $40a = -1.64$, so $a = -0.041$. By symmetry, $b = 0.041$. Thus we are 90% confident that the difference between the sample and population means falls within the interval $[-0.041, 0.041]$. As predicted, this is a narrower interval than the case where we used the t-distribution (where the population standard deviation was unknown).

5.5 Central Limit Theorem

Take n independent samples $Y_1, \ldots, Y_n$ from a population with population mean μ and population variance σ^2. Then we showed earlier that the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ has mean $\mathbb{E}(\bar{Y}) = \mu$ and variance $Var(\bar{Y}) = \sigma^2/n$. In the case where the population is normally distributed, then $\bar{Y}$ is also normally distributed, i.e. $\bar{Y} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$.

What is the distribution of $\bar{Y}$ if the population distribution is not normal? It turns out that as long as the sample size is large ($n \geq 30$ is often used as a guideline for "large"), the sample mean $\bar{Y}$ can be approximated by a normal distribution. This remarkable result is known as the *central limit theorem*.

Central Limit Theorem

Let $Y_1, \ldots, Y_n$ be independent and identically distributed (iid) random variables with $\mathbb{E}(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$. Let $\bar{Y}$ be the sample mean, i.e.

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Define U_n by

$$U_n = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$$

U_n is a normalized version of the sample mean $\bar{Y}$, which we obtain by subtracting the mean and dividing by the standard deviation. Then the distribution of U_n converges to

that of a standard normal distribution as $n \to \infty$. Mathematically, this means that the CDF of U_n converges to the standard normal CDF, i.e.

$$\lim_{n\to\infty} \mathbb{P}(U_n \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Practically, the means that for a large sample size (approx $n \geq 30$), the sample mean $\bar{Y}$ is asymptotically normally distributed with mean μ and variance σ^2/n.

The proof of the central limit theorem is complicated and will be omitted for the sake of time. It involves the use of either moment generating functions or characteristic functions and some clever approximations.

Let's do some examples.

Example. Suppose standardized test scores historically have a mean of 60 and variance of 64. New versions are created for every administration of the test. A new version of the test is given to 100 students, and the mean score of those 100 students is 58. How likely is it that there is something wrong with new version of the test.

We will compute the probability that the sample mean is less than or equal to 58. Let $\bar{Y}$ be the mean of the sample of 100 students who take the new version. By the central limit theorem, since our sample size is large (≥ 30), the sample mean is an approximately normal random variable with mean $\mu = 60$ and variance $\sigma^2/n = 64/100$. The standard deviation of the sample mean is the square root of the variance, i.e. $8/10 = 0.8$. Converting to the standard normal random variable:

$$\mathbb{P}(\bar{Y} \leq 58) = \mathbb{P}\left(Z \leq \frac{58-60}{0.8}\right) = \mathbb{P}(Z \leq -2.5) = 0.0062$$

This probability is so small that it is unlikely that it was drawn from a population with mean 60 and variance 64. Thus it is highly likely that it was drawn from a population with different characteristics, i.e. it is highly likely that there is something wrong with the test.

Example. Service times for customers in a retail store are independent random variables with mean 1.5 minutes and variance 1.0 minutes. Approximately what is the probability that 100 customers can be served in less than 2 hours?

Let Y_i be the service time for the ith customer. The distributions of the service times is unknown, but if we had to model them we might choose an exponential distribution (although we would have to change either the mean or the variance if we were to do this). Then we want:

$$\mathbb{P}\left(\sum_{i=1}^{100} Y_i \leq 120\right) = \mathbb{P}\left(\frac{1}{100}\sum_{i=1}^{100} Y_i \leq \frac{120}{100}\right) = \mathbb{P}(\bar{Y} \leq 1.20)$$

Since n is large (i.e. ≥ 30), by the central limit theorem, $\bar{Y}$ is approximately normally distributed with mean $\mu = 1.5$ and variance $\sigma^2/n = 1/100 = 0.01$. Thus, converting to the standard normal random variable, we have:

$$\begin{aligned}\mathbb{P}(\bar{Y} \leq 1.20) &\approx \mathbb{P}\left(Z \leq \frac{1.20 - 1.5}{\sqrt{0.01}}\right) \\ &= \mathbb{P}\left(Z \leq \frac{-0.30}{0.1}\right) \\ &= \mathbb{P}(Z \leq -3.0) = 0.0013\end{aligned}$$

This probability is so small that it is virtually impossible to serve 100 customers in less than 2 hours.

6 Estimation

6.1 Introduction

The purpose of statistics is to make inferences about populations based on data from a small sample of that population. Populations are characterized by probability distributions which can be described by numerical parameters. The population mean μ and population variance σ^2 are parameters common to all probability distributions. In addition, some populations can be described by other natural parameters. One example is the population of all registered voters in Rhode Island. If we are interested in the yes/no question "Are you going to vote for Gina Raimondo?", a natural parameter is p, the proportion of the population who plans on voting for Raimondo.

The setup is exactly the same as with sampling. Suppose we are studying a population whose distribution is characterized by a parameter of interest which we will denote θ (this could be the population mean, population variance, proportion of voters supporting Raimondo, or some other parameter). We will take n independent, identically distributed samples $Y_1, \ldots, Y_n$ from our population. An *estimator* is a function of our n samples which is designed to give us information about the population parameter. There are two types of estimators we will discuss:

1. A *point estimator* produces a single number which we think is close to the parameter of interest. We will learn several ways to quantitatively evaluate the "goodness" of a point estimator

2. An *interval estimate* produces an interval (often called a confidence interval) in which we believe our parameter of interest lies. We will learn how to construct confidence intervals which include the parameter of interest with a given probability.

6.2 Point Estimators

We start our discussion with point estimators. We have already met one point estimator, the sample mean $\bar{Y}$. The sample mean is an estimator since it is a function of our n samples. It is an estimator for the population mean. For another example, suppose we are polling n voters out of a population of registered voters and asking them if they are voting for Gina Raimondo. The parameter of interest is p, the proportion of registered voters who will vote for Raimondo. Let Y be the number of voters in our sample who are voting for Raimondo. Then the sample proportion $\hat{p} = Y/n$ is an estimator for the population proportion p. (The "hat" over the p indicates that it is an estimator for the parameter p).

Since we are applied mathematicians, we need quantitative tools to tell whether an estimator is any good. Let $\hat{\theta}$ be an estimator for the parameter θ. The first criterion we can use to evaluate an estimator is *bias*. We would like the expected value of our estimator to be the actual value of the parameter we are trying to estimate, i.e. $\mathbb{E}(\hat{\theta}) = \theta$. An estimator which has this property is called *unbiased*

Bias of an Estimator

Let $\hat{\theta}$ be an estimator for a parameter θ. The estimator $\hat{\theta}$ is *unbiased* if $\mathbb{E}(\hat{\theta}) = \theta$. Otherwise, the estimator $\hat{\theta}$ is *biased*. The *bias* of $\hat{\theta}$ is given by

$$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Let's look at the estimators we have seen so far. The sample mean $\bar{Y}$ is an estimator for the population mean μ (here we abandon our "hat"-convention, and call this $\bar{Y}$ rather than $\hat{\mu}$). We showed in the last section that $\mathbb{E}(\bar{Y}) = \mu$, thus the sample mean is an unbiased estimator for the population mean.

What about the sample proportion estimator we use in polling? Suppose we poll n voters in a population, and let Y be the number of voters in our sample who are voting for Gina Raimondo. As long as the sample is small enough (less than 1/20 of the population size), we can take Y to be a binomial random variable with parameters n and p. For our estimator $\hat{p} = Y/n$,

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{Y}{n}\right) = \frac{1}{n}\mathbb{E}(Y) = \frac{np}{n} = p$$

where we have used the expected value of a binomial random variable. Since $\mathbb{E}(\hat{p}) = p$, this estimator is unbiased as well. We will see an example of a biased estimator later.

A perhaps better measure of the "goodness" of an estimator is its *mean square error*, the average of the square distance of the estimator from the parameter of interest.

Mean Square Error

Let $\hat{\theta}$ be an estimator for a parameter θ. The *mean square error* of $\hat{\theta}$ is defined by

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

If $Bias(\hat{\theta})$ is the bias of $\hat{\theta}$ and $Var(\hat{\theta})$ is the variance of $\hat{\theta}$, then

$$MSE(\hat{\theta}) = [Bias(\hat{\theta})]^2 + Var(\hat{\theta})$$

Note that the MSE can be divided into two components: variance and bias squared. Both of these are positive. We have discussed above how low bias is a good quality for an estimator. Low variance is a desired quality for an estimator as well since ideally we would like the distribution of the estimator to cluster tightly about the parameter of interest. In general, if we keep the sample size fixed, there is a trade-off between bias and variance. For a given MSE, if we wish our estimator to have lower bias, then we must accept a higher variance

and vice versa.

To show the relationship between MSE, bias, and variance, we take the definition of MSE and add and subtract $\mathbb{E}(\hat{\theta})$ inside the parentheses.

$$\begin{aligned}
MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta})) + (\mathbb{E}(\hat{\theta}) - \theta)]^2 \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\underbrace{\mathbb{E}(\hat{\theta}) - \theta)}_{\text{constant}}] + \mathbb{E}[\underbrace{(\mathbb{E}(\hat{\theta}) - \theta)^2}_{\text{constant}}] \\
&= Var(\hat{\theta}) + 2(\mathbb{E}(\hat{\theta}) - \theta)\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))] + (\mathbb{E}(\hat{\theta}) - \theta)^2 \\
&= Var(\hat{\theta}) + 2(\mathbb{E}(\hat{\theta}) - \theta)(\underbrace{\mathbb{E}(\hat{\theta}) - \mathbb{E}(\hat{\theta})}_{=0}) + [Bias(\hat{\theta})]^2 \\
&= Var(\hat{\theta}) + [Bias(\hat{\theta})]^2
\end{aligned}$$

Let's look at the MSE of the two estimators we discussed above.

1. Sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

   We showed that this estimator is unbiased, so the MSE is equal to the variance. We showed in a previous section that the variance of the sample mean is σ^2/n, where σ^2 is the population variance. Thus we have

   $$MSE(\bar{Y}) = \frac{\sigma^2}{n}$$

   Note that the MSE goes to 0 as $n \to \infty$, i.e. the error of our estimator decreases as our sample gets larger. This makes intuitive sense that a larger sample provides a better estimator for the population mean.

2. Sample proportion. $\hat{p} = \frac{Y}{n}$.

   Recall that we are assuming that $Y \sim$ Binomial(n, p). We showed above that this estimator is also unbiased, thus once again the MSE is equal to the variance. Recalling that the variance of a binomial random variable is $np(1-p)$,

   $$\begin{aligned}
   MSE(\hat{p}) &= Var\left(\frac{Y}{n}\right) \\
   &= \frac{1}{n^2} Var(Y) \\
   &= \frac{np(1-p)}{n^2} \\
   &= \frac{p(1-p)}{n}
   \end{aligned}$$

Sometimes we are interested in studying the difference between two populations. Here are some examples of that.

1. Suppose we are interested in whether Brown University first-year students or seniors get more sleep. In this case, the parameter of interest is the *difference* in the mean amount of sleep between first-years and seniors. If μ_1 and σ_1^2 are the mean and variance of the amount of sleep of first-years and μ_2 and σ_2^2 are the same for seniors, then mathematically our parameter of interest is $\mu_1 - \mu_2$. If we take a sample of n_1 first-year students and n_2 seniors and compute the sample means $\bar{Y}_1$ and $\bar{Y}_2$, then $\bar{Y}_1 - \bar{Y}_2$ is an estimator for $\mu_1 - \mu_2$. By linearity of expectation and our result for the expected value of the sample mean, the expected value of the sample mean is $\mu_1 - \mu_2$, so this estimator is unbiased. For the variance of this estimator, we use the formula for the variance of a sum (assuming the samples are independent), and recall that constants are squared when they are pulled out of the variance:

$$\begin{aligned} Var(\bar{Y}_1 - \bar{Y}_2) &= Var(\bar{Y}_1) + Var(-\bar{Y}_2) \\ &= Var(\bar{Y}_1) + (-1)^2 Var(\bar{Y}_2) \\ &= Var(\bar{Y}_1) + Var(\bar{Y}_2) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

2. Suppose we are interested in the preference for Gina Raimondo in rural versus urban voters in Rhode Island. The parameter of interest here is the difference in the proportion of Raimondo supporters between rural and urban areas. First, we have to define "rural" and "urban". This is admittedly tricky in a small state like Rhode Island, but we could for example take "urban" to mean living in a city with population of 40,000 or more (in Rhode Island, this would include Woonsocket, East Providence, Pawtucket, Cranston, Warwick, and Providence). What are advantages or drawbacks to this definition? If p_1 is the proportion of Raimondo supports in rural areas and p_2 is the proportion of Raimondo supporters in urban areas, then the parameter of interest is $p_1 - p_2$. Suppose we sample n_1 voters from rural areas and n_2 voters from urban areas. Let Y_1 and Y_2 be the proportion of rural and urban voters (respectively) who support Raimondo. Then

$$\hat{p}_1 - \hat{p}_2 = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}$$

is an estimator for $p_1 - p_2$. By linearity of expectation and the result for a single population, the expected value of this estimator is $p_1 - p_2$, so this estimator is unbiased. What is the variance of this estimator? As above, using the formula for the variance of a sum of two independent random variables,

$$\begin{aligned} Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + Var(-\hat{p}_2) \\ &= Var(\hat{p}_1) + (-1)^2 Var(\hat{p}_2) \\ &= Var(\hat{p}_1) + Var(\hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \end{aligned}$$

We summarize these common estimators for population mean and proportion in the following table:

Parameter of Interest	Sample Size	Estimator	Expected Value	Variance	Standard Deviation
μ	n	$\bar{Y}$	μ	$\frac{\sigma^2}{n}$	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\frac{p(1-p)}{n}$	$\sqrt{\frac{p(1-p)}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\bar{Y}_1 - \bar{Y}_2$	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

The standard deviation of an estimator is sometimes called the *standard error*. You may see that term in the scientific literature, but we will not use it in class. Note that the expected values and variances in the table above hold regardless of the distribution of the underlying population. In the case that the population is normal, the sample mean estimators $\bar{Y}$ and $\bar{Y}_1 - \bar{Y}_2$ have a normal distribution, as discussed in the chapter on sampling distributions. However, for large sample sizes (approximately $n \geq 30$), the central limit theorem comes into play. Thus for large sample sizes, the sample mean estimators $\bar{Y}$ and $\bar{Y}_1 - \bar{Y}_2$ are approximately normally distributed regardless of the distribution of the underlying population. For a binomial population, the population proportion estimators $\hat{p}$ and $\hat{p}_1 - \hat{p}_2$ are also approximately normal for large sample sizes. How large a sample size do we need in this case? It depends on p. The farther p is from $1/2$, the larger sample size n we need for the binomial distribution to be approximately normal. In a homework problem, we showed that this is the case when $0 \leq 3\sqrt{pq/n} \leq 1$, where $q = 1 - p$. We then showed that an easier rule to check is that the binomial distribution is approximately normal when $n \geq 9p/q$ and $n \geq 9q/p$.

As another example, let's look at our estimators for population variance. Recall from the previous section that the sample variance is given by:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

where $\bar{Y}$ is the sample mean. The factor of $(n-1)$ in the denominator seems peculiar. It seems more natural to divide by n and use the following estimator for the population variance:

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

We will show that while they are both estimators for the sample variance, S'^2 is biased whereas S^2 is unbiased. Since S^2 is unbiased, we call it the sample variance.

First we show S'^2 is biased by computing its expected value. This is done in several steps.

1. First we find a nice formula for $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2) \\ &= \sum_{i=1}^{n} Y_i^2 - 2\bar{Y}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n}\bar{Y}^2 \\ &= \sum_{i=1}^{n} Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2 \\ &= \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\end{aligned}$$

2. Next we take the expected value of this. By linearity of expectation,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] &= \mathbb{E}\left(\sum_{i=1}^{n} Y_i^2\right) - n\mathbb{E}(\bar{Y}^2) \\ &= \sum_{i=1}^{n}\mathbb{E}(Y_i^2) - n\mathbb{E}(\bar{Y}^2)\end{aligned}$$

3. Next we use the Magic Variance Formula "in reverse" to compute $\mathbb{E}(Y_i^2)$ and $\mathbb{E}(\bar{Y}^2)$.

$$\begin{aligned}Var(Y_i) &= \mathbb{E}(Y_i^2) - [\mathbb{E}(Y_i)]^2 \\ \mathbb{E}(Y_i^2) &= Var(Y_i) + [\mathbb{E}(Y_i)]^2 \\ &= \sigma^2 + \mu^2\end{aligned}$$

   Similarly,

$$\begin{aligned}\mathbb{E}(\bar{Y}^2) &= Var(\bar{Y}) + [\mathbb{E}(\bar{Y})]^2 \\ &= \frac{\sigma^2}{n} + \mu^2\end{aligned}$$

4. We then plug these into the expression from step 3.

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] &= \sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\ &= (n-1)\sigma^2\end{aligned}$$

5. Finally we use this result to compute the expected value of S'^2.

$$\begin{aligned}\mathbb{E}(S'^2) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] \\ &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] \\ &= \frac{1}{n}(n-1)\sigma^2 \\ &= \frac{n-1}{n}\sigma^2\end{aligned}$$

Since $\mathbb{E}(S'^2) \neq \sigma^2$, this estimator is biased. For large n, however, $(n-1)/n \approx 1$, so the bias is minimal. We can convert this to an unbiased estimator by multiplying by $n/(n-1)$. This is legitimate since n is a known constant (the sample size) and not one of the parameters we are trying to estimate:

$$\begin{aligned}\mathbb{E}\left(\frac{n}{n-1}S'^2\right) &= \frac{n}{n-1}\mathbb{E}(S'^2) \\ &= \frac{n}{n-1}\frac{n-1}{n}\sigma^2 \\ &= \sigma^2\end{aligned}$$

So we have found an unbiased estimator for σ^2. But we also have:

$$\begin{aligned}\frac{n}{n-1}S'^2 &= \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= S^2\end{aligned}$$

Thus S^2 is an unbiased estimator for the population variance σ^2.

6.3 Interval Estimators

Once again, suppose we are studying a population whose distribution is characterized by a parameter of interest which we will denote θ. We will take n independent, identically distributed samples $Y_1, \ldots, Y_n$ from our population. An *interval estimator* or *confidence interval* uses the samples $Y_1, \ldots, Y_n$ to construct a confidence interval $[\hat{\theta}_L, \hat{\theta}_U]$. (The subscripts L and U denote the lower and upper endpoints of the interval, and the "hat" on the θ reminds us that this is an estimator). Note that the confidence interval is a random variable, and is a function of our n sample points. If we take a different sample, we will get a different confidence interval.

We would like the confidence interval to have the following properties:

1. It should contain the parameter of interest θ.
2. It should be relatively narrow (otherwise we haven't learned much).
3. We should be able to calculate the probability that our confidence interval will enclose our parameter of interest. This probability is called the *confidence coefficient*.

Suppose we have constructed a confidence interval $[\hat{\theta}_L, \hat{\theta}_U]$ for our parameter θ from our n samples $Y_1, \ldots, Y_n$. Then the confidence coefficient is denoted $(1-\alpha)$, and so we have[25]:

$$\mathbb{P}(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

The confidence interval $[\hat{\theta}_L, \hat{\theta}_U]$ is sometimes called a *two-sided confidence interval*. We can also construct one-sided confidence intervals, although we will not do so in this course.

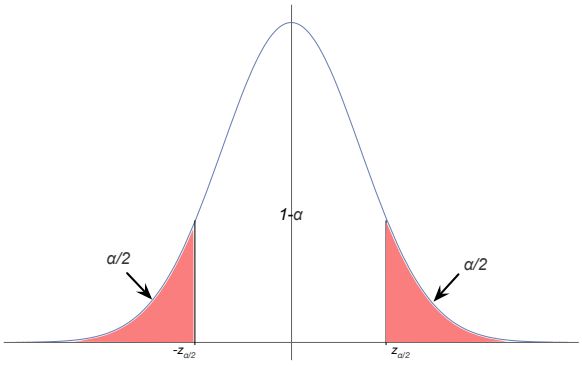
6.3.1 Confidence Intervals for Large Sample Sizes

Earlier we discussed the common unbiased estimators $\bar{Y}$ (sample mean) and $\hat{p}$ (sample proportion), as well as the equivalent estimators for the difference of two populations. For large sample sizes ($n \geq 30$ for the sample mean, and using our binomial rule for the sample proportion), the central limit theorem tells us that these estimators are all normally distributed with mean and standard deviation given in the table above. Thus, if we convert to the standard normal random variable, we can construct a confidence interval with desired confidence coefficient $(1-\alpha)$ using the Z distribution. Let's see how we can do this.

Let θ be our parameter of interest (either $\mu, p, \mu_1 - \mu_2$, or $p_1 - p_2$), and let $\hat{\theta}$ be the appropriate unbiased estimator from the table above. In all cases, since the estimator is unbiased, the expected value $\mathbb{E}(\hat{\theta}) = \theta$. Let $\sigma_{\hat{\theta}}$ be the standard deviation of $\hat{\theta}$ (which we can look up in the table above). Then:

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

is (approximately) a standard normal random variable. Suppose we want a confidence interval for θ which has a confidence coefficient $(1-\alpha)$. We will always use a two-sided, symmetric confidence interval (although you do not have to do this).



[25]The reasons for the notation $(1-\alpha)$ will be more evident when we discuss hypothesis testing; roughly speaking, α is the probability of a false positive result.

Converting to the standard normal random variable, this is equivalent to finding values $-z_{\alpha/2}$ and $z_{\alpha/2}$ such that $\mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = (1-\alpha)$. This is illustrated in the picture above. To find $-z_{\alpha/2}$ and $z_{\alpha/2}$, we can use the Z table. For example, if $(1-\alpha) = 0.95$, then $\alpha/2 = 0.025$. Consulting the Z table, we find that $-z_{\alpha/2} = -1.96$. By symmetry of the standard normal distribution about its mean of 0, $z_{\alpha/2} = 1.96$.

To find our confidence interval, we convert from the standard normal distribution back to the distribution of our estimator, which is Normal$(\theta, \sigma_{\hat{\theta}})$. Substituting for Z, we have

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Now we just manipulate this to get it into the form we want.

$$\begin{aligned}
\mathbb{P}\left(-z_{\alpha/2}\sigma_{\hat{\theta}} \leq \hat{\theta} - \theta \leq z_{\alpha/2}\sigma_{\hat{\theta}}\right) &= 1 - \alpha \\
\mathbb{P}\left(-\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq -\theta \leq -\hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right) &= 1 - \alpha \\
\mathbb{P}\left(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right) &= 1 - \alpha
\end{aligned}$$

where to get the last line we multiplied by -1 and flipped all the inequalities. Thus the $(1-\alpha)$ confidence interval for θ is:

$$[\hat{\theta}_L, \hat{\theta}_U] = [\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}]$$

We can also write this in terms the population standard deviation σ and the sample size n by substituting $\sigma_{\hat{\theta}} = \sigma/\sqrt{n}$.

$$[\hat{\theta}_L, \hat{\theta}_U] = \left[\hat{\theta} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\theta} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Let's do some examples.

Example. A ball bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation $\sigma = 0.1$ mm. We would like to estimate the mean μ using a 90% confidence interval. To do this, we take a sample size of $n = 9$ ball bearings and use the sample mean $\bar{Y}$ as our estimator for μ. Suppose we measure $\bar{Y} = 10.0$ mm. What is the 90% confidence interval for μ?

Although this is not a large sample size, we can still find a confidence interval in this case because we are assuming that the population is normally distributed. Thus the sample mean $\bar{Y}$ will always be normally distributed. Since we know the population standard deviation, we can convert to the standard normal Z distribution directly without appealing to approximations which require a large sample size. First compute the standard deviation of our estimator:

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{9}} = \frac{0.1}{3} = 0.0333$$

Since $1-\alpha = 0.9$, $\alpha = 0.1$, thus $\alpha/2 = 0.05$. Looking at the Z table, we have to choose between a probability of 0.0495 ($z_{\alpha/2} = 1.65$) and a probability of 0.0505 ($z_{\alpha/2} = 1.64$). We will choose $z_{\alpha/2} = 1.65$ to allow for more "wiggle room". Thus our 90% confidence interval for μ is given by:

$$\begin{aligned}[\bar{Y}_L, \bar{Y}_U] &= [\bar{Y} - z_{\alpha/2}\sigma_{\bar{Y}}, \bar{Y} + z_{\alpha/2}\sigma_{\bar{Y}}] \\ &= [\bar{Y} - (1.65)(0.0333), \bar{Y} + (1.65)(0.0333)] \\ &= [10.0 - (1.65)(0.0333), 10.0 + (1.65)(0.0333)] \\ &= [9.945, 10.055]\end{aligned}$$

Although we cannot say for certain that the true population parameter μ falls within this range, we can say that the probability that it does is 0.90. If we repeated this procedure, i.e. took 9 more samples, generated the sample mean, and constructed a 90% confidence interval, we would get a different confidence interval, but the probability that μ would fall in that confidence interval would still be 0.90.

Note that constructing this confidence interval required knowledge of the population standard deviation (which we then divided by n to get the standard deviation of the estimator). Often we do not know this parameter, but we wish to construct a confidence interval anyway. What to we do in that case? We use the sample standard deviation (found, for example, using the estimator S) in place of the population standard deviation. If this sounds to you like cheating, you are right! The key is that n is large. If the population is large, the sample standard deviation is close to the population standard deviation and there is very little loss of accuracy if we use S in place of σ in the formula for the confidence interval. We will justify this approximation mathematically later on, but for now just recall that the t distribution (which we use when the population standard deviation is unknown) approaches the standard normal Z distribution as the number of degrees of freedom increases.

Example. The shopping times of $n = 64$ randomly selected customers at a local supermarket were recorded. The sample mean and sample variance of the 64 shoppers were 33 minutes and 256 minutes, respectively. Find a 98% confidence interval for μ, the true average shopping time per customer.

In this case, the parameter of interest is μ. Our estimator is $\bar{Y}$, the sample mean. In this experiment, we sampled $n = 64$ customers and found a sample mean $\bar{Y} = 33$ and a sample variance $S^2 = 256$. Since the sample is large (more than 30 customers), by the central limit theorem, we can assume that $\bar{Y}$ is normally distributed. We do not know the population variance, so we will use the sample variance in place of the population variance. Since the population is large, this is a reasonable assumption. Thus we can use the formula above for the confidence interval, where for the standard deviation of the estimator we use:

$$\sigma_{\bar{Y}} = \frac{S}{\sqrt{n}} = \frac{\sqrt{256}}{\sqrt{64}} = \frac{16}{8} = 2$$

To find $z_{\alpha/2}$, we look at our Z table. Since $(1-\alpha) = 0.98$, $\alpha = 0.02$, thus $\alpha/2 = 0.01$. Looking at our Z table, the closest probability we have to 0.01 is 0.0099 (0.0102 is the next

closest, but since it is farther away, we will use 0.099), and the value of Z corresponding to that gives us $z_{\alpha/2} = 2.33$. We now have everything we need to construct our 98% confidence interval.

$$\begin{aligned}[\bar{Y}_L, \bar{Y}_U] &= [\bar{Y} - z_{\alpha/2}\sigma_{\bar{Y}}, \bar{Y} + z_{\alpha/2}\sigma_{\bar{Y}}] \\ &= [33 - (2.33)(2), \bar{Y} + (2.33)(2)] \\ &= [28.34, 37.66]\end{aligned}$$

Here is another example, this time dealing with the difference between sample proportions.

Example. Two brands of light bulbs, denoted brand A and brand B, are guaranteed to last for at least 1 year. In a random sample of 50 light bulbs of brand A, 12 were found to fail before the 1 year period ended. In an independent random sample of 60 light bulbs of brand B, 12 were also found to fail before the 1 year period ended. Give a 98% confidence interval for the difference $p_1 - p_2$ between the proportion of failures of the two brands during the 1 year period.

The parameter of interest here is $p_1 - p_2$, and we use as our estimator $\hat{p}_1 - \hat{p}_2$. Evaluating our estimators, we get $\hat{p}_1 = 12/50 = 0.24$ and $\hat{p}_2 = 12/16 = 0.20$, thus $\hat{p}_1 - \hat{p}_2 = 0.24 - 0.20 = 0.04$. To construct our confidence interval, we need the standard deviation of our estimator. From the table above (or by deriving it from the variance of the binomial distribution), we have

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

But we have a problem. The standard deviation involves the true parameter p, and that is what we are trying to estimate! No worries, we will just do what we did before. Since the population is large, we will use the estimator $\hat{p}$ in place of p in the expression for the standard deviation.

$$\begin{aligned}\sigma_{\hat{p}_1 - \hat{p}_2} &\approx \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= \sqrt{\frac{(0.24)(0.76)}{50} + \frac{(0.20)(0.80)}{60}} \\ &= 0.0795\end{aligned}$$

We found $z_{\alpha/2}$ for a 98% confidence interval in the previous example, thus we have $z_{\alpha/2} = 2.33$. Using the formula, our 98% confidence interval is:

$$\begin{aligned}[(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sigma_{\hat{p}_1 - \hat{p}_2}, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sigma_{\hat{p}_1 - \hat{p}_2}] &= [0.04 - (2.33)(0.0795), 0.04 + (2.33)(0.0795)] \\ &= [0.04 - 0.185, 0.04 + 0.185] \\ &= [-0.145, 0.225]\end{aligned}$$

Note that our 98% confidence overlaps 0. If the true parameter $p_1 - p_2$ were 0, then there would be no difference in the 1-year failure rates of the two brands of light bulbs! Thus we cannot exclude that possibility with 98% confidence.

6.3.2 Experimental Design: Selecting the Sample Size

If you are designing an experiment, one of the parameters you must choose is the sample size. As an example, suppose you are polling registered voters in Rhode Island and asking them if they are voting for Gina Raimondo. You use the estimator $\hat{p}$ for p, the true proportion of Raimondo voters. As the number n of people polled (the size of your sample) increases, the variance of your estimator $\hat{p}$ decreases, and there is a higher probability that your estimator will be close to the true proportion. In terms of confidence intervals, if you desire 95% confidence that your interval contains the true parameter p, as you increase the sample size, the confidence interval becomes narrower. There is a trade-off, however, as the sample size gets larger. Collecting a larger sample is more expensive and consumes more resources. Is there any way mathematically to determine ahead of time the sample size we need to attain a desired level of accuracy for our estimator?

Suppose we want to estimate the average diameter μ of a ball bearing produced by a ball bearing machine, and we wish the probability that our estimate is within 0.02 mm of the true mean to be 0.95. Assuming the diameters of the ball bearings are normally distributed (or taking a large enough sample that the central limit theorem applies), by the 68-95-99.7 rule we want the true value μ to be within 2 standard deviations of our estimator $\bar{Y}$ (This is the standard deviation of the *estimator*, not the population standard deviation). Thus we want:

$$2\sigma_{\bar{Y}} = 2\frac{\sigma}{\sqrt{n}} = 0.02$$

We can solve the above equation for n as long as we know σ.

$$n = \left(\frac{2\sigma}{0.02}\right)^2$$

What do we do if the population standard deviation σ is not know, as is often the case. One possibility is to use the sample standard deviation S from a previously taken sample in place of the population standard deviation σ. As long as we take a large enough sample, this is a reasonable estimate. Another alternative is to use the range (largest value minus smallest value) of a previous sample. If we assume that the population is roughly normally distributed, then we know from the 68-95-99.7 rule that 95% of the population will lie within two standard deviations of the mean. If our sample is large enough, then the sample range should be approximately 4σ, since we are assuming the range represents two standard deviations both above and below the mean. Thus we can approximate σ as one-fourth of the range.

In our ball bearing example, suppose in a previous sample of 50 ball bearings we had a range of 0.4 mm. Dividing this by 4, we can estimate $\sigma = 0.4/4 = 0.1$. Plugging this info our formula above, we get:

$$n = \left(\frac{2 \cdot 0.1}{0.02}\right)^2 = 10^2 = 100$$

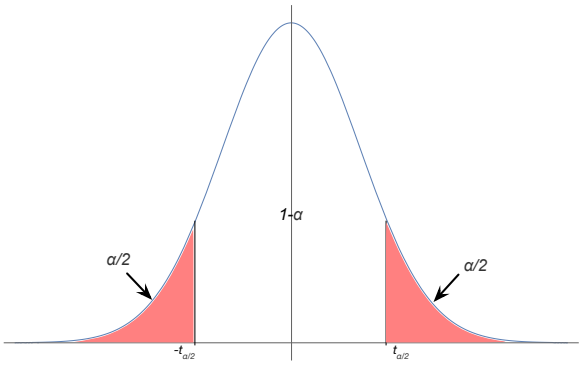
6.4 Small Sample Confidence Intervals for Population Mean

In this section we will discuss confidence intervals for the estimators of population means μ and $\mu_1 - \mu_2$. The confidence interval estimators in the previous section relied on the fact that the sample size was large. We used the large sample size assumption in two places. First, if the underlying sample size is not normal, we need a large sample size to apply the central limit theorem and conclude that the sample mean estimator $\bar{Y}$ is approximately normal. Second, the confidence interval formula relies on the standard deviation of the estimator, and the formula for that standard deviation involves the population standard deviation. This population standard deviation is usually unknown, so we estimate it with the sample standard deviation S. This approximation is only valid for a large sample size. What do we do when the sample size is small?

We will assume in this section that the population is normally distributed, since the sample size will be too small for the central limit theorem to apply. Let $Y_1, \ldots, Y_n$ be n samples from a normal population, and let $\bar{Y}$ and S^2 be the sample mean and variance. We would like to construct a confidence interval for the population mean μ in the case where the population variance σ^2 is unknown and the sample is too small to approximate σ^2 with S^2. Recall from the chapter on sampling distributions that

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$$

has a t-distribution with $(n-1)$ degrees of freedom. We will construct a confidence interval in the same method as for large sample sizes, but we will use the t-distribution in place of the standard normal distribution.



As with the confidence interval involving the Z distribution, given a confidence coefficient $(1-\alpha)$, we will look for a two-sided confidence interval by dividing α by 2 and looking for values $-t_{\alpha/2}$ and $t_{\alpha/2}$ such that

$$\mathbb{P}(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$$

The value $t_{\alpha/2}$ can be found from the t-table with $(n-1)$ df. The confidence interval formula is derived in exactly the same was as the large sample case, and the only differences between

the formulas is that the population standard deviation σ is replaced by the sample standard deviation S and the Z value is replaced by a t value.

$$[\bar{Y}_L, \bar{Y}_U] = \left[\bar{Y} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{Y} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

Since the t distribution has thicker tails than the normal distribution, confidence intervals involving small sample sizes which use the t distribution will be wider than those using the Z distribution.

Example. You are a rocket scientist, and you conduct an experiment which involves measuring the launch velocity of a model rocket. Suppose 8 measurements are taken. The sample mean is 29.59 m/s, and the sample standard deviation is 0.391 m/s. Find a 95% confidence interval for the true launch velocity of the model rocket.

We will assume that the launch velocities are normally distributed. Since we have a small sample size, we have to use the t distribution to construct our confidence interval. Since there are 8 observations, we want the t distribution with 7 df. Since $(1-\alpha) = 0.95$, $\alpha/2 = 0.025$. From the t-table, we find that $t_{\alpha/2} = t_{0.025} = 2.365$. Thus our 95% confidence interval is

$$\begin{aligned}[\bar{Y}_L, \bar{Y}_U] &= \left[29.59 - (2.365)\frac{0.391}{\sqrt{8}}, 25.959 + (2.365)\frac{0.391}{\sqrt{8}}\right] \\ &= [29.59 - 0.327, 29.59 + 0.327]\end{aligned}$$

We can similarly handle the case where we want to compare the means of two normal populations. If the sample sizes are large and we know the standard deviations of both populations, we can use the confidence interval methods described in the previous section. If the standard deviations are not know, we will again rely on the t distribution. In this case, the algebra is a lot more messy. *I think the result is interesting and useful, so I am presenting it here. It is complicated enough that I will not put it on an exam.*

Suppose we want to compare the means μ_1 and μ_2 of two populations. We will assume that they have the *same* variance σ_2, but that this variance is unknown. (This assumption is important for this model.) Take a sample of size n_1 from the first population and a sample of size n_2 from the second population. Compute the sample means $\bar{Y}_1$ and $\bar{Y}_2$, and construct the estimator $\bar{Y}_1 - \bar{Y}_2$. We would like to construct a confidence interval for this estimator. First, we compute the sample variance S_1^2 for the first sample and the sample variance S_2^2 for the second sample. An unbiased estimator for the population variance σ^2 can be obtained by pooling the sample data from both samples to obtain the *pooled variance estimator* S_p^2:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

This estimator is a weighted average of S_1^2 and S_2^2, with the larger sample being given a higher weight. The weights $(n_1 - 1)$ and $(n_2 - 1)$ are used in place of n_1 and n_2 so that this

is an unbiased estimator (similar to the factor of $n-1$ in the denominator for the sample variance estimator S^2). We can show that the quantity

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t distribution with $n_1 + n_2 - 2$ degrees of freedom (df). Thus, in a similar fashion to the confidence intervals we constructed earlier, the confidence interval for $(\mu_1 - \mu_2)$ is given by:

$$\left[(\bar{Y}_1 - \bar{Y}_2) - t_{\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{Y}_1 - \bar{Y}_2) + t_{\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

6.5 Consistency

We talked in previous sections about desirable properties for estimators. Ideally, we would like our estimators to be unbiased. However, we showed in a homework problem that in certain circumstances, a biased estimator may in fact have lower mean square error (MSE) than an unbiased estimator, so bias is not the entire story. Mean square error, which is the sum of bias squared and variance, is a good measure of the accuracy of an estimator, but it does not take into account the size of the sample we are taking. Ideally, we would like to define an estimator as "good" if the probability of "missing" the true parameter value goes to 0 as the sample size gets large. This concept of "goodness" of an estimator is called *consistency*.

We use the same setup as always. Consider a population which can be characterized by a parameter θ. Take n samples $Y_1, \ldots, Y_n$ from the population, and construct from these an estimator $\hat{\theta}_n$ for θ. Note the subscript n indicates that we have a different estimator for each value of n. For example, if we want to estimate the population mean μ, then we can use the sample mean as an estimator. Writing the sample mean with a subscript n to denote the number of samples we are taking from our population, we have our estimator for μ:

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

An estimator $\hat{\theta}_n$ is *consistent* for θ if the probability that $\hat{\theta}_n$ misses the true parameter by even a small amount approaches 0 as n approaches infinity. Mathematically we write this as follows.

Consistency Estimator

Let $\hat{\theta}_n$ be an estimator for a parameter θ, where $\hat{\theta}_n$ is obtained from a sample of size n from the underlying population. The estimator $\hat{\theta}_n$ is *consistent* for θ if for all $\epsilon > 0$ (no matter how small),

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta| \geq \epsilon) = 0$$

This type of convergence is called *convergence in probability*.

Like many criteria in mathematics, this is difficult to check. In the case of an unbiased estimator, we have an equivalent characterization of consistency which is much easier to verify. This is one of the reasons we like unbiased estimators, even though we have seen that unbiased estimators are not always "better".

Consistency of an Unbiased Estimator

Let $\hat{\theta}_n$ be an *unbiased* estimator for a parameter θ, where $\hat{\theta}_n$ is obtained from a sample of size n from the underlying population. Then $\hat{\theta}_n$ is consistent for θ if

$$\lim_{n\to\infty} Var(\hat{\theta}_n) = 0$$

To see that this is true, we can use Chebyshev's Inequality. For $\epsilon > 0$, Chebyshev's Inequality gives us

$$\mathbb{P}(|\hat{\theta}_n - \mathbb{E}(\theta)| \geq \epsilon) \leq \frac{Var(\hat{\theta}_n)}{\epsilon^2}$$

Since $\hat{\theta}_n$ is unbiased (this is critical here), we know $\mathbb{E}(\theta) = \theta$, so we can substitute θ for $\mathbb{E}[\theta)$ above to get:

$$\mathbb{P}(|\hat{\theta}_n - \theta| \geq \epsilon) \leq \frac{Var(\hat{\theta}_n)}{\epsilon^2}$$

Taking the limit of both sides,

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta| \geq \epsilon) \leq \frac{1}{\epsilon^2} \lim_{n\to\infty} Var(\hat{\theta}_n) = 0$$

since we are assuming that the variance of our estimator goes to 0 as n goes to infinity.

We can apply this to see that the sample mean $\bar{Y}$ is a consistent estimator for the population mean μ. Recall that the sample mean is given by:

$$\bar{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

We have shown that this is unbiased, i.e. $\mathbb{E}\bar{Y_n} = \mu$. We also know from the section on sampling distributions that $Var(\bar{Y_n}) = \sigma^2/n$, which goes to 0 as n goes to infinity. Thus $\bar{Y_n}$ is a consistent estimator for μ. This result is known as the *Weak Law of Large Numbers*, and is one of the fundamental results of probability. Recall from the beginning of the course that when we defined expected value, we said that we could think of it intuitively as taking the average of a large number of experiments. In other words, the expected value is approximately what we get if we repeat an experiment n times, add up the results, and divide by n. The law of large numbers makes this intuition mathematically precise. $\bar{Y_n}$ is the empirical mean, i.e. what we get when we "add-them-up-and-divide-by-n". As n gets large, $\bar{Y_n}$ approaches the true expected value μ in the sense that the probability that $\bar{Y_n}$ "misses"

μ becomes arbitrarily small.

We can similarly show that our unbiased estimator for variance S^2 (which we can write S_n^2 to indicate the sample size of n) is a consistent estimator for the population variance.

6.6 Construction of Estimators

So far we have defined what an estimator is and discussed qualities we would like to have in our estimators (unbiased, low MSE, consistent). But we have only discussed three estimators: $\bar{Y}$, S^2, and $\hat{p}$. There are many more parameters we might like to estimate. How can we construct estimators for them? In this section we look at two methods of constructing estimators: the method of moments and the maximum likelihood estimator (MLE).

6.6.1 Method of Moments

This method is based on the fact that the sample mean $\bar{Y}$ is a consistent estimator for the population mean μ. A population parameter is often a function of the population mean. For example, if we have a population which has a Poisson distribution, then the Poisson parameter λ is equal to the population mean μ. Thus we can estimate λ by the sample mean $\bar{Y}$, i.e. $\hat{\lambda} = \bar{Y}$ is an estimator for the Poisson parameter λ.

In its most simple form, the method of moments works as follows:

1. Write the parameter of interest in terms of the population mean, i.e. solve for the parameter in terms of the population mean μ.
2. Substitute $\bar{Y}$ for μ to get the method of moments estimator.

Let's do some examples.

Example. Suppose we have a population which has a uniform distribution on the interval $[0, b]$. Find the method of moments estimator for b.

The population mean μ is the expected value of a uniform random variable, so $\mu = \frac{b-0}{2} = \frac{b}{2}$. Solving for b we get $b = 2\mu$. Substituting $\bar{Y}$ for μ we get $b = 2\bar{Y}$. Thus our method of moments estimator for b is:

$$\hat{b} = 2\bar{Y}$$

Since $\mathbb{E}(\bar{Y}) = \mu$, $\hat{b}$ is an unbiased estimator for b. What is its variance?

$$\begin{aligned} Var(\hat{b}) &= Var(2\bar{Y}) \\ &= 4Var(\bar{Y}) \\ &= 4\frac{\sigma^2}{n} \\ &= 4\frac{b^2}{12n} \\ &= \frac{b^2}{3n} \end{aligned}$$

where we used both the variance of $\bar{Y}$ and the variance of the uniform distribution on $[0, b]$. Since there is a factor of n in the denominator, the variance of $\hat{b}$ goes to 0 as n goes to infinity, thus the method of moments estimator is a consistent estimator for b.

Example. Suppose we have a population which has a geometric distribution with parameter p. Find the method of moments estimator for p.

The population mean μ is the expected value of a geometric random variable, so $\mu = 1/p$. Solving for p, we get $p = 1/\mu$. For the method of moments, we substitute $\bar{Y}$ for μ, thus the method of moments estimator is:

$$\hat{p} = \frac{1}{\bar{Y}}$$

Example. Suppose we have a population which has a Poisson distribution with parameter λ. Find the method of moments estimator for λ.

Since the population mean is $\mu = \lambda$, the method of moments estimator is $\hat{\lambda} = \bar{Y}$.

Sometimes we cannot construct an method of moments estimator in this way, i.e. we cannot solve for the parameter in terms of the population mean μ. For example, suppose the population is uniformly distributed on $[a, b]$, and we want to estimate both a and b. The population mean is $\mu = \frac{a+b}{2}$, but we cannot solve for either a or b in terms of μ without having an expression in terms of the other parameter. The idea here is to estimate the variance using the method of moments, and then we have two equations for the two unknowns a and b. Since the algebra gets messy really quickly, we will not be pursuing this any further. All method of moments estimators we will use will only involve the sample mean $\bar{Y}$.

6.6.2 Maximum Likelihood Estimator (MLE)

The method of moments is very intuitive but often does not lead to the best estimators. A second way of constructing estimators is called the maximum likelihood estimator (MLE). Let's look an an example to illustrate how the MLE works.

Example. You have a bag which contains three marbles. You know they are either red or white, but you do not know how many of each are in the bag. For some reason, you are forbidden from opening the bag and looking at the marbles. However, you are permitted to sample two marbles without replacement. Suppose we draw out two red balls. What is a good estimate of the number of red balls in the bag?

Since we drew two red balls, the bag either contains two red balls or three red balls. If the bag contains two red balls, then the probability of our draw is:

$$\frac{\binom{2}{2}\binom{1}{0}}{\binom{3}{2}} = \frac{1}{3}$$

If the bag contains three red balls, then the probability of our draw is:

$$\frac{\binom{3}{2}}{\binom{3}{2}} = 1$$

A reasonable estimate for the number of red balls in the bag is 3, since that maximizes the probability of our draw.

The method illustrated in the example above is known as the method of *maximum likelihood*, since we choose the parameter which maximizes the probability of attaining our sample. Before we give the formal definition and more examples, we need to formally define the concept of likelihood.

Suppose we have a population whose distribution can be characterized by a parameter θ. The pmf or density function for the population will depend on θ, so the parameter θ will appear in the formula for the pmf or density function. We will indicate this dependence on θ as a subscript θ. Let $p_\theta(y)$ or $f_\theta(y)$ be the pmf or density function for our population. Now take a sample of size n, denoted $Y_1, Y_2, \ldots, Y_n$, from our population. Then we define the *likelihood* of our sample as follows.

Likelihood of a Sample

Let $Y_1, Y_2, \ldots, Y_n$ be a sample drawn from a population with pmf or density $p_\theta(y)$ or $f_\theta(y)$. Then the *likelihood* of the sample is given by:

$$L(Y_1, Y_2, \ldots, Y_n|\theta) = p_\theta(Y_1)p_\theta(Y_2)\cdots p_\theta(Y_n) \qquad \text{discrete distribution}$$
$$L(Y_1, Y_2, \ldots, Y_n|\theta) = f_\theta(Y_1)f_\theta(Y_2)\cdots f_\theta(Y_n) \qquad \text{continuous distribution}$$

In other words, we plug the samples into the pmf or density and multiply them together.

The maximum likelihood estimator (MLE) chooses the value of the parameter θ which maximizes the likelihood of our sample.

Maximum Likelihood Estimator

Let $Y_1, Y_2, \ldots, Y_n$ be a sample drawn from a population parameterized by θ, and let $L(Y_1, Y_2, \ldots, Y_n|\theta)$ be the likelihood of the sample. Then the *maximum likelihood estimator* (MLE) for θ, which is sometimes written $\hat{\theta}_{MLE}$, is the value of θ which maximizes the likelihood $L(Y_1, Y_2, \ldots, Y_n|\theta)$.

Let's do some examples of this. We will redo the three examples from the method of moments section.

Example. Suppose we have a population which has a uniform distribution on the interval $[0, b]$. Take a sample $Y_1, \ldots, Y_n$ from this distribution. Find the MLE for b.

First we need to find the likelihood function $L(Y_1, \ldots, Y_n|b)$. The density function for a uniform distribution on $[0, b]$ is

$$f_\theta(y) = \begin{cases} \frac{1}{b} & 0 \le y \le b \\ 0 & \text{otherwise} \end{cases}$$

Thus the likelihood function is

$$\begin{aligned} L(Y_1, Y_2, \ldots, Y_n|\theta) &= f_\theta(Y_1) f_\theta(Y_2) \cdots f_\theta(Y_n) \\ &= \begin{cases} \frac{1}{b^n} & \text{if } 0 \leq Y_i \leq b \text{ for all samples } Y_i \\ 0 & \text{otherwise, i.e. if for any sample } Y_i > b \text{ or } Y_i < 0 \end{cases} \end{aligned}$$

The MLE is the value of b which maximizes this likelihood function. We do not want the likelihood to be 0, since that cannot be a maximum, thus we need to have $0 \leq Y_i \leq b$ for all samples Y_i. Since b is in the denominator in our expression above, to maximize the likelihood we want to make b as small as possible without causing the likelihood to be 0. In other words, we want the *narrowest* interval for the uniform distribution that we can get away with, i.e. the smallest value for b for which the interval can contain our sample points. How do we do this? If we choose b to be the largest of our sample points, we have maximized our likelihood! In other words, our MLE estimator is the largest order statistic of our data:

$$\hat{b}_{MLE} = \max(Y_1, Y_2, \ldots, Y_n) = Y_{(n)}$$

This is very different from our method of moments estimator above. Is this estimator unbiased? The expected value of this estimator is a little tricky to compute, but we can do it if we think in terms of CDFs. Let $Y \sim \text{Uniform}[0, b]$, and let $F(y)$ be the CDF of Y, i.e. $F(y) = \mathbb{P}(Y \leq y)$. Let $Y_{(n)} = \max(Y_1, Y_2, \ldots, Y_n)$, and let $F_{(n)}$ be the CDF for $Y_{(n)}$. Then

$$\begin{aligned} F_{(n)}(y) &= \mathbb{P}(Y_{(n)} \leq y) \\ &= \mathbb{P}(\max(Y_1, Y_2, \ldots, Y_n) \leq y) \\ &= \mathbb{P}(Y_i \leq y \text{ for all } i = 1, \ldots, n) \\ &= F(Y_1) F(Y_2) \ldots F(Y_n) \end{aligned}$$

We can get the CDF of Y by integrating the uniform density:

$$F(y) = \begin{cases} 0 & y < 0 \\ \frac{y}{b} & 0 \leq y \leq b \\ 1 & y > 1 \end{cases}$$

Thus as long as all the samples are in the interval $[0, b]$ we have

$$F_{(n)}(y) = \left(\frac{y}{b}\right)^n$$

To get the density $f_{(n)}$ of $Y_{(n)}$ we take the derivative of the CDF with respect to y.

$$\begin{aligned} f_{(n)}(y) &= \frac{d}{dy} F_{(n)}(y) \\ &= n y^{n-1} \frac{1}{b^n} \end{aligned}$$

The density is 0 outside the interval $[0, b]$, so with the appropriate limits the density becomes

$$f_{(n)}(y) = \begin{cases} n y^{n-1} \frac{1}{b^n} & 0 \leq y \leq b \\ 0 & \text{otherwise} \end{cases}$$

So to get the expected value of $Y_{(n)}$ we use the formula for the expected value of a continuous random variable.

$$\begin{aligned}
\mathbb{E}(Y_{(n)}) &= \int_0^b y f_{(n)}(y) dy \\
&= \int_0^y y n y^{n-1} \frac{1}{b^n} \\
&= \frac{n}{b^n} \int_0^b y^n dy \\
&= \frac{n}{b^n} \frac{y^{n+1}}{n+1} \Big|_0^b \\
&= \frac{n}{b^n} \frac{b^{n+1}}{n+1} \\
&= \frac{n}{n+1} b
\end{aligned}$$

Since this is not equal to b, the MLE is a biased estimator for the parameter b. However, we can convert this to an unbiased estimator by multiplying it by $(n+1)/n$. Thus the estimator

$$\frac{n+1}{n} Y_{(n)} = \frac{n}{n+1} \max(Y_1, Y_2, \ldots, Y_n)$$

is an unbiased estimator for b.

Example. Suppose we have a population which has a geometric distribution with parameter p. Take samples $Y_1, \ldots, Y_n$ from this distribution. Find the MLE for p.
Recall that the pmf for the geometric distribution with parameter p is:

$$p(k) = \mathbb{P}(Y = k) = (1-p)^{k-1} p$$

Then our likelihood function is:

$$L(Y_1, \ldots, Y_n | p) = \prod_{i=1}^n (1-p)^{Y_i - 1} p$$

To maximize this, we need to use calculus. We will differentiate with respect to p and set the derivative equal to zero. Since this requires an annoying amount of algebra, we will use a trick. Since the log function is strictly increasing, the log of the likelihood function and the likelihood function attain their maximum at the same value of p. So we can find the maximum of the log likelihood function instead, and this will give us the correct maximum for p.

$$\begin{aligned}
\log L(Y_1, \ldots, Y_n | p) &= \log \prod_{i=1}^n (1-p)^{Y_i - 1} p \\
&= \sum_{i=1}^n \log p + \sum_{i=1}^n \log(1-p)^{Y_i - 1} \\
&= n \log p + \sum_{i=1}^n (Y_i - 1) \log(1-p)
\end{aligned}$$

Taking the derivative with respect to p:

$$\frac{d}{dp}\log L(Y_1,...,Y_n|p) = \frac{n}{p} - \frac{1}{1-p}\left(\sum_{i=1}^{n} Y_i - n\right)$$

For this to be 0 we require:

$$\begin{aligned}
\frac{1}{1-p}\left(\sum_{i=1}^{n} Y_i - n\right) &= \frac{n}{p} \\
p\left(\sum_{i=1}^{n} Y_i - n\right) &= n(1-p) \\
p\sum_{i=1}^{n} Y_i - np &= n - np \\
p &= \frac{n}{\sum_{i=1}^{n} Y_i} = \frac{1}{\bar{Y}}
\end{aligned}$$

Thus the MLE for the parameter p is $1/\bar{Y}$. This is the same estimator as we found by the method of moments above.

Example. Suppose we have a population which has a Poisson distribution with parameter λ. Take samples $Y_1, \ldots, Y_n$ from this distribution. Find the MLE for λ.

Using the Poisson pmf, The likelihood function is:

$$\begin{aligned}
L(Y_1,\ldots,Y_n|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{Y_i}}{Y_i!} \\
&= e^{-n\lambda}\lambda^{\sum_{i=1}^{n} Y_i} \prod_{i=1}^{n} \frac{1}{Y_i!} \\
&= e^{-n\lambda}\lambda^{n\bar{Y}} \prod_{i=1}^{n} \frac{1}{Y_i!}
\end{aligned}$$

To maximize this with respect to λ, we will maximize the log likelihood function.

$$\begin{aligned}
\log L(Y_1,\ldots,Y_n|\lambda) &= \log(e^{-n\lambda}) + \log(\lambda^{n\bar{Y}}) + \log\left(\prod_{i=1}^{n} \frac{1}{Y_i!}\right) \\
&= -n\lambda + n\bar{Y}\log(\lambda) + \log\left(\prod_{i=1}^{n} \frac{1}{Y_i!}\right)
\end{aligned}$$

Taking the derivative with respect to λ:

$$\frac{d}{d\lambda}\log L(Y_1,\ldots,Y_n|\lambda) = -n + \frac{n\bar{Y}}{\lambda}$$

Setting this equal to 0, we get $\lambda = \bar{Y}$, which is the same estimator we got using the method of moments.

7 Hypothesis Testing

7.1 Introduction

Statistical hypothesis testing is a formal procedure for conducting scientific inquiry according to the scientific method. In the scientific method, a scientist poses a hypothesis based on observations from nature. She then conducts an experiment to test that hypothesis, and the results of that experiment either support or refute the hypothesis. From our mathematical standpoint, a *hypothesis* is a claim about one or more parameters of a population (or more than one population) of interest. For example, we might claim that a parameter equals a certain value or falls within a certain range. We then take a sample from our population and use an appropriate estimator to estimate the parameter of interest. We then use statistical techniques to either support or refute our hypotheses, and to determine the probability that we made the correct decision.

Here are two examples of hypothesis tests.

1. You are a pollster for a major news outlet, and are interested in the preferences of urban vs. rural voters in Rhode Island. Your hypothesis is that rural voters are less likely to prefer Gina Raimondo. The parameter of interest is $p_1 - p_2$, the difference in the proportion of Raimondo supporters in the two populations. You then sample registered voters from both populations and construct the estimator $\hat{p}_1 - \hat{p}_2$. Using statistics, you will be able quantify the probability that you correctly accept or reject your hypothesis based on the value of the estimator. In particular, you will be able to determine the probability that you made the *wrong* decision in accepting or rejecting your hypothesis.

2. You are the principal investigator for a trial of a new drug to treat hypertension (high blood pressure). You claim that your new drug will reduce systolic blood pressure by 10 mmHg compared to a placebo pill. The parameter of interest is $\mu_1 - \mu_2$, the difference in mean blood pressure reductions between the mythical populations of all people with hypertension who would receive the drug and all people with hypertension who would receive a placebo. Your hypothesis is that $\mu_1 - \mu_2 \geq 10$. To test this, you will do a double-blind study and randomly assign 100 patients to receive the drug and 100 patients to receive the placebo pill. You will use the estimator $\bar{Y}_1 - \bar{Y}_2$ to test your hypothesis. Using statistics, you will be able quantify the probability that you correctly accept or reject your hypothesis based on the value of the estimator.

Hypothesis testing uses the ideas we learned in the previous sections – sampling, estimators, and confidence intervals – to make inferences about a general population using samples drawn from that population and to quantify how confident we are with these inferences.

7.2 Elements of a Hypothesis Test

There are four elements of any hypothesis test. We will go through an example as we discuss them.

Example. Suppose we are interested in the voting preference in Rhode Island for the gubernatorial election.

1. We start with a hypothesis that we would like to support. This is called the *alternative hypothesis*. For example, our alternative hypothesis could be "More than 50% of the voters in Rhode Island support Gina Raimondo". In mathematical terms, our population of interest is the registered voters in Raimondo, and our parameter of interest is p, the proportion of registered voters who support Raimondo. The alternative hypothesis can be stated in terms of p as $p > 0.5$.

2. We will obtain support for our alternative hypothesis by showing (to a specified degree of confidence) that the *null hypothesis*, the converse of the null hypothesis, is false. You may have heard of this as "rejecting the null hypothesis". The null hypothesis in this case is "50% or fewer voters in Rhode Island support Raimondo". Mathematically, we could state this as $p \leq 0.5$ if we like. However, it turns out that this is not a useful way to state the null hypothesis. Recall that we wish to find evidence to *reject* the null hypothesis. The only way to reject the hypothesis $p \leq 0.5$ is for our estimator $\hat{p}$ to produce a value over 0.5. Consider the hypothesis $p = 0.5$. Suppose our estimator $\hat{p}$ is 0.6, and we decide that this is sufficient evidence to reject the hypothesis $p = 0.5$. If we will reject $p = 0.5$, we will certainly also reject $p \leq 0.5$, since if we are willing to reject $p = 0.5$, we are even more willing to reject anything lower. Thus it suffices to take $p = 0.5$ as our null hypothesis.

3. A *test statistic* is something we can actually measure which we will use to either reject or not reject the null hypothesis. In general, the test statistic will be one of our common estimators, such as $\bar{Y}$ or $\hat{p}$, or the equivalent estimators for the difference between two populations. In this case, since we are interested the population proportion p, we will use $\hat{p} = Y/n$ for our estimator, where Y is the number of people out of a sample of size n who prefer Raimondo.

4. Armed with a test statistic, the only thing left is to decide when we will reject the null hypothesis. A *rejection region* (RR) specifies the values of the test statistic for which the null hypothesis will be rejected. The RR is given as a range a values. In this case, since we will reject the null hypothesis if $\hat{p}$ is *high*, the rejection region will look like $\hat{p} \geq k$, where k is a threshold we will choose. A lower value of k means that we are more likely to reject the null hypothesis; in particular, we are more likely to reject the null hypothesis when it is in fact true, so we are more likely to commit a false positive error. Conversely, a higher value of k means that we are less likely to reject the null hypothesis; in fact, we are more likely to fail to reject the null hypothesis when in fact the null hypothesis is false, thus we are more likely to commit a false positive error. Thus there is no idea value of k for us to choose. We will discuss later how to pick a k based on the amount of false positives and false negatives we are willing to accept.

We can summarize the elements of any hypothesis test in the following table. Although the null hypothesis is "more fundamental", I list the alternative hypothesis first, since that lets us find the form of the null hypothesis.

Elements of a Hypothesis Test

---

1. Alternative hypothesis, H_a
2. Null hypothesis, H_0
3. Test statistic
4. Rejection region (RR)

Before we go any further, we will take several real-world examples of hypothesis tests, and specify their four parameters. We will discuss how to choose rejection regions and how to quantify our confidence in rejecting the null hypothesis in the next sections. For completeness, the first one will be the example we did above. We will use these examples throughout this section.

1. You are a pollster who is interested in the voting preference in Rhode Island for the gubernatorial election. The population of interest is the number of registered voters in Rhode Island, and the parameter of interest is the p, the proportion of voters who are Gina Raimondo supporters.
   (a) Alternative hypothesis, $H_a : p > 0.5$
   (b) Null hypothesis, $H_0 : p = 0.5$
   (c) Test statistic, $\hat{p} = Y/n$
   (d) Rejection region (RR), $\{\hat{p} > k\}$
2. You are the principal investigator for a trial of a new drug to treat hypertension. The populations of interest are all patients with hypertension who receive you drug and all patients with hypertension who receive a placebo. The parameter is interest is the difference in mean blood pressure reduction $\mu_1 - \mu_2$ between these two populations.
   (a) Alternative hypothesis, $H_a : \mu_1 - \mu_2 > 10$
   (b) Null, $H_0 : \mu_1 - \mu_2 = 10$
   (c) Test statistic, $\bar{Y_1} - \bar{Y_2}$
   (d) Rejection region (RR), $\{\bar{Y_1} - \bar{Y_2} > k\}$
3. You are a mechanical engineer and have designed a ball bearing machine which produces ball bearings which are 5 mm in diameter. Since your customers demand precision, the machine needs to be recalibrated if the average ball bearing diameter deviates from 5 mm by more than 1 percent. You suspect there might be something wrong with the machine. The population of interest is all ball bearings produces by your machine. The parameter of interest is μ, the average ball bearing diameter.

(a) Alternative hypothesis, $H_a : \mu \neq 5$

(b) Null hypothesis, $H_0 : \mu = 5$

(c) Test statistic, $\bar{Y}$

(d) Rejection region (RR), $|\bar{Y} - 5| > k$

Note that there is something unusual about the third example. For the first two examples, the alternative hypothesis was that the parameter of interest is *above* a certain value. These are known as upper-tail hypothesis tests, since we will reject the null hypothesis if the test statistic falls in the upper tail of the appropriate probability distribution. Similarly, we could have a lower-tail hypothesis test if the alternative hypothesis is that the parameter of interest is *below* a certain value. In this case, the null hypothesis would be rejected if the test statistic falls in the lower tail of the appropriate probability distribution.

The final example is known as a *two-tailed hypothesis test*. The alternative hypothesis is that the parameter of interest is not equal to some value, i.e. that it is either above or below that value. The null hypothesis, as always, is that the parameter of interest is equal to a specific value (in this case, we do not require an additional argument to show that this makes sense). The rejection region is stated as $|\bar{Y} - 5| > k$, indicating that we will reject the null hypothesis if the test statistic deviates from a specific value by more than a certain amount.

We will first discuss hypothesis tests where the sample we take is large, then we will consider the case where the sample is small. Before we do that, we will define the two types of error which can result from a hypothesis test.

Types of Error

---

1. A *type I error* is made if the null hypothesis is rejected when in fact the null hypothesis is true. The probability of making a type I error is denoted by α:

$$\begin{aligned}\alpha &= \mathbb{P}(\text{reject null hypothesis when null hypothesis is true}) \\ &= \mathbb{P}(\text{test statistic lies in rejection region when null hypothesis is true})\end{aligned}$$

2. A *type II error* is made if the null hypothesis is accepted when the alternative hypothesis is true. The probability of making a type II error is denoted by β:

$$\begin{aligned}\beta &= \mathbb{P}(\text{accept null hypothesis when alternative hypothesis is true}) \\ &= \mathbb{P}(\text{test statistic lies outside rejection region when alternative hypothesis is true})\end{aligned}$$

Some statisticians (especially biostatisticians) will use the term *power*, which is defined as $1 - \beta$.

7.3 Large Sample Hypothesis Tests

In this section, we will test hypotheses about the mean μ or proportion p of a population. This includes the cases where we are interested in the difference between two populations. We will take a large enough sample that we can assume (by the central limit theorem) that the test statistic is normally distributed. Thus we can use the Z distribution (standard normal) in our computations.

Let's look at one-tailed hypothesis tests first. We will do an upper-tailed test, but this is similar for lower-tailed tests. Suppose we have a parameter of interest θ (e.g. μ, p, or the equivalent for the difference of two populations). We wish to test the alternative hypothesis $\theta > \theta_0$ (upper-tailed test), so the null hypothesis is $\theta = \theta_0$. The test statistic is the estimator $\hat{\theta}$, where we choose the appropriate estimator based on the parameter of interest. Since this is an upper-tail test, the rejection region will be of the form $\hat{\theta} > k$, where k will be chosen later. Thus the parameters for the test are:

1. Alternative hypothesis, $H_a : \theta > \theta_0$
2. Null hypothesis, $H_0 : \theta = \theta_0$
3. Test statistic, $\hat{\theta}$
4. Rejection region (RR), $\hat{\theta} > k$

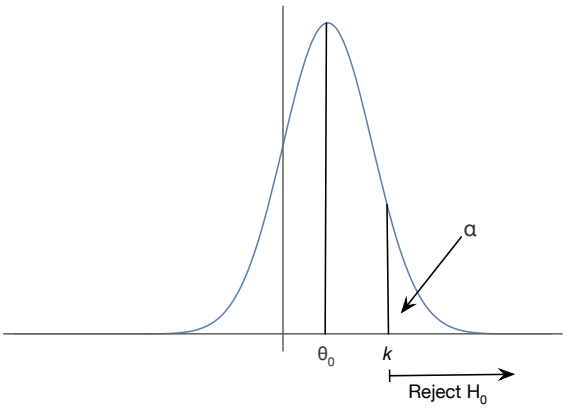
To determine the parameter k in the rejection region, we fix the level α of type I error which we are willing to accept. If the null hypothesis is true, then the estimator $\hat{\theta}$ has a normal distribution with mean θ_0 and standard deviation $\sigma_{\hat{theta}}$. Since α is the probability of incorrectly rejecting the null hypothesis when it is in fact true, we want to choose the rejection region such that the probability that $\hat{\theta} \geq k$ to be α. Converting to the standard normal random variable, we want:

$$\begin{aligned} \mathbb{P}(\hat{\theta} \geq k) &= \alpha \\ \mathbb{P}\left(\frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} \geq \frac{k - \theta_0}{\sigma_{\hat{\theta}}}\right) &= \alpha \\ \mathbb{P}\left(Z \geq \frac{k - \theta_0}{\sigma_{\hat{\theta}}}\right) &= \alpha \end{aligned}$$

Looking at the Z table, we choose z_α such that $\mathbb{P}(Z \geq z_\alpha) = \alpha$. (On our Z table, this is equivalent to $\mathbb{P}(Z \leq -z_\alpha) = \alpha$). Thus we have:

$$\begin{aligned} \frac{k - \theta_0}{\sigma_{\hat{\theta}}} &= z_\alpha \\ k &= \theta_0 + z_\alpha \sigma_{\hat{\theta}} \end{aligned}$$

This is shown in the diagram below:



Recall that $\sigma_{\hat{\theta}}$ is the standard deviation of the estimator, which is computed from the the population standard deviation and the sample size. If we do not know the population standard deviation, we can use the sample standard deviation S in place of the population standard deviation σ since the sample size is large.

Example. You are a pollster who is interested in the voting preference in Rhode Island for the gubernatorial election. The population of interest is the number of registered voters in Rhode Island, and the parameter of interest is the p, the proportion of voters who support Gina Raimondo. Suppose you sample 100 voters and 60 of them favor Raimondo. Does the evidence support Raimondo being favored at a level of 0.05?

The parameters of the test are given above. The test statistic is $\hat{p} = 0.6$. To find the appropriate rejection region at a level of $\alpha = 0.05$ we need to find the appropriate value of z_α from the Z table. Looking at the table, we find that $z_\alpha = 1.64$ (we could also have chosen 1.65). We do not know the standard deviation $\sigma_{\hat{p}}$ for the test statistic $\hat{p}$, but since we are centering our confidence interval about the null hypothesis ($p = 0.5$), we estimate the standard deviation of $\hat{p}$ by assuming that the null hypothesis holds, i.e.

$$\sigma_{\hat{p}} \approx \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.5)(0.5)}{100}} = 0.05$$

From this we calculate:

$$k = 0.5 + z_\alpha \sigma_{\hat{p}} = 0.5 + 1.64(0.05) = 0.582$$

Thus the rejection region is

$$\{\hat{p} \geq 0.582\}$$

Since our test statistic is 0.6, it falls inside the rejection region, thus we can reject our null hypothesis with a level of 0.05, so we are 95% confident that Raimondo is favored in the population at large.

Similarly, we can do this for a two-tailed hypothesis test. Here we "split" the α between the

upper and lower tails of the normal distribution.

$$\begin{aligned} \mathbb{P}(|\hat{\theta} - \theta_0| \geq k) &= \alpha \\ \mathbb{P}\left(\left|\frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}\right| \geq \frac{k}{\sigma_{\hat{\theta}}}\right) &= \alpha \\ \mathbb{P}\left(|Z| \geq \frac{k}{\sigma_{\hat{\theta}}}\right) &= \alpha \end{aligned}$$

Splitting the α evenly between the upper and lower tails, we want

$$\mathbb{P}\left(Z \leq -\frac{k}{\sigma_{\hat{\theta}}}\right) = \alpha/2$$

and

$$\mathbb{P}\left(Z \geq \frac{k}{\sigma_{\hat{\theta}}}\right) = \alpha/2$$

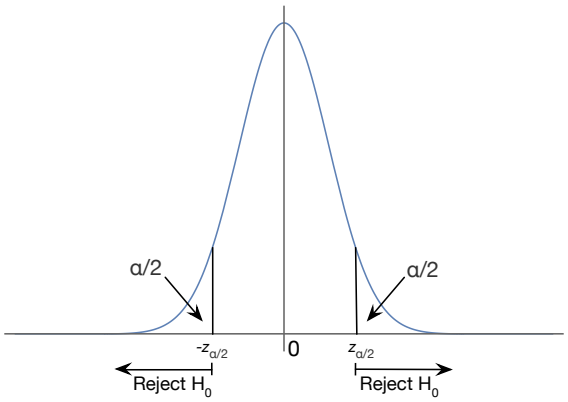
In the same way as before, we can use the Z table to find the appropriate value of $z_{\alpha/2}$. Thus we have:

$$\begin{aligned} \frac{k}{\sigma_{\hat{\theta}}} &= z_{\alpha/2} \\ k &= z_{\alpha/2}\sigma_{\hat{\theta}} \end{aligned}$$

Our rejection region is:

$$\{|\hat{\theta} - \theta_0| \geq z_{\alpha/2}\sigma_{\hat{\theta}}\}$$

The diagram below shows the rejection region for a two-tailed test in terms of the standard normal Z distribution.



Let's do our ball bearing example.

Example. You are a mechanical engineer and have designed a ball bearing machine which produces ball bearings which are 5 mm in diameter. You suspect there might be something wrong with the machine. You sample 64 ball bearings from the machine, and obtain a sample mean of 4.98 mm and a sample standard deviation of 0.1 mm. Can you conclude at a level of 0.95 that there is something wrong with the machine?
The parameters of the test are given above. The test statistic is $\bar{Y} = 4.98$. To find the appropriate rejection region at a level of $\alpha = 0.05$, since this is a two-sided hypothesis test, we need to find the appropriate value of $z_{\alpha/2}$ from the Z table. Looking at the table, we find that $z_\alpha = 1.96$. We do not know the population standard deviation, but we can estimate it by using S in place of σ. The standard deviation of the estimator is therefore approximately:

$$\sigma_{\bar{Y}} \approx \frac{S}{\sqrt{n}} = \frac{0.1}{\sqrt{64}} = \frac{0.1}{8} = 0.0125$$

Thus our value of k is:

$$k = z_{\alpha/2}\sigma_{\bar{Y}} = 1.96(0.0125) = 0.0245$$

The rejection region is therefore:

$$\{|\bar{Y} - 5| \geq 0.0245\} = \{\bar{Y} \leq 4.9755 \text{ or } \bar{Y} \geq 5.0245\}$$

Since out test statistic does not fall within our rejection region, we do not reject the null hypothesis, so for now you conclude that you do not have to do maintenance on the machine.

7.4 Large Sample Hypothesis Tests and Type II Error

We have discussed how to choose the rejection region based on the level of type I error we are willing to accept. Now we will discuss type II error. We will only quantify type II error for one-tailed hypothesis tests. For two-tailed tests, this process is arduous, thus will be omitted. The discussion below will concern upper-tail hypothesis tests. Lower-tail tests are similar, except everything is "flipped".

Recall that a type II error is made if we accept the null hypothesis when in fact the null hypothesis is false. In other words, our test statistic falls *outside* the rejection region, even though the null hypothesis is false and should be rejected. When evaluating type II error, we need to make an additional decision. We need to specify a *specific* value of the alternative hypothesis to be a *positive test threshold* (my own term) which we will use to compute β.

Let's use the same upper-tail setup as before. Suppose we have a parameter of interest θ. We wish to test the alternative hypothesis $\theta > \theta_0$ (upper-tailed test), so the null hypothesis is $\theta = \theta_0$. The test statistic is the estimator $\hat{\theta}$, and the rejection region is of the form $\{\hat{\theta} > k\}$. For now, assume we have chosen k according to our desired α using the methods of the previous section. To reiterate, our test has the following parameters:

1. Alternative hypothesis, $H_a : \theta > \theta_0$
2. Null hypothesis, $H_0 : \theta = \theta_0$

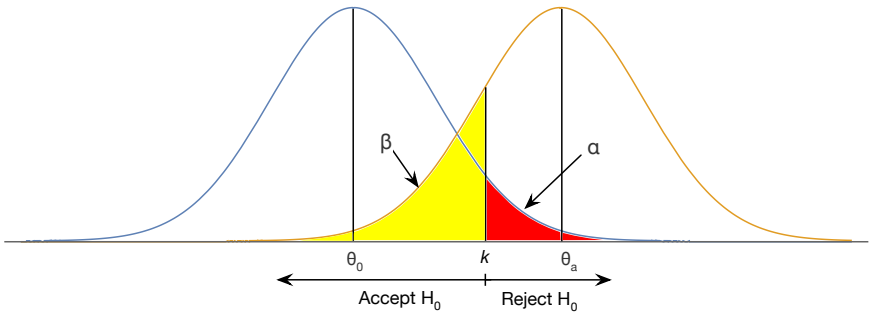
3. Test statistic, $\hat{\theta}$

4. Rejection region (RR), $\{\hat{\theta} > k\}$

We will choose a specific value θ_a for the alternative hypothesis (our positive test threshold). Since this is an upper tail test, we must have $\theta_a > \theta_0$. Then β, the probability of a type II error, is defined as the probability of accepting the null hypothesis when the true value of the parameter is in fact θ_a.

It seems odd that we must specify a value for θ_a in order to do this, but this has a nice interpretation. The probability of accepting the null hypothesis if the true parameter is θ_a is β. If the true value of θ is greater than θ_a, the probability of accepting the null hypothesis is less than β. Thus if the true value of θ is θ_a or greater, the probability that we will incorrectly accept the null hypothesis is at most β.

Now that we've gotten that out of the way, how we we actually find β? Take a look a the following picture:



β is the probability that the test statistic lies outside the rejection region when the true value of the parameter is θ_a. This time, the estimator has a normal distribution with mean θ_a and standard deviation $\sigma_{\hat{\theta}}$. Thus, for the upper tail test here, we want the probability that our test statistic is less than k, the threshold of the rejection region.

$$\begin{aligned}\beta &= \mathbb{P}(\hat{\theta} \leq k) \\ &= \mathbb{P}\left(\frac{\hat{\theta} - \theta_a}{\sigma_{\hat{\theta}}} \leq \frac{k - \theta_a}{\sigma_{\hat{\theta}}}\right) \\ &= \mathbb{P}\left(Z \leq \frac{k - \theta_a}{\sigma_{\hat{\theta}}}\right)\end{aligned}$$

Since we know θ_a and $\sigma_{\hat{\theta}}$, and since we have already computed the boundary of the rejection region k, we can use the Z table to solve for k
Let's go back and do our polling example from above.

Example. You are again a pollster who is interested in the voting preference in Rhode Island for the gubernatorial election. The population of interest is the number of registered voters in Rhode Island, and the parameter of interest is the p, the proportion of voters who are Gina Raimondo supporters. You sample 100 voters and 60 of them favor Raimondo. You create a hypothesis test with $\alpha = 0.05$, for which we have seen that the rejection region is $\{\hat{p} \geq 0.582\}$. For a value of the alternative hypothesis $p_a = 0.60$, calculate β, the probability of a type II error.

In this case, since we are assuming the alternative hypothesis $p_a = 0.60$ is true, we estimate the standard deviation $\sigma_{\hat{p}}$ of the test statistic $\hat{p}$ using the value of p from the alternative hypothesis.

$$\sigma_{\hat{p}} \approx \sqrt{\frac{p_a(1-p_a)}{n}} = \sqrt{\frac{(0.6)(0.4)}{100}} = 0.049$$

Plugging in the value of $k = 0.58$ we determined in the example above, and using $p_a = 0.60$ and $\sigma_{\hat{p}} = 0.049$:

$$\begin{aligned}
\beta &= \mathbb{P}\left(Z \leq \frac{k - p_a}{\sigma_{\hat{p}}}\right) \\
&= \mathbb{P}\left(Z \leq \frac{0.58 - 0.60}{0.049}\right) \\
&= \mathbb{P}(Z \leq -0.41) \\
&= 0.3409
\end{aligned}$$

7.5 Sample Size Selection

If you are a scientist devising a hypothesis test, you don't want to just run the test and calculate β after the fact the way we did in the example above. If you did that, β might be too large, and your test could be meaningless! You would like a procedure where, if you specify the maximum values of α and β you are willing to tolerate, you can compute the size of the sample n you need to attain these values. Along the way, you will also compute k, the threshold for the rejection region. The following derivation is for an upper-tail hypothesis test. A similar derivation will work for a lower-tail test. The two-tailed test will not be discussed.

For simplicity, we will do the computation for a hypothesis test involving the sample mean. We can similarly do this for a hypothesis test involving the sample proportion. As an experimenter, you are conducting an upper tail hypothesis test with the following four parameters:

1. Alternative hypothesis, $H_a : \mu > \mu_0$
2. Null hypothesis, $H_0 : \mu = \mu_0$
3. Test statistic, $\bar{Y}$
4. Rejection region (RR), $\{\bar{Y} > k\}$

In addition, you need to choose three more parameters:

1. θ_a, a specific value of the alternative hypothesis. As before, we require $\theta_a > \theta_0$
2. α, the maximum type I error you are willing to accept
3. β, the maximum type II error you are willing to accept

The population standard deviation is denoted σ. If we do not know σ, we can estimate it by the sample standard deviation S. In the real world, to obtain S, we can do a pilot study for the express purposes of computing S. We just need to make sure that the pilot study is large enough that the S we obtain is a good estimator for σ.

As above, we will write the appropriate equations for α and β. Since we will have two equations and only two unknowns (k and n), we can solve them for the unknowns. Using the definition of α and referring to the picture above:

$$\begin{aligned}
\alpha &= \mathbb{P}(\text{test statistic is in rejection region when null hypothesis is true}) \\
&= \mathbb{P}(\bar{Y} \geq k \text{ when } \mu = \mu_0) \\
&= \mathbb{P}\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{k - \mu_0}{\sigma/\sqrt{n}}\right) \\
&= \mathbb{P}(Z \geq z_\alpha)
\end{aligned}$$

where z_α is chosen from our Z-table so that $\mathbb{P}(Z \geq z_\alpha) = \alpha$. Using the definition of β and again referring to the picture above:

$$\begin{aligned}
\beta &= \mathbb{P}(\text{test statistic is outside rejection region when alternative hypothesis is true}) \\
&= \mathbb{P}(\bar{Y} \leq k \text{ when } \mu = \mu_a) \\
&= \mathbb{P}\left(\frac{\bar{Y} - \mu_a}{\sigma/\sqrt{n}} \leq \frac{k - \mu_a}{\sigma/\sqrt{n}}\right) \\
&= \mathbb{P}(Z \leq -z_\beta)
\end{aligned}$$

where z_β is chosen from our Z-table so that $\mathbb{P}(Z \leq -z_\beta) = \beta$. We use the negative sign for convenience, since the value of z we are looking for will always fall to the left of 0 on a graph of the standard normal distribution. We then have two equations we can solve simultaneously:

$$\begin{aligned}
\frac{k - \mu_0}{\sigma/\sqrt{n}} &= z_\alpha \\
\frac{k - \mu_a}{\sigma/\sqrt{n}} &= -z_\beta
\end{aligned}$$

If we solve both equations for k, we get:

$$\begin{aligned}
k &= \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \\
k &= \mu_a - z_\beta \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

We can use the first equation in the set above to find k, the boundary of the rejection region. This is the same equation we derived in the section on large sample hypothesis tests. Setting these two equations equal to each other:

$$\begin{aligned}
\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} &= \mu_a - z_\beta \frac{\sigma}{\sqrt{n}} \\
(z_\alpha + z_\beta)\frac{\sigma}{\sqrt{n}} &= \mu_a - \mu_0 \\
\sqrt{n} &= \frac{(z_\alpha + z_\beta)\sigma}{\mu_a - \mu_0} \\
n &= \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_a - \mu_0)^2}
\end{aligned}$$

7.6 p-values

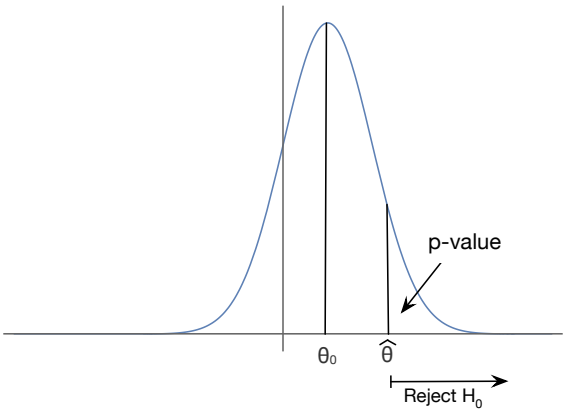
The parameter α is the probability of making a type I error, i.e. rejecting the null hypothesis when it is in fact true. Generally, experimenters choose small values of α to maximize their confidence that a positive result is not a false positive, i.e. due to chance alone. Typical values of α you will see in the scientific literature are 0.05 and 0.01, although this is a little arbitrary.

Another way of reporting the probability of a type I error is the p-value. This is what is most often given in the scientific literature.

p-value

The p-value, or attained significance level, of a test is the smallest level of significance α for which the test statistic indicates that the null hypothesis should be rejected.

In other words, for any α greater than or equal to the p-value, the null hypothesis can be rejected using the test statistic. For α less than the p-value, the null hypothesis cannot be rejected. This provides more information than just saying that the null hypothesis was rejected for a certain value of α. The following picture may be useful. It shows an upper-tail hypothesis test, but this can be computed for any hypothesis test. The null hypothesis is given by θ_0, and the test statistic is $\hat{\theta}$.



Let's look at our voter polling example from before.

Example. You are a pollster who is interested in the voting preference in Rhode Island for the 2016 gubernatorial election. The population of interest is the number of registered voters in Rhode Island, and the parameter of interest is the p, the proportion of voters who support Gina Raimondo. Suppose you sample 100 voters and 60 of them favor Raimondo. What is the p-value for this test? (It is a little confusing that p is the sample proportion and we also have a p-value for the test, but this notation is standard in both cases; we will write the p-value as "p-value" rather than just p to avoid this confusion).

We want the smallest value of α for which we will reject the null hypothesis. The value we obtained for $\hat{p}$ is 0.6. We estimated the standard deviation $\sigma_{\hat{p}}$ of the estimator $\hat{p}$ as 0.050 in the example above, where we recall that we are assuming that the null hypothesis is true. Thus we have:

$$\begin{aligned} p\text{-value} &= \mathbb{P}(\hat{p} \geq 0.6 \text{ given null hypothesis is true}) \\ &= \mathbb{P}(\hat{p} \geq 0.6 \text{ given } p = 0.5) \\ &= \mathbb{P}\left(\frac{\hat{p} - 0.5}{\sigma_{\hat{p}}} \geq \frac{0.6 - 0.5}{\sigma_{\hat{p}}}\right) \\ &= \mathbb{P}\left(Z \geq \frac{0.1}{0.05}\right) \\ &= \mathbb{P}(Z \geq 2.00) \\ &= 0.0228 \end{aligned}$$

We obtain a p-value of 0.0228, which is the smallest value of α for which the test statistic indicates that we can reject the null hypothesis. Above, we showed that we can reject the null hypothesis with an α of 0.05. This is consistent with our p-value, since 0.05 is greater than the p-value.

7.7 Small sample hypothesis testing for the population mean

In the previous section, we discussed hypothesis testing procedures for large samples. The large sample assumption is important because that guarantees that the test statistic has a

normal distribution (by the central limit theorem). The large sample size also allows us to estimate the population standard deviation using the sample standard deviation.

What happens when the sample size is not large? As in the section on confidence intervals, if we have a population that is normally distributed with unknown standard deviation, we can use the t-distribution in place of the Z distribution. Suppose we have a normally-distributed population, and we take a small sample ($n < 30$) from that population. The population standard deviation is unknown, so we estimate it with the sample standard deviation S. Suppose we are are doing an upper-tailed hypothesis test. Let $\mu = \mu_0$ be the null hypothesis, and $\mu > \mu_0$ be the alternative hypothesis. The test statistic is $\bar{Y}$, the standard unbiased estimator for μ. Then to find the appropriate rejection region based on our desired level α, we follow the same procedure above except that we use the t distribution with $n - 1$ df in place of the Z distribution. This works because

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$$

has a t distribution with $n - 1$ df. For example, if we are doing an upper-tail test with desired level α, then the rejection region is given by:

$$\begin{aligned} k &= \mu_0 + t_\alpha \sigma_{\bar{Y}} \\ &= \mu_0 + t_\alpha \frac{S}{\sqrt{n}} \end{aligned}$$

All we did to get this formula was replace the z with a t. We will return to our rocket science example from the section on confidence intervals.

Example. You are a rocket scientist, and you conduct an experiment which involves measuring the launch velocity of a model rocket. You claim that the launch velocity of the model rocket is more than 29 m/s. Suppose 8 measurements are taken. The sample mean is 29.59 m/s, and the sample standard deviation is 0.391 m/s. Is the claim supported at the 0.025 level of significance?

This is an upper-tailed hypothesis test with the following parameters:

1. Alternative hypothesis, $H_a : \mu > 29$
2. Null hypothesis, $H_0 : \mu = 29$
3. Test statistic, $\bar{Y}$
4. Rejection region (RR), $\{\bar{Y} > k\}$

We assume that the launch velocities are normally distributed (a normally distributed pop-

ulation is essential for use to use the t-distribution). For the rejection region, we have:

$$\begin{aligned} k &= \mu_0 + t_\alpha \sigma_{\bar{Y}} \\ &= 29 + t_\alpha \frac{S}{\sqrt{n}} \\ &= 29 + t_{0.025} \frac{0.391}{\sqrt{8}} \\ &= 29 + 2.365(0.138) \\ &= 29.326 \end{aligned}$$

The rejection region is $\{\bar{Y} \geq 29.326\}$. Since our measurement lies in the rejection region, we can reject the null hypothesis with a level of 0.025. Thus the claim is supported at the 0.025 level of significance.

7.8 Power of Hypothesis Tests

In the previous sections, we have discussed large-sample and small-sample hypothesis tests for various test statistics. We learned how to specify a rejection region for a desired α and how to compute β for a specific value of the alternative hypothesis. How did we decide on those test statistics, and how do we know we selected the best rejection region. In other words, how "good" are the tests?

So far, we have used the parameters α and β, the probabilities of type I and type II error, to measure the "goodness" of a hypothesis test. Recall that to compute β we had to choose a specific value for the alternative hypothesis θ_a. We would like a function which gives us the error of the test given the true value of the parameter. The function we use is called the power of a hypothesis test.

Power of a Hypothesis Test

Suppose we have a hypothesis test, with test statistic $\hat{\theta}$ and rejection region RR. Then the *power* of the hypothesis test, denoted $Power(\theta)$ is the probability that the test statistic $\hat{\theta}$ will lie in the rejection region if the true parameter value is θ, i.e.

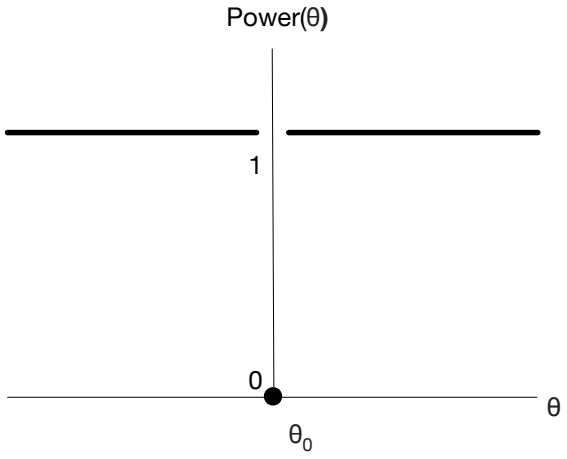
$$Power(\theta) = \mathbb{P}(\hat{\theta} \text{ lies in RR when true parameter value is } \theta)$$

The power function relates α and β in the following way. Suppose our null hypothesis is θ_0, and we have chosen the specific value θ_a for our alternative hypothesis. Let α be the probability of a type I error, and let $\beta(\theta_a)$ be the probability of a type II error when the true value is θ_a. Then we observe the following:

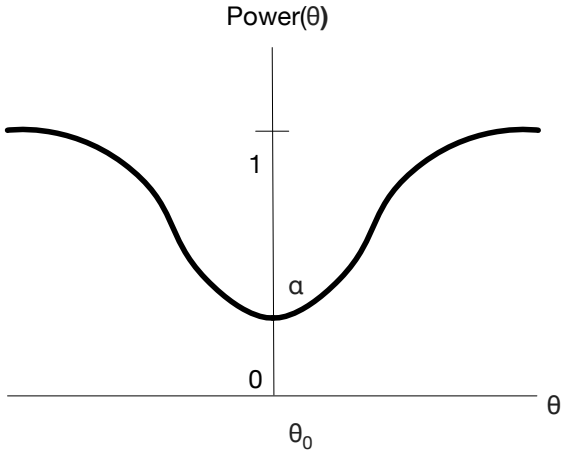
1. $Power(\theta_0) = \alpha$ since $Power(\theta_0)$ is the probability of rejecting the null hypothesis when it is true.

2. $Power(\theta_a) = 1 - \beta(\theta_a)$, since $\beta(\theta_a)$ is the probability of accepting the null hypothesis when θ_a is true.

For an ideal test, the power function would be 0 at θ_0 and 1 for all possible values of the alternative hypothesis θ_a. Here is what this would look like graphically for a two-tailed hypothesis test.



No hypothesis test, however, is perfect. Realistically, the power curve for a two-tailed hypothesis test will look more like this:



What we would like is a test which maximizes the power function for a given α. This is called the *most powerful α-level test*. Before we state the condition under which we have such a test, we need one more set of definitions.

Simple and Composite Hypotheses

Suppose we have a random sample taken from a population with parameter θ, and consider a hypothesis specifying the value of θ. If the hypothesis uniquely specifies the distribution of the population, we call our hypothesis a *simple hypothesis*. Otherwise, we call it a *composite hypothesis*.

Let's look at examples of simple and composite hypotheses.

Example.

1. Suppose we have a population which is normally distributed with mean μ and variance 1. Since the distribution of the population is uniquely specified by μ, a hypothesis involving μ such as the null hypothesis $\mu_0 = 0$, is a simple hypothesis.

2. Suppose we have a population which is normally distributed with mean μ and unknown variance σ^2. Here we need both μ and σ^2 to specify the distribution of the population, so a hypothesis involving μ such as the null hypothesis $\mu_0 = 0$, is a composite hypothesis.

3. Suppose we have a population with is exponentially distributed with parameter λ. Then a hypothesis involving λ is a simple hypothesis since λ uniquely specifies the distribution of the population. Since the population mean $\mu = 1/\lambda$, a hypothesis involving the mean μ is also a simple hypothesis.

Consider a hypothesis test where we are testing a simple null hypothesis $\theta = \theta_0$ against a simple alternative hypothesis $\theta = \theta_a$. We would like to choose a rejection region such that:

1. $Power(\theta_0) = \alpha$

2. $Power(\theta_a)$ is as large as possible

In other words, we are looking for the most powerful α-level test. The following theorem tells us how to derive the most powerful α-level test in this case.

Neyman-Pearson Lemma

Suppose we have a population whose distribution parameterized by θ. Suppose we have a hypothesis test in which we wish to test the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_a : \theta = \theta_a$. We do this by taking a random sample $Y_1, \ldots, Y_n$ drawn from the population. Let $L(Y_1, \ldots, Y_n|\theta)$ be the likelihood function for the sample when the value of the parameter is θ. Then, for a given α, the test that maximizes the power for the alternative hypothesis θ_a has a rejection region given by:

$$\frac{L(Y_1, \ldots, Y_n|\theta_0)}{L(Y_1, \ldots, Y_n|\theta_a)} < k$$

where the value of k is chosen so that the test has the desired α. The ratio of likelihood functions is called a *likelihood ratio*. Such a test is the most powerful α-level test for H_0 versus H_a.

Let's look at an example application of this theorem.

Example. Suppose Y is a single observation from a population parameterized by θ with probability density function

$$f_\theta(y) = \begin{cases} \theta y^{\theta-1} & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the most powerful hypothesis test with $\alpha = 0.05$ to test the null hypothesis $H_0 : \theta = 2$ against the alternative hypothesis $H_a : \theta = 1$.

Since the distribution of the population is uniquely determined by the parameter θ, both hypotheses are simple, so we can use the Neyman-Pearson lemma to derive the most powerful test. Since there is only one sample, the likelihood ratio is given by:

$$\frac{L(Y|\theta_0)}{L(Y|\theta_a)} = \frac{f_{\theta_0(Y)}}{f_{\theta_a(Y)}} = \frac{2(Y)}{1(Y^0)} = 2Y$$

So the rejection region is of the form $2Y < k$, where we will determine k based on the desired level $\alpha = 0.05$. Dividing by 2, we get $Y < k/2$, and letting $m = k/2$, the rejection region is of the form $Y < m$. We determine m based on the definition of α.

$$\begin{aligned} 0.05 = \alpha &= \mathbb{P}(Y \text{ lies in RR when the null hypothesis is true}) \\ &= \mathbb{P}(Y \text{ lies in RR when } \theta = 2) \\ &= \mathbb{P}(Y < m \text{ when } \theta = 2) \\ &= \int_0^m 2y dy \\ &= m^2 \end{aligned}$$

Thus we have $m^2 = 0.05$, so $m = \sqrt{0.05} = 0.2236$. Thus the rejection region for the 0.05-most powerful test is:

$$\{Y < 0.2236\}$$

In other words, among all hypothesis tests for $H_0 : \theta = 2$ versus $H_a : \theta = 1$ based on a sample size of 1 and $\alpha = 0.05$, this test has the largest possible value for $Power(\theta_a) = Power(1)$. Equivalently, this test has the smallest type II error given a sample size of 1 and $\alpha = 0.05$ for this specific pair of alternative and null hypotheses. What is the actual value of $Power(1)$

in this case. Using the definition of the power of a hypothesis test,

$$\begin{aligned} Power(1) &= \mathbb{P}(Y \text{ lies in RR when } \theta = 1) \\ &= \mathbb{P}(Y < 0.2236 \text{ when } \theta = 1) \\ &= \int_0^{0.2236} 1 dy \\ &= 0.2236 \end{aligned}$$

The value 0.2236 is the maximum value of the power of the test among all tests with $\alpha = 0.05$. But for this test, $\beta = 1 - 0.2236 = 0.7764$, which is very large. So this test is not very good. However, no other test with these same parameters is any better.

What about our hypothesis tests from the previous section. It turns out that they are the uniformly best hypothesis tests for the given situations. We give one example (without proof) below.

Example. Suppose $Y_1, \ldots, Y_n$ are samples drawn from a normal distribution with unknown mean μ and known variance σ^2. We wish to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_a : \mu > \mu_0$. Then the α-most powerful test is given by $\bar{Y} > k$, where

$$k = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

This is the test statistic and rejection region we used earlier for large-sample hypothesis testing. Thus our large sample hypothesis test using the Z-distribution is in fact the best hypothesis test we can construct. The proof of this result is an application of the Neyman-Pearson lemma. It is relatively straightforward, but the calculus and algebra are messy, so we will omit that here.

7.9 Likelihood Ratio Tests

In the previous section, we learned how to construct the most powerful α-level test for simple hypotheses. In this case, the distribution of the population is known except for the value of a single parameter θ, and the null and alternative hypothesis are specified in terms of θ. The Neyman-Pearson lemma shows us how to use a likelihood ratio test to construct the most powerful α-level test.

In many cases, we are interested in testing hypotheses involving one parameter, but the population has more than one unknown parameter. We have already encountered this in the section on small-sample hypothesis tests for the mean of normally distributed populations where the population variance is unknown. In this case, the parameter of interest is the population μ. We don't care about the unknown population standard deviation σ, so it is called a *nuisance parameter*. We are also interested in cases where we do not have to choose a specific value for the alternative hypothesis, like we did when we used the Neyman-Pearson lemma. In both these cases (multiple unknown parameters and more complicated alternative hypotheses), we can use a *likelihood ratio test*.

Here is the setup for a likelihood ratio test:

1. We have a population which is parameterized by a set of parameters $\Theta = (\theta_1, \ldots, \theta_n)$. For example, if the population is normally distributed, it is parameterized by $\Theta = (\mu, \sigma)$.

2. We take a sample of size n of independent samples $Y_1, \ldots, Y_n$ from the population.

3. Given specific values of the parameters $\Theta = (\theta_1, \ldots, \theta_n)$, the likelihood function for our sample is denoted $L(Y_1, \ldots, Y_n|\Theta)$. For a normally distributed population, our likelihood function will depend on $\Theta = (\mu, \sigma)$.

4. The null hypothesis states that Θ lies in a particular set of values Ω_0, where the alternative hypothesis states that Θ lies in another set of values Ω_a, where Ω_0 and Ω_a must be disjoint (it does not make sense otherwise). Note that the null and alternative hypotheses no longer need to be single points. They do not have to be simple hypotheses since they can contain unknown parameters or multiple values of a parameter. For example, if we have a population which is exponentially distributed with parameter λ, then if we want to test the null hypothesis $H_0 : \lambda = \lambda_0$ versus the alternative hypothesis $\lambda \neq \lambda_0$, then we would have $\Omega_0 = \{\lambda_0\}$ and $\Omega_a = \{\lambda > 0 : \lambda \neq \lambda_0\}$.

5. The *parameter space* is defined to be $\Omega = \Omega_0 \cup \Omega_a$, which is all possible values of the parameters. In the exponential example $\Omega = \{\lambda > 0\}$, which is all possible values of an exponential parameter.

6. We define:

$$L(\hat{\Omega}_0) = \max_{\Theta \in \Omega_0} L(Y_1, \ldots, Y_n|\Theta)$$
$$L(\hat{\Omega}) = \max_{\Theta \in \Omega} L(Y_1, \ldots, Y_n|\Theta)$$

   We can think of $L(\hat{\Omega}_0)$ as the "best explanation" for the observed data given the null hypothesis is true, i.e. $\Theta \in \Omega_0$. $L(\hat{\Omega}_0)$ is the "best explanation" for the observed data given all possible values of Θ. If $L(\hat{\Omega}_0) = L(\hat{\Omega})$, then the "best explanation" of the observed data is the null hypothesis, so we should accept the null hypothesis. If $L(\hat{\Omega}_0) < L(\hat{\Omega})$, the "best explanation" of observed data is found inside Ω_a, and we should consider rejecting the null hypothesis in favor of the alternative hypothesis.

7. A *likelihood ratio test* is based on the likelihood ratio $L(\hat{\Omega}_0)/L(\hat{\Omega})$.

Likelihood Ratio Test

Given the setup above, define the likelihood ratio λ by

$$\lambda = \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} = \frac{\max_{\Theta \in \Omega_0} L(Y_1, \ldots, Y_n|\Theta)}{\max_{\Theta \in \Omega} L(Y_1, \ldots, Y_n|\Theta)}$$

A likelihood ratio test of the null hypothesis $H_0 : \Theta \in \Omega_0$ versus the alternative hypothesis $H_a : \Theta \in \Omega_a$ has the likelihood ratio λ as a test statistic, and the rejection region is given by $\{\lambda \leq k\}$. The specific value of k is chosen so that α is a desired level. It can be shown that $0 \leq \lambda \leq 1$. A value of λ close to 0 indicates that the likelihood of the sample is much smaller under H_0 than H_a, which favors rejection of the null hypothesis.

The following example of a likelihood ratio test is presented without proof, and justifies our t-test for small samples drawn from a normally distributed population with unknown variance.

Example. Suppose $Y_1, \ldots, Y_n$ are samples drawn from a normal distribution with unknown mean μ and *unknown* variance σ^2. In this case, n is small. We wish to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_a : \mu > \mu_0$. Then if we use the likelihood ratio test described above, we obtain a hypothesis test with test statistic $\bar{Y}$ and rejection region $\bar{Y} > k$. The value is k is given by

$$k = \mu_0 + t_\alpha \frac{S}{\sqrt{n}}$$

where S is the sample standard deviation computed from our unbiased estimator for the sample variance. Thus the likelihood ratio test for this scenario is exactly the t-test we discussed earlier. The proof of this is several pages of messy calculus and algebra, and will be omitted.

Unfortunately, in most cases, the likelihood ratio does not give us a known distribution such as the t-distribution. It can be shown that in cases where the sample size is large and the underlying distribution is "nice" (this is intentionally vague, but covers many real-world cases), the likelihood ratio λ has a chi-square distribution, thus we can construct hypothesis tests using likelihood ratios.

Chi-Square Likelihood Ratio Test

Take n samples $Y_1, \ldots, Y_n$ from a population parameterized by $\Theta = (\theta_1, \ldots, \theta_k)$. Let

$$L(\Theta) = L(Y_1, \ldots, Y_n | \Theta)$$

be the likelihood function, and let λ be the likelihood ratio as computed above. Let r_0 be the number of free parameters (parameters in the vector Θ which are not specified) in the null hypothesis $H_0 : \Theta \in \Omega_0$, and let r be the number of free parameters in $\Omega = \Omega_0 \cup \Omega_a$. Then the large n, the test statistic $-2\log(\lambda)$ has an approximately chi-square distribution with $r_0 - r$ degrees of freedom (df). (This is the natural logarithm, sometimes written as ln). Since there is a negative sign in front of the test statistic, the rejection region is given by:

$$-2\log(\lambda) > \chi^2_\alpha$$

where χ^2_α is found in the chi-square table based on $r - r_0$ df.

Let's do one final example using this likelihood ratio test.

Example. Suppose you are the quality control engineer for the ACME widget factory, and you wish to compare the number of defective widgets produced per day by two different factories. You observe the number of defective widgets produced per day by each factory for 100 days and find sample means $\bar{X} = 20$ and $\bar{Y} = 22$ for the two factories. Assume the number of defective widgets produced per day by the first factory is a Poisson distribution with parameter θ_1 and the number of defective widgets produced per day by the second factory is a Poisson distribution with parameter θ_2. Use the likelihood ratio test to test $H_0 : \theta_1 = \theta_2$ versus $H_a : \theta_1 \neq \theta_2$ with level $\alpha = 0.01$.

Let $X_1, \ldots, X_n$ be the number of defective widgets from days 1 to n from the first factory, and let $Y_1, \ldots, Y_n$ be the same for the second factory. In this problem, $n = 100$. The population parameter is $\Theta = (\theta_1, \theta_2)$. The null hypothesis is $\Omega_0 = \{(\theta_1, \theta_2) : \theta_1 = \theta_2 = \theta)$, where θ is unknown. Then the likelihood function is the product of all the Poisson pmfs.

$$\begin{aligned} L(\theta_1, \theta_2) &= L(Y_1, \ldots, Y_n | \theta_1, \theta_2) \\ &= \prod_{i=1}^{100} \frac{e^{-\theta_1} \theta_1^{X_i}}{X_i!} \prod_{i=1}^{100} \frac{e^{-\theta_2} \theta_2^{Y_i}}{X_i!} \\ &= \frac{1}{k} \theta_1^{\sum X_i} e^{-n\theta_1} \theta_2^{\sum Y_i} e^{-n\theta_2} \end{aligned}$$

where $k = X_1! \cdots X_n! Y_1! \cdots Y_n!$ and $n = 100$. First, we compute the maximum likelihood under the null hypothesis, i.e. where $\theta_1 = \theta_2$. In this case, the likelihood function is a function of a single parameter θ:

$$L(\theta) = \frac{1}{k} \theta^{\sum X_i + \sum Y_i} e^{-2n\theta}$$

This is the likelihood function for a Poisson random variable with parameter θ. We have $2n$ samples $X_1, \ldots, X_n, Y_1, \ldots, Y_n$. Since the mean of the Poisson random variable is the same as its parameter θ, the maximum likelihood estimator (MLE) for θ is just the sample mean of these $2n$ samples. (The MLE for the population mean is always the sample mean). So $L(\theta)$ is maximized at the MLE:

$$\hat{\theta} = \frac{1}{2n} \left(\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i \right) = \frac{1}{2} (\bar{X} + \bar{Y})$$

Thus the maximum of the likelihood function for the null hypothesis is:

$$\begin{aligned} L(\hat{\Omega}_0) &= \frac{1}{k} \hat{\theta}^{\sum X_i + \sum Y_i} e^{-2n\hat{\theta}} \\ &= \frac{1}{k} \hat{\theta}^{n\bar{X} + n\bar{Y}} e^{-2n\hat{\theta}} \end{aligned}$$

Our alternative hypothesis is $\Omega_a = (\theta_1, \theta_2) : \theta_1 \neq \theta_2$, so for our parameter space Ω we have $\Omega = (\theta_1, \theta_2) : \theta_1, \theta_2 > 0$. The likelihood function $L(\theta_1, \theta_2)$ on Ω is maximized when both

θ_1 and θ_2 are equal to their maximum likelihood estimators, which are $\bar{Y_1}$ and $\bar{Y_2}$, thus the maximum of $L(\theta_1, \theta_2)$ over Ω is given by:

$$L(\hat{\Omega}) = \frac{1}{k}\bar{X}^{\sum X_i}e^{-n\bar{X}}\theta_2^{\bar{Y}}e^{-n\bar{Y}}$$

Taking the likelihood ratio gives us:

$$\begin{aligned}\lambda &= \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} \\ &= \frac{\frac{1}{k}\hat{\theta}^{n\bar{X}+n\bar{Y}}e^{-2n\hat{\theta}}}{\frac{1}{k}\bar{X}^{\sum X_i}e^{-n\bar{X}}\theta_2^{\bar{Y}}e^{-n\bar{Y}}} \\ &= \frac{\hat{\theta}^{n\bar{X}+n\bar{Y}}}{\bar{X}^{n\bar{X}} + \bar{Y}^{n\bar{Y}}}\end{aligned}$$

For this specific problem, $n = 100$, $\bar{X} = 20$, $\bar{Y} = 22$, and $\hat{\theta} = \frac{1}{2}(\bar{X} + \bar{Y}) = 21$. Thus the observed value for λ is:

$$\lambda = \frac{21^{100(20+22)}}{20^{(100)(20)} + 22^{(100)(22)}}$$

This has lots of annoying large exponents, but we only care about the log of this.

$$-2\log(\lambda) = -2[4200\log(21) - 2000\log(20) - 2200\log(22)] = 9.53$$

This quantity has a χ^2 distribution. The number of free parameters in $\Omega = (\theta_1, \theta_2) : \theta_1, \theta_2 > 0$ is 2, since we know neither θ_1 nor θ_2. The number of free parameters in $\Omega_0 = \{(\theta_1, \theta_2) : \theta_1 = \theta_2 = \theta)$ is only 1. Thus $-2\log(\lambda)$ has a χ^2 distribution with $2 - 1 = 1$ degrees of freedom (df). Looking at the chi-square table, we see that $\chi^2_{0.01} = 6.635$ for 1 df. The rejection region is thus:

$$-2\log(\lambda) > 6.635$$

Since our value of $-2\log(\lambda) = 9.53$ lies inside the rejection region, we will reject the null hypothesis at a significance level of 0.01, thus we conclude that the number of defective widgets produced by the two factories is indeed different.

8 Additional Topics

8.1 Sampling from Probability Distributions

We will first talk about how to generate samples from a probability distribution. Why is this useful? Suppose we wish to simulate a radioactive decay process on a computer. A reasonable model for the times between subsequent decays of our radioactive isotope is the exponential distribution. Thus we need to be able to generate samples from an exponential distribution. For common distributions, such as the exponential distribution, there are built-in routines in Matlab or SciPy (python) to do this. It is useful, however, to know how this works since many times the distribution you want to sample does not have a built-in routine.

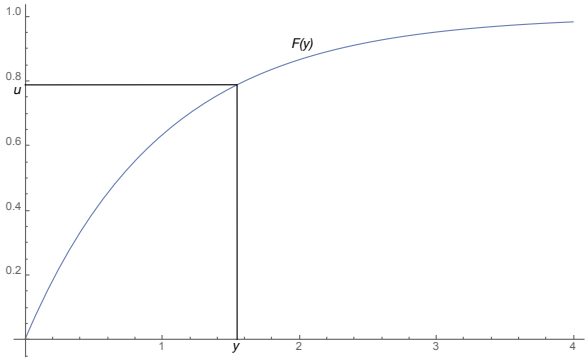
For simplicity, we will only discuss sampling from continuous random variables. The same ideas work for discrete random variable, with a few alterations. We will also assume that we have a method for generating samples from the Uniform[0, 1] distribution. This is a very interesting problem, and there are many ways to do it, most of which are not straightforward. If you find this interesting, the topic is discussed in courses on computational probability and in some computer science courses. We will discuss two methods of sampling from probability distributions: the inverse CDF method and rejection sampling.

8.1.1 Inverse CDF method

Suppose we have a population whose distribution is characterized by a CDF $F(y)$. Recall that for a random variable Y, the CDF is defined as $F(y) = \mathbb{P}(Y \leq y)$. We would like to generate samples from the population. The inverse CDF method works as follows:

1. Generate U, a sample from the Uniform[0, 1] distribution.
2. Let Y be the largest number x such that $F(x) \leq u$. If the CDF $F(y)$ is *strictly* increasing, then $F(y)$ is invertible so we can let $Y = F^{-1}(U)$.

It is perhaps easier to see this on a picture. This is an example of the inverse CDF method used on a population which has a exponential distribution with parameter $\lambda = 1$.



Why does this work. For simplicity, let's consider only the case where the CDF $F(y)$ is invertible. (Invertibility of the CDF is not required; in particular, it works for discrete random variables, whose CDF is not invertible). Let $U \sim \text{Uniform}[0, 1]$. Then we claim the distribution of $F^{-1}(U)$ is F. To see this, we look at the CDF of $F^{-1}(U)$.

$$\begin{aligned}\mathbb{P}(F^{-1}(U) \leq y) &= \mathbb{P}(U \leq F(y)) \\ &= F(y)\end{aligned}$$

where we used the fact that for a Uniform[0, 1] random variable U, the CDF is $\mathbb{P}(U \leq u) = u$ for $u \in [0, 1]$.

Now that we've seen the picture, let's use the inverse CDF method to sample from an exponentially distributed population.

Example. Suppose we have a population which has an exponential distribution with parameter λ. Let U be a Uniform[0, 1] random variable. Use the inverse CDF method to generate a sample from the population in terms of U.

First we need to find the CDF for the population. Integrating the density function from 0 to y:

$$\begin{aligned}F(y) &= \int_0^y \lambda e^{-\lambda t} dt \\ &= e^{-\lambda t}\Big|0^y \\ &= 1 - e^{-\lambda y}\end{aligned}$$

With appropriate bounds, the CDF is:

$$F(y) = \begin{cases} 1 - e^{-\lambda y} & y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

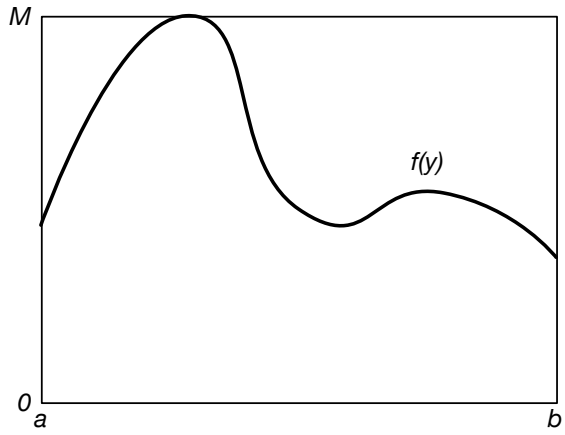
The CDF $F(y)$ is strictly increasing, it has an inverse. Since we want $Y = F^{-1}(U)$, we take $F(Y) = U$ and solve for Y.

$$\begin{aligned}1 - e^{-\lambda Y} &= U \\ e^{-\lambda Y} &= 1 - U \\ -\lambda Y &= \log(1 - U) \\ Y &= -\frac{1}{\lambda}\log(1 - U)\end{aligned}$$

For the exponential distribution this is very straightforward. In general this method works if we can easily compute the CDF. For continuous distributions, if the integral of the density has a nice closed form (as in the exponential case), the inverse CDF method usually works very well. For discrete distributions, this method also works well since to find the discrete CDF, all we have to do is add up the probabilities of the appropriate simple events. For cases where the CDF does not have a closed form (such as the normal distribution) or cases where the CDF is hard to invert, this method is not so good. For many of those cases, rejection sampling is the way to go.

8.1.2 Rejection Sampling

Rejection sampling is based on the "dartboard principle". Here's how it works. Imagine you have a continuous probability density function $f(x)$ you wish to sample from. Furthermore, imagine that the density function is nonzero only on the interval $[a, b]$. Put the density function on a rectangular dartboard. The bounds of the dartboard are from a to b in the x-direction and from 0 to M in the y-direction, where M is the maximum of $f(x)$ on $[a, b]$[26]. Here is an example of a possible dartboard.



Now throw darts uniformly at the dartboard until your dart lands under the density curve $f(x)$. In other words, reject all darts which do not land under the density curve (this is why we call this rejection sampling). Then your dart will be uniformly distributed in the region between the x-axis and the density curve, and the x-coordinate of your dart will be distributed according to the density $f(x)$. Intuitively, this works since there is more room on the board for (nonrejected) darts to land where the density curve is highest, i.e. where the probability density is greatest.

Let's show mathematically that this actually works. We have all the tools we need from the section on multivariate distributions! Let (X, Y) be the position of a nonrejected dart. Then the pair (X, Y) is uniformly distributed on the region between the x-axis and the density curve. Since we are assuming that the density $f(x)$ is zero outside $[a, b]$, the area of the region is $\int_a^b f(x)dx = 1$ since $f(x)$ is a probability density function. Since the joint density function of a uniform distribution is the reciprocal of the area of the region, for our joint density function we have joint density:

$$f(x, y) = \begin{cases} 1 & a \leq x \leq b, 0 \leq y \leq f(x) \\ 0 & \text{otherwise} \end{cases}$$

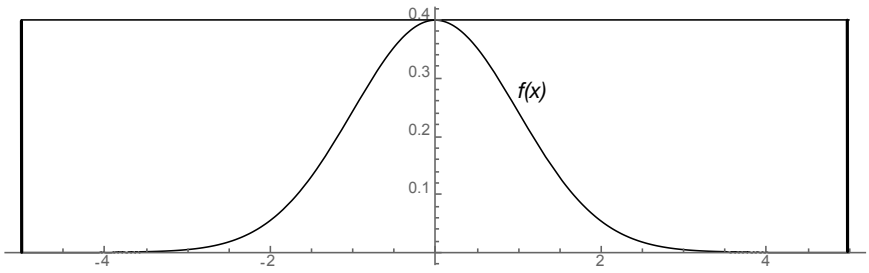
We claim the marginal density of X is the density $f(x)$. To see this, all we have to do is

[26]Recall that a continuous function has an absolute maximum on a closed interval $[a, b]$ and that density functions are nonnegative; thus this rectangular dartboard will have finite size.

integrate the joint density in y.

$$\begin{aligned} f_X(x) &= \int_0^{f(x)} 1 dy \\ &= y\Big|_0^{f(x)} \\ &= f(x) \end{aligned}$$

Thus the rejection sampling method produces works as advertised. There are a few disadvantages to the method, however. First, we need to have a bounded region for the density function. Many useful probability distributions, such as the normal distribution, are unbounded. To use rejection sampling for the standard normal distribution, for example, we have to impose artificial bounds on the density. For example, we could impose the bounds [-5, 5]. Since the probability is exceedingly low that a sample will be more than 5 standard deviations from the mean, this is not too unreasonable; it is important, however, to know that in doing this we are reducing the probability of extreme outliers to 0, which may not be what we want to do, especially if we are taking large numbers of samples. In addition, depending on the shape of the density function, we might have to throw many darts in order for one to not be rejected. This presents no problem theoretically, but might be a problem computationally. Take a look at the following picture of the standard normal density between -5 and 5.



From the picture, the area of the rectangular dartboard is approximately 4, and the area under the curve is approximately 1, since $f(x)$ is a probability density function. Thus we expect only 1/4 of our darts to be accepted, so we will have to throw on average 4 darts to get a single sample.

One way around this computation inefficiency is to note that the x-position of our darts does not have to be uniform. In fact, it makes sense to select the x-position of our darts according to a density function $g(x)$ which is similar to the one we are trying to simulate and is easy to take samples from. We can think of this as throwing darts at a non-square dartboard.

Once again, suppose we are trying to generate a sample from a probability density function $f(x)$, where $f(x)$ is zero outside a closed interval $[a, b]$. Suppose we can generate samples from another probability density function $g(x)$, either using the inverse CDF method or some other method. The function $g(x)$ will give us the shape of our dartboard. For this to work,

the function $f(x)$ must fit entirely on our dartboard. To make this happen, we scale $g(x)$ (if needed) by a constant factor M so that $f(x) \leq Mg(x)$ for all $x \in [a, b]$. We now throw darts uniformly at this dartboard.

At this point, the dartboard analogy breaks down a bit, and it is easier to just give the algorithm.

1. Choose Y uniformly from the interval $[0, 1]$.
2. Choose X according to probability density $g(x)$.
3. If $Y \leq f(X)/Mg(X)$, then accept the sample (X, Y). X is then the desired sample from the distribution $f(x)$.
4. Otherwise reject the sample and repeat from step 1.

Mathematically, why does this work. The proof uses Bayes' theorem. Let A be the event that the sample is accepted. Then by Bayes' theorem (fudging a little because density functions are not exactly probabilities but are close enough),

$$\mathbb{P}(X = x|A) = \frac{\mathbb{P}(A|X = x)P(X = x)}{\mathbb{P}(A)}$$

Since Y is a uniform random variable on [0, 1],

$$\begin{aligned}\mathbb{P}(A|X = x) &= \mathbb{P}\left(Y \leq \frac{f(x)}{Mg(x)}\right) \\ &= \frac{f(x)}{Mg(x)}\end{aligned}$$

where we used the density of the uniform random variable on [0, 1]. The probability $\mathbb{P}(X = x) = g(x)$, since X is chosen according to density function $g(x)$. As mentioned above, this is not quite accurate since densities are not really probabilities, but this is good enough for our purposes. By the Law of Total Probability for continuous random variables (essentially the same as the discrete case, except we replace summation with integration),

$$\begin{aligned}\mathbb{P}(A) &= \int_a^b \mathbb{P}(A|X = x)\mathbb{P}(X = x)dx \\ &= \int_a^b \frac{f(x)}{Mg(x)}g(x)dx \\ &= \frac{1}{M}\int_a^b f(x)dx \\ &= \frac{1}{M}\end{aligned}$$

where we used the fact that $f(x)$ is a density function, thus integrates to 1 over the interval $[a, b]$. Thus the probability of accepting a sample is $1/M$. Putting all of this together,

$$\begin{aligned}\mathbb{P}(X = x|A) &= \frac{\mathbb{P}(A|X = x)P(X = x)}{\mathbb{P}(A)} \\ &= \frac{\frac{f(x)}{Mg(x)}g(x)}{\frac{1}{M}} \\ &= f(x)\end{aligned}$$

Thus the rejection sampling method produces a sample which is distributed according to the desired distribution $f(x)$. The probability of accepting a sample is $1/M$. Thus if we treat the rejection sampling procedure as a sequence of Bernoulli trials with probability of success, then on average it should take M trials for a sample to be accepted. (We model the number of trials needed for the first success as a geometric random variable, and use the formula for the expected value of a geometric distribution.)

8.2 Monte Carlo Methods

Rejection sampling is an example of a Monte Carlo method. Monte Carlo methods are a class of computational algorithms where numerical results are obtained by repeated random sampling. In rejection sampling, we sample repeatedly from a probability distribution until the sample is accepted, then the x-coordinate of the accepted sample is the value we seek. Let's look at Monte Carlo techniques more generally, and then apply them to numerical integration.

There is no standard definition of Monte Carlo methods. In general, the idea is to approximate an expected value with the sample mean of simulated random variables. Ideally, the expected value is a probability, an integral, or something else you care about. Consider a random variable X having probability density function $f(x)$ defined on the interval $[a, b]$. Let $g(x)$ be a real-valued function of x. Then we can compute the expected value of $g(X)$ using the formula we learned in class in the continuous random variable section:

$$\mathbb{E}(g(X)) = \int_a^b g(x)f(x)dx$$

Now let's take a sample of n random variables $X_1, \ldots, X_n$ from the density $f(x)$ and compute the sample mean of $g(X_1), \ldots, g(X_n)$, which we will call $\bar{g}_n$, where the subscript n denotes the number of samples we took.

$$\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

Taking the expected value of $\bar{g}$, we get:

$$\begin{aligned}\mathbb{E}(\bar{g}_n) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(g(X_i)) \\ &= \mathbb{E}(g(X))\end{aligned}$$

since all the X_i have the same distribution. Thus $\bar{g}_n$ is an unbiased estimator for the expected value $\mathbb{E}(g(X))$. This is called the *Monte Carlo estimator*. As long as the variance of $g(X)$ is finite[27], then this is a consistent estimator for $g(X)$. Thus by the Law of Large Numbers, the Monte Carlo estimator $\bar{g}_n$ converges to the expected value $\mathbb{E}(g(X))$ in probability as $n \to \infty$. We can use Monte Carlo techniques for numerical integration. We will give two Monte Carlo methods, the first based on a dartboard and second based on our Monte Carlo estimator which we defined above.

Suppose we want to integrate a function $f(x)$ from a to b. Let's throw n darts uniformly at the same rectangular dartboard we used above, and let Y be the number of darts which land under the curve. Then $Y \sim \text{Binomial}(n, p)$, where p is the probability of a dart landing under the curve. The probability that a dart lands under the curve is the ratio of the area of the curve to the area of the rectangle, i.e.

$$p = \frac{\int_a^b f(x)dx}{M(b-a)}$$

Let $\hat{p}_n = Y/n$ be the unbiased estimator for p, where the subscript n denotes the number of darts thrown. We know from the section on estimation that $\hat{p}_n$ is a consistent estimator for p, i.e. $\hat{p}_n$ converges to p (in probability) as $n \to \infty$. By linearity,

$$M(b-a)\hat{p}_n = M(b-a)\frac{Y}{n}$$

is an unbiased estimator for $\int_a^b f(x)dx$ and converges to $\int_a^b f(x)dx$ in probability as $n \to \infty$. Thus we can estimate the integral by $M(b-a)\frac{Y}{n}$. It is easy to compute this, since all we have to do is generate uniform points in the rectangle and verify that the y-coordinate lies under the curve.

Just as we did above in the case of sampling from probability distributions, Monte Carlo integration does not require that samples be chosen uniformly in a rectangle, i.e. we don't have to use a "dartboard" to do it. Why might it be useful to use a different approach? Suppose we want to integrate the standard normal density function from -5 to 5. If we enclose that in a rectangular dartboard and throw darts uniformly at random, many of the darts will lie in the region of the tails of the normal distribution where there is very little area. It makes sense that for a good estimate of the are under the curve, darts thrown near the center "matter" more than darts thrown to the edges since there is more area there. Thus it might be a good idea to throw darts whose x-coordinate distributed such that more darts hit the center than the edges of the dartboard. This idea is called *importance sampling*. The algorithm is based on our Monte Carlo Estimator for expected values.

Let $f(x)$ be a continuous function, and suppose we want to estimate the integral $\int_a^b f(x)dx$. Let $g(x)$ be any probability density function defined on $[a, b]$. Take n independent samples $X_1, \ldots, X_n$ using the density $g(x)$ (this can be done with inverse CDF, rejection sampling,

[27]It suffices for $\mathbb{E}(g(X)$ to be finite, but that is beyond the scope of this course.

or any other method). Then the following is an unbiased estimator for $\int_a^b f(x)dx$:

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)}{g(X_i)}$$

What is the expected value of this estimator? Since the X_i are distributed according to $g(x)$, using linearity of expectation and the formula for the expected value of a function of a random variable,

$$\begin{aligned}
\mathbb{E}(\hat{I}_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{f(X_i)}{g(X_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_a^b \frac{f(x)}{g(x)}g(x)dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_a^b f(x)dx \\
&= \int_a^b f(x)dx
\end{aligned}$$

Thus $\hat{I}_n$ is unbiased. By the Law of Large Numbers, as $n \to \infty$, the estimator $\hat{I}_n$ converges to the integral $\int_a^b f(x)dx$ (in probability). Thus $\hat{I}_n$ is a consistent estimator for the integral $\int_a^b f(x)dx$.

8.3 Linear Regression

Linear regression is one of the mathematical methods most often used by professional statisticians. Linear regression is an inferential procedure which is used when one random variable Y, called the dependent variable, has a mean which is a function of one or more nonrandom variables $x_1, \ldots, x_n$, called the independent variables. (The names "dependent" and "independent" are used in the scientific sense to indicate the roles the variables play, and do not imply anything about independence in the probabilistic sense.) What are some examples of this?

1. Y is the stopping distance for an automobile of a particular make and model. Y is a random variable, but its mean depends on the velocity x of the car before the brakes are applied.

2. Y is the elongation of a strip of metal subject to a stretching force. Y is a random variable whose mean depends on the applied force x_1 and the temperature x_2.

3. Y is the sale price of homes in a certain neighborhood over the past month. Y is a random variable whose mean depends, among other things, on x, the size of the home in square feet.

In this section we will only consider a single independent variable x. Consider a scatter-plot of n samples of the random variable Y corresponding to n values of the independent variable

x. We could, for example, plot home prices in a certain neighborhood versus the size of the home. Intuitively, you would expect that Y would roughly increase as x increases, since bigger homes are more likely to sell for more. A purely deterministic, linear model, which we can write as

$$Y = \beta_0 + \beta_1 x$$

cannot possibly fit the data, because that would imply that all the points lie exactly on a straight line. Instead, we will use a probabilistic model which says that the data lies along a straight line, but that there is random error involved so that the points do not lie exactly along the line. This model would look like:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is a random variable modeling the error in the problem. The distribution of ϵ is unknown, but we will take it to have mean 0 and variance σ_2. It is often useful (and in many cases quite accurate) to take ϵ to have a normal distribution. Since $\mathbb{E}(\epsilon) = 0$, the expected value of Y is

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x$$

since everything other than ϵ is a constant. Thus, although Y itself does not fall on a straight line, its expected value does.

A *linear statistical model* models the expected value $\mathbb{E}(Y)$ as a linear function of one or more unknown parameters $\beta_0, \beta_1, \ldots, \beta_n$. Note that the word *linear* refers to the fact that $\mathbb{E}(Y)$ is a linear function of the parameters β_i. $\mathbb{E}(Y)$ is not necessarily a linear function of the independent variables. If we like, to make this clear we could write the model as $\mathbb{E}(Y) = x\beta_1 + \beta_0$, to show that the independent variable plays the role of the coefficient here. We shall consider here only *simple linear regression* models, ones in which $\mathbb{E}(Y)$ is a linear function of only two unknown parameters β_0 and β_1. The following are all examples of simple linear regression models:

$$\begin{aligned}
\mathbb{E}(Y) &= \beta_0 + \beta_1 x \\
\mathbb{E}(Y) &= \beta_0 + \beta_1 x^2 \\
\mathbb{E}(Y) &= \beta_0 + \beta_1 \sqrt{x} \\
\mathbb{E}(Y) &= \beta_0 + \beta_1 \log(x)
\end{aligned}$$

In all the above cases, $\mathbb{E}(Y)$ is a linear function of the unknown parameters. The thing involving the independent variable x is merely the coefficient of β_1. We can do whatever we want (take a power, root, log, exponential, etc) to the independent variable x since its values are always known. A *multiple linear regression model* is still a linear function of unknown parameters, but we have more than two unknown parameters. The following are examples of multiple linear regression models with three unknown parameters:

$$\begin{aligned}
\mathbb{E}(Y) &= \beta_0 + \beta_1 x_1 + \beta_2 x^2 \\
\mathbb{E}(Y) &= \beta_0 + \beta_1 x + \beta_2 x^2
\end{aligned}$$

Recall that the independent variables are the coefficients of the linear model. In the first case, the coefficients are x and x^2, which are two powers of the same independent variable. In the second case, the coefficients are two different independent variables. We will only concern ourselves with simple linear regression models here.

The parameters β_0 and β_1 are population parameters which are unknown. We will use the method of least squares to estimate these parameters. Roughly, this is what we are doing. Y represents a measurement from a population whose distribution is unknown, but whose expected value depends on an independent, nonrandom variable x via a simple linear regression relationship $\mathbb{E}(Y) = \beta_0 + \beta_1 x$ with unknown parameters β_0 and β_1. We will estimate β_0 and β_1 in the following way. Take n samples $Y_1, \ldots, Y_n$ from the population, along with their corresponding values of $x_1, \ldots, x_n$. In our real estate example, we select n homes. Their sale price corresponds to the Y_i, and their size corresponds to the x_i. In the automobile braking example, we select n different velocities x_i and measure the braking distance Y_i corresponding to each one. In all cases, we have n ordered pairs (x_i, Y_i). Plot these on a graph, and draw the "best-fit" straight line (you can just eyeball this!) The y-intercept and slope of this "best-fit" line are our estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the population parameters β_0 and β_1.

Our estimator for the unknown parameters is nothing more than the "best-fit" line, which we are familiar with from high school science class! Since we want to quantify how good the estimators are, the "eyeball method" is not good enough. We need a method to mathematically construct the best-fit straight line through a set of points. This method is known as the *method of least squares*, and is the method used by Microsoft Excel and other software packages to do this.

8.3.1 Method of Least Squares

Suppose we have n points (x_i, y_i), and we wish to fit the best possible straight line to them. Since this is an estimator, we will denote the best possible straight line by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

In the "eyeball method", we want to make the line as close as possible to all the points. Mathematically, here's what we do. The x-value x_i corresponds to the observed y-value y_i. If we plug x_i into our linear model, we get our estimated y-value:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

For x_i, the deviation of the observed value from the estimated value, called the *error*, is $y_i - \hat{y}_i$. We want to minimize the *sum of square errors* (SSE) of the n samples. That is, we want to find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$SSE = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=0}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

To do this requires multivariable calculus. (We could also do the minimization numerically, using the method of gradient descent). Essentially, we take the partial derivatives with

respect to both $\hat{\beta}_0$ and $\hat{\beta}_1$, set them equal to 0, and solve the resulting equations for the two parameters. The algebra is lengthy and so the details will be omitted; if you like you can work through them or look up the calculations in any book on statistics. The result is presented in the box below.

Least Squares Estimators for Simple Linear Regression

Suppose that a random variable Y and a nonrandom variable x are related by a linear regression model:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x$$

Take n samples $(x_1, y_1), \ldots, (x_n, y_n)$. Then the least squares estimators of the parameters β_0 and β_1 are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the two variables, and

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Calculating these estimators is sufficiently annoying that it should only be done on a computer! If you have studied linear algebra, there is actually a nice form for these involving matrix multiplication. Instead of doing an example (by hand) of a least squares regression, we will instead discuss properties of these estimators such as bias and variance.

8.3.2 Properties of Least Squares Estimators

Recall the model that we are using for simple linear regression. Y is a random variable whose value depends on an independent variable x with a degree of error given by ϵ.

$$Y = \beta_0 + \beta_1 x + \epsilon$$

The error ϵ has mean 0 and variance σ^2. In particular, we assume that the variance of the error does not depend on the independent variable x. Thus the expected value of Y is a linear function of x:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x$$

We will show that the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of the parameters β_0 and β_1. To form our estimators, recall that we took n samples $Y_1, \ldots, Y_n$ from

the population. To each sample we have a corresponding value for the independent variable $x_1, \ldots, x_n$. The dependent variable Y_i and the independent variable x_i are related according to our model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i is the value of the error for the ith sample (x_i, Y_i). The errors ϵ_i are independent and have the same distribution as ϵ. Since the expected value of ϵ_i is 0, note that

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i$$

Since $\beta_0 + \beta_1 x_i$ is a constant and thus adding it does not affect variance, the variance of Y_i is given by

$$Var(Y_i) = Var(\epsilon_i) = \sigma^2$$

First let's find the bias of $\hat{\beta}_1$. We can simplify the expression for $\hat{\beta}_1$ as follows[28]:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
&= \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}} \\
&= \frac{\sum(x_i - \bar{x})Y_i - \bar{Y}\sum(x_i - \bar{x})}{S_{xx}} \\
&= \frac{\sum(x_i - \bar{x})Y_i}{S_{xx}}
\end{aligned}$$

since $\sum(x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$. Using linearity of expectation, we take the expected value of $\hat{\beta}_1$:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \mathbb{E}\left[\frac{\sum(x_i - \bar{x})Y_i}{S_{xx}}\right] \\
&= \frac{\sum(x_i - \bar{x})\mathbb{E}(Y_i)}{S_{xx}} \\
&= \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} \\
&= \beta_0 \frac{\sum(x_i - \bar{x})}{S_{xx}} + \beta_1 \frac{\sum(x_i - \bar{x})x_i}{S_{xx}}
\end{aligned}$$

Note that $\sum(x_i - \bar{x}) = 0$ and

$$\begin{aligned}
S_{xx} &= \sum(x_i - \bar{x})^2 \\
&= \sum(x_i^2 - x_i\bar{x} - x_i\bar{x} + \bar{x}^2) \\
&= \sum(x_i - \bar{x})x_i - \bar{x}\sum(x_i - \bar{x}) \\
&= \sum(x_i - \bar{x})x_i
\end{aligned}$$

[28]Since all sums are from 1 to n, we will omit the indexes on the sum for simplicity.

since the second sum in the second-to-last line is 0. Plugging these into the expression for the expected value of $\hat{\beta}_1$, we obtain:

$$\mathbb{E}(\hat{\beta}_1) = 0 + \beta_1 \frac{S_{xx}}{S_{xx}} = \beta_1$$

Thus $\hat{\beta}_1$ is an unbiased estimator for β_1. To compute its variance, we use the formula for the variance of a sum of independent random variables, together with the facts that a constant is squared when it is pulled out.

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left[\frac{\sum(x_i - \bar{x})Y_i}{S_{xx}}\right] \\ &= \frac{1}{S_{xx}^2}\sum Var\left[(x_i - \bar{x})Y_i\right] \\ &= \frac{1}{S_{xx}^2}\sum (x_i - \bar{x})^2 Var(Y_i) \end{aligned}$$

Since $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and the variance of ϵ_i is σ^2, $Var(Y_i) = \sigma^2$, since $\beta_0 + \beta_1 x_i$ is constant thus does not affect the variance. Putting all this together, we have:

$$\begin{aligned} Var(\hat{\beta}_1) &= \frac{1}{S_{xx}^2}\sigma^2 \sum (x_i - \bar{x})^2 \\ &= \frac{S_{xx}}{S_{xx}^2}\sigma^2 \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Since $\hat{\beta}_1$ is unbiased, its mean squared error (MSE) is equal to its variance.

Let's do the exact same thing for the other estimator $\hat{\beta}_0$. Recall that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$. By linearity of expectation,

$$\begin{aligned} \mathbb{E}(\hat{\beta}_0) &= \mathbb{E}(\bar{Y} - \hat{\beta}_1 \bar{x}) \\ &= \mathbb{E}(\bar{Y}) - \bar{x}\mathbb{E}(\hat{\beta}_1) \\ &= \mathbb{E}(\bar{Y}) - \beta_1 \bar{x} \end{aligned}$$

where we used the result from above. All we need to do is compute the expected value of $\bar{Y}$. Since our model is more complicated than in the one in the sampling distribution section (i.e. Y depends on the independent random variable x as well as the random error ϵ), this expected value is not just the population mean. By linearity of expectation, using the expected value of Y_i which we found earlier,

$$\begin{aligned} \mathbb{E}(\bar{Y}) &= \frac{1}{n}\sum \mathbb{E}(Y_i) \\ &= \frac{1}{n}\sum (\beta_0 + \beta_1 x_i) \\ &= \beta_0 + \beta_1 \frac{1}{n}\sum x_i \\ &= \beta_0 + \beta_1 \bar{x} \end{aligned}$$

Plugging this in above, we get:

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

Thus $\hat{\beta}_0$ is an unbiased estimator for β_0.

Now we find its variance. This is a little trickier since $\bar{Y}$ and $\hat{\beta}_1$ are not necessarily independent. Recalling that the general form for the variance of a sum involves the covariance,

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{Y} - \hat{\beta}_1\bar{x}) \\ &= Var(\bar{Y}) + Var(-\bar{x}\hat{\beta}_1) + 2Cov(\bar{Y}, -\bar{x}\hat{\beta}_1) \\ &= Var(\bar{Y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}Cov(\bar{Y}, \hat{\beta}_1) \end{aligned}$$

where we used the fact that constants are squared when they are pulled out of the variance, but are pulled out without alteration from the covariance. We computed the variance of $\hat{\beta}_1$ above. Since the samples Y_i are independent, the variance of $\bar{Y}$ is given by:

$$Var(\bar{Y}) = \frac{1}{n^2}\sum Var(Y_i) = \frac{\sigma^2}{n}$$

where we used the fact that $Var(Y_i) = \sigma^2$, which we derived above. Finally we compute the covariance of $\bar{Y}$ and $\hat{\beta}_1$. Here we use the fact that the covariance is linear in each argument, i.e.

$$Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$$

This can be shown using the definition of covariance or using the Magic Covariance Formula. Thus we have:

$$\begin{aligned} Cov(\bar{Y}, \hat{\beta}_1) &= Cov\left(\frac{1}{n}\sum Y_i, \frac{1}{S_{xx}}\sum (x_i - \bar{x})Y_i\right) \\ &= \frac{1}{nS_{xx}}\sum_i (x_i - \bar{x})Cov(Y_i, Y_i) + \frac{1}{nS_{xx}}\sum_i\sum_{j\neq i}(x_j - \bar{x})Cov(Y_i, Y_j) \\ &= \frac{1}{nS_{xx}}\sum_i (x_i - \bar{x})Var(Y_i) \end{aligned}$$

since $Cov(Y_i, Y_j) = 0$ for $i \neq j$ by independence of the samples Y_i, and $Cov(Y_i, Y_i) = Var(Y_i)$. Since we know that the variance of Y_i is σ^2, this becomes:

$$\begin{aligned} Cov(\bar{Y}, \hat{\beta}_1) &= \frac{1}{nS_{xx}}\sum_i (x_i - \bar{x})Var(Y_i) \\ &= \frac{\sigma^2}{nS_{xx}}\sum_i (x_i - \bar{x}) = 0 \end{aligned}$$

So the covariance of $\bar{Y}$ and $\hat{\beta}_1$ is 0, which is really convenient! Note that this does *not* imply that $\bar{Y}$ and $\hat{\beta}_1$ are independent. Now we have everything we need to compute the variance of $\hat{\beta}_0$.

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{Y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}Cov(\bar{Y}, \hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} - 0 \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \end{aligned}$$

8.3.3 Estimation of the Variance of the Error

The expressions for the variances of the least squares estimators are both in terms of σ^2, the variance of the error term ϵ. In almost every case, this variance is unknown, thus we will use the samples themselves to estimate σ^2. Recall that if we use $\bar{Y}$ as an estimator of the population mean, we showed that the estimator

$$\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

is an unbiased estimator for the population variance. Things are little more complicated here because of the dependence on x. Recall that for each sample Y_i we are estimating the mean $\mathbb{E}(Y_i)$ with the estimator

$$\hat{(Y_i)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Thus it seems reasonable to suppose that an estimator for the variance σ^2 of the error is based on the sum of squares error $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$. In fact, the estimator

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \frac{1}{n-2} SSE$$

is an unbiased estimator for σ^2. Note that the 2 in the denominator of S^2 is the same as the number of parameters β_i which occur in the model. Since by linearity

$$\mathbb{E}(S^2) = \frac{1}{n} \mathbb{E}(SSE)$$

it suffices to find the expected value of the SSE.

$$\begin{aligned} \mathbb{E}(SSE) &= \mathbb{E}\left[\sum (Y_i - \hat{Y}_i)^2 \right] \\ &= \mathbb{E}\left[\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right] \\ &= \mathbb{E}\left[\sum (Y_i - \bar{Y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 \right] \end{aligned}$$

where we used the fact that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ (this is the definition of $\hat{\beta}_0$). Thus we have:

$$\begin{aligned}
\mathbb{E}(SSE) &= \mathbb{E}\left[\sum [(Y_i - \bar{Y}) + \hat{\beta}_1(\bar{x} - x_i)]^2\right] \\
&= \mathbb{E}\left[\sum (Y_i - \bar{Y}) - 2\hat{\beta}_1 \sum (Y_i - \bar{Y})(\bar{x} - x_i) + \hat{\beta}_1^2 \sum (\bar{x} - x_i)^2\right] \\
&= \mathbb{E}\left[\sum (Y_i - \bar{Y}) - 2\hat{\beta}_1 \sum (Y_i - \bar{Y})(\bar{x} - x_i) + \hat{\beta}_1^2 S_{xx}\right]
\end{aligned}$$

Since we have:

$$\sum (Y_i - \bar{Y})(\bar{x} - x_i) = S_{xy} = \hat{\beta}_1 S_{xx}$$

and

$$\begin{aligned}
\sum (Y_i - \bar{Y}))^2 &= \sum (Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2) \\
&= \sum (Y_i)^2 - 2\bar{Y} \sum Y_i + n\bar{Y}^2 \\
&= \sum (Y_i)^2 - 2n\bar{Y}^2 + n\bar{Y}^2 \\
&= \sum (Y_i)^2 - n\bar{Y}^2
\end{aligned}$$

we have:

$$\begin{aligned}
\mathbb{E}(SSE) &= \mathbb{E}\left[\sum (Y_i)^2 - n\bar{Y}^2 - 2\hat{\beta}_1^2 S_{xx} + \hat{\beta}_1^2 S_{xx}\right] \\
&= \mathbb{E}\left[\sum (Y_i)^2 - n\bar{Y}^2 - \hat{\beta}_1^2 S_{xx}\right] \\
&= \sum \mathbb{E}(Y_i^2) - n\mathbb{E}(\bar{Y}^2) - S_{xx}\mathbb{E}(\hat{\beta}_1^2)
\end{aligned}$$

We know the last term from above. For the first two terms, recall that for any random variable X, since $Var(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$ by the Magic Variance Formula, $\mathbb{E}(X^2) = Var(X) + [\mathbb{E}(X)]^2$. Thus we have:

$$\begin{aligned}
\mathbb{E}(Y_i^2) &= Var(Y_i) + [\mathbb{E}(Y_i)]^2 \\
&= \sigma^2 + (\beta_0 + \beta_1 x_i)^2
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(\bar{Y}^2) &= Var(\bar{Y}) + [\mathbb{E}(\bar{Y})]^2 \\
&= \frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(\hat{\beta_1}^2) &= Var(\hat{\beta}_1) + [\mathbb{E}(\hat{\beta}_1)]^2 \\
&= \frac{\sigma^2}{S_{xx}} + \beta_1^2
\end{aligned}$$

Plugging both of these in, we get:

$$\begin{aligned}\mathbb{E}(SSE) &= \sum(\sigma^2 + (\beta_0 + \beta_1 x_i)^2) - n\left(\frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2\right) - S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) \\ &= n\sigma^2 + \sum(\beta_0 + \beta_1 x_i)^2 - \sigma^2 - n(\beta_0 + \beta_1 \bar{x})^2 - \sigma^2 - \beta_1^2 S_{xx} \\ &= (n-2)\sigma^2\end{aligned}$$

where in the last line, everything else cancels if we expand all the squares, evaluate the sum, and simplify. Thus we conclude that S^2 is indeed an unbiased estimator for the variance σ^2 of the error term ϵ.

Now that we have unbiased estimators for β_0 and β_1 and know (or know how to estimate) their variance, we can use them to construct confidence intervals and hypothesis tests for the β_0 and β_1. It is reasonable (and useful) to assume that the error ϵ is normally distributed. This is beyond the scope of this course, but these ideas (and more!) can be found in a book (or further course) on statistics.

8.4 Bayesian Statistics

The final topic we will discuss in this class is Bayesian statistics. We will only have time for a brief introduction, since this could be the subject of an entire course. Since Bayesian statistics is currently very popular (and since Nate Silver's election prediction methodology relies heavily on Bayesian ideas), it is worth taking a look at what it's all about.

8.4.1 Introduction

Bayesian statistics is a fundamentally different approach to statistics from the frequentist philosophy. In the frequentist approach, which we have used thus far in this course, we consider a population to described by a probability distribution with one or more unknown parameters. These parameters have "true" values which we do not know, but which we can estimate by taking samples from the population and computing estimators such as the sample mean. The more samples we take, the closer our estimate is to the "true" parameter value.

In the Bayesian approach, these population parameters do not have "true" values but are themselves distributed according to a probability distribution. We start with a "best guess" as to the distribution of the population parameter (the *prior distribution*). We then take samples from our population and use Bayes' theorem to "update" the population distribution to obtain a *posterior distribution*. We can repeat this process as many times as we want. The idea is that each time we collect more data, we learn more about the probability distribution of the parameter of interest, and

Let's see how this works mathematically. First, recall Bayes' theorem for two events A and B:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

The denominator is typically unknown, so we can expand it using the Law of Total Probability, using any partition of the probability space. The partition $\{A, A^c\}$ is typically used, giving us the following form of Bayes' theorem.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}$$

We can rewrite this as follows:

$$\mathbb{P}(A|B) = \frac{1}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}\mathbb{P}(B|A)\mathbb{P}(A)$$

Suppose A is the event we are interested in studying. The prior probability of A is $\mathbb{P}(A)$, the probability of A occurring before we know any additional information. Suppose we then observe that B occurs. We are interested in updating the probability that A occurs now that we know this new information. In other words, we want to know $\mathbb{P}(A|B)$. The term $\mathbb{P}(B|A)$ above is the probability that our new event B occurs given A occurs; this is the likelihood of B given A. The fraction in front on the right hand side is just a constant. Thus we can write:

$$\mathbb{P}(A|B) \propto \mathbb{P}(B|A)\mathbb{P}(A)$$

This proportion captures the essence of Bayesian statistics: *the posterior distribution is proportional to the product of the prior distribution and the likelihood*. Let's now rework this in terms of probability density functions and population parameters.

Suppose we have a population whose distribution is parameterized by θ (as before, θ could be a population mean, proportion, variance, or some other parameter). We will write then density function for the population as $f_\theta(y)$. (The population could be discrete, in which case we have a pmf $p_\theta(y)$). Since we are now Bayesian statisticians, there is no true value for θ. Rather, we assume that θ is distributed according to some probability density function which we will call $g(\theta)$[29].

Now we take n independent samples $Y_1, \ldots, Y_n$ from the population. We form the likelihood function of our data like we have in previous sections:

$$L(Y_1, \ldots, Y_n|\theta) = \prod_{i=1}^{n} f_\theta(Y_i)$$

If the population is described by a pmf $p\theta(y)$, the likelihood function is the joint probability of our data. If the population is described by a density $p\theta(y)$, the likelihood function is not quite a probability (since the probability of a point is 0), but we will treat it as if it were one. The posterior density is the conditional density of the parameter θ given our samples $Y_1, \ldots, Y_n$, which we will denote:

$$g(\theta|Y_1, \ldots, Y_n)$$

[29]This could be a pmf or a density; for the purposes of our discussion, we will assume that the population parameters are continuous random variables, so are described by a probability density function.

Using Bayes' theorem with these densities (pretending that densities are probabilities), we get:

$$\begin{aligned} g(\theta|Y_1,\ldots,Y_n) &= \frac{\mathbb{P}(Y_1,\ldots,Y_n|\theta)g(\theta)}{\mathbb{P}(Y_1,\ldots,Y_n)} \\ &= \frac{L(Y_1,\ldots,Y_n|\theta)g(\theta)}{\int L(Y_1,\ldots,Y_n|\theta)g(\theta)d\theta} \end{aligned}$$

where the integral in the denominator is taken over all possible values of θ. Since the denominator is a constant, we get the same proportionality:

$$g(\theta|Y_1,\ldots,Y_n) \propto L(Y_1,\ldots,Y_n|\theta)g(\theta)$$

In other words, the posterior density is proportional to the product of the prior density and the likelihood of the data. Let's think about how we could use this in a real-world example.

Example. You are a pollster interested in the proportion p of voters in Rhode Island who plan on voting for Gina Raimondo. If you were a frequentist statistician, you would take a sample of, say, 100 voters and use $\hat{p}$ the proportion of Raimondo supporters in your sample, to estimate p. This time, we are a Bayesian statistician. First we need to choose a prior distribution for p. Much of the art of Bayesian statistics involves what to choose for the prior distribution. One choice (sometimes called the naive choice) is that since we have no information at all, we could choose a Uniform[0, 1] distribution for p as our prior. Now we poll voters from our population. Using the poll data and Bayes' theorem (the version above), we can update the probability distribution for p using our poll data. This posterior distribution for p does not give us an exact value for p (it is still a probability distribution), but since it will not be uniform, it will give us a better understanding of the proportion p of voters who prefer Raimondo. We can keep taking polls, updating our probability distribution for p each time, to get a better and better sense of p

8.4.2 Conjugate Distributions

In some cases, there is nice relationship between the prior and the posterior distributions (this is not in general the case). This will depend on the distribution of underlying population and what is chosen for the prior distribution. For a given population distribution parameterized by θ, if the posterior probability distribution $g(\theta|Y_1,\ldots,Y_n)$ and the prior probability distribution $g(\theta)$ will belong to the same family, we say that the prior and posterior are *conjugate distributions*. We will look at two examples of this.

1. Binomial population

   You are a pollster interested in the proportion p of voters in Rhode Island who plan on voting for Gina Raimondo. Take a sample of size n from the population. Let Y be

the number of voters in your sample you prefer Raimondo. Then $Y \sim \text{Binomial}(n, p)$. Recall the pmf for Y is:

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y} \qquad\qquad y = 0, 1, \ldots, n$$

Let's express this as a function of p. It takes the form

$$g(p) = K p^a (1-p)^b$$

where a and b are constants related to the distribution of p, and K is a normalizing constant. This is not a probability distribution for p, but we could turn it into a probability distribution on $[0, 1]$ by choosing the normalizing constant K so that $g(p)$ integrates to 1 over $[0, 1]$. The usual conjugate prior for a binomial population is the beta distribution, which is a slightly altered version of this:

$$g(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$$

where α and β are chosen to reflect your existing belief of the distribution of p, and $B(\alpha, \beta)$ is a normalizing constant chosen so that $g(p)$ integrates to 1 over $[0, 1]$. The values α and β are called *hyperparameters* of the prior distribution, to distinguish them from p, which is a parameter of the population. Note that in the special case where $\alpha = 1$ and $\beta = 1$, we have $g(p) = 1$, which is the uniform distribution on $[0, 1]$. If Y is a beta random variable with parameters α and β, then the mean of Y is:

$$\mathbb{E}(Y) = \frac{\alpha}{\alpha + \beta}$$

Let's use Bayes' theorem (as above) to find the posterior distribution of p given our binomial data Y from our n samples and a prior distribution of Beta$[\alpha, \beta]$. Bayes' theorem says that:

$$g(p|Y) = \frac{L(Y|p)g(p)}{\int L(Y|p)g(p)dp}$$

Note that we write our data as Y instead of $Y_1, \ldots, Y_n$ since Y is a binomial random variable incorporating data from n samples. Let's look at each of these pieces in turn. For the likelihood of the data:

$$L(Y|p) = \binom{n}{Y} p^Y (1-p)^{n-Y}$$

The prior distribution for p is the beta distribution given above. Multiplying this by the likelihood, we get:

$$\begin{aligned} L(Y|p)g(p) &= \binom{n}{Y} p^Y (1-p)^{n-Y} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} \\ &= \binom{n}{Y} \frac{1}{B(\alpha, \beta)} p^{\alpha+Y-1}(1-p)^{\beta+(n-Y)-1} \end{aligned}$$

Putting this all together:

$$\begin{aligned} g(p|Y) &= \frac{\binom{n}{Y}\frac{1}{B(\alpha,\beta)}p^{\alpha+Y-1}(1-p)^{\beta+(n-Y)-1}}{\int_0^1 \binom{n}{Y}\frac{1}{B(\alpha,\beta)}p^{\alpha+Y-1}(1-p)^{\beta+(n-Y)-1}dp} \\ &= \frac{p^{\alpha+Y-1}(1-p)^{\beta+(n-Y)-1}}{\int_0^1 p^{\alpha+Y-1}(1-p)^{\beta+(n-Y)-1}dp} \\ &= \frac{p^{\alpha+Y-1}(1-p)^{\beta+(n-Y)-1}}{B(\alpha+Y-1,\beta+(n-Y)-1)} \end{aligned}$$

Where the last line follows since the denominator is the integral of the numerator, so is the normalizing constant. Comparing this to the beta density, we see that the posterior distribution is also a beta distribution, with hyperparameters $\alpha + Y$ and $\beta + (n - Y)$ (The 1s are not part of the hyperparameters; look back at the density for the beta distribution to see why this is the case). Thus we have updated our prior based on our data and obtained a posterior in the same family as the prior but with updated hyperparameters. Essentially, we have added the number of "sucesses" (Raimondo supporters) to α and the number of "failures" (non-Raimondo supporters) to β.

We can use the mean of the posterior distribution as an estimate for p. We call this the Bayes estimator for p, denoted $\hat{p}_B$. The Bayes estimator is given by:

$$\hat{p}_B = \frac{\alpha + Y}{\alpha + Y + \beta + (n - Y)} = \frac{\alpha + Y}{\alpha + \beta + n}$$

We can write this estimator in a slightly different form:

$$\begin{aligned} \hat{p}_B &= \frac{\alpha}{\alpha+\beta+n} + \frac{Y}{\alpha + Y + \beta + (n-Y)} \\ &= \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta} + \left(\frac{n}{\alpha+\beta+n}\right)\frac{Y}{n} \end{aligned}$$

This is a weighted average of mean of the beta prior and the sample proportion Y/n (the MLE for p). The prior mean is given less weight for larger sample sizes, whereas the weight of the sample proportion increases as the sample size gets larger. Since $\mathbb{E}(Y/n) = p$, the Bayes estimator is not an unbiased estimator for p. In most cases, Bayes estimators are not unbiased.

2. Normal population

This time, suppose we have a normal population with unknown mean μ and *known* variance σ_0^2. If we choose a normal distribution as a prior for the population parameter μ, then the posterior distribution for μ will also be normal. Mathematically, here's how this goes.

Suppose we choose as our prior for the parameter μ a normal distribution with mean η and variance δ^2. That is, our prior is Normal(η, δ). Let $Y_1, \ldots, Y_n$ be a random sample from the population, and compute the sample mean $\bar{Y}$. Then the posterior distribution for μ is also a normal distribution with mean η^* and variance δ^* given by:

$$\eta^* = \frac{\delta^2 n\bar{Y} + \sigma_0^2 \eta}{n\delta^2 + \sigma_0^2}$$
$$\delta^{*2} = \frac{\sigma_0^2 \delta^2}{n\delta^2 + \sigma_0^2}$$

The Bayes estimator $\hat{\mu}_B$ is the mean of the normal posterior distribution. We can separate the sum in the numerator to write it as:

$$\hat{\mu}_B = \frac{n\delta^2}{n\delta^2 + \sigma_0^2}\bar{Y} + \frac{\sigma_0^2}{n\delta^2 + \sigma_0^2}\eta$$

Thus the Bayes estimator for the population mean is a weighted average of the sample mean $\bar{Y}$ (the MLE for μ) and the mean of the prior η. Again, as the sample size n increases, the weight given to $\bar{Y}$ increases, while the weight given to the prior mean η decreases.