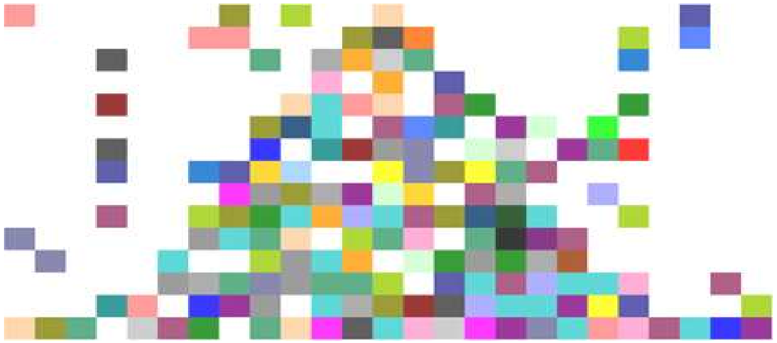
Probability and Stochastic Processes

by
William O. Ray



William Ray, Tulsa Oklahoma, Spring 2017

Table of Contents

1. Introduction

Herodotus

The history of the study of probability is inextricably interwoven with the study of games of chance. Many of the earliest attempts to model games of chance were motivated by the mundane desire to 'beat the house' and develop a winning strategy for games in casinos. Conversely, the casinos were equally motivated to devise games that guaranteed – in the long run – that the casino would make money. These origins sometimes leave one with the impression that the study of probability is the study of various kinds of more or less dissolute behaviors. Even modern examples often have names like 'drunkard's walk' as though mathematicians have taken a kind of perverse delight in unsavory applications of the theory. However much truth there might be in this, it is true that games of chance provide a set of examples that are easy to understand and explain. As with most of mathematics, tools developed to solve one set of problems often turn out to solve many kinds of problems. Problems as apparently different as 'balls in urns,' network switches, Mendelian genetics and sorting algorithms in a computer program often turn out to have mathematical similarities even though the contexts appear very different.

As applications go, games of chance are an especially appealing class of problems for introducing probabilistic concepts since they have simple and easily understood rules. The same mathematical model that describes, say, the arrival of calls in a telephony network, might also describe a simple casino game. The latter has the advantage that one needs minimal technical knowledge to grasp the problem. By way of example, consider dice games.

According to the Greek historian Herodotus dice were invented in the fifth century BCE by the Lydians of Asia Minor. However, dice at least 2000 years older have been found in Egyptian ruins and there is some evidence that dice are as much as 6000 years old. The Greeks and Romans used the familiar modern cubical dice with spots marking the different sides, but also used animal bones such as sheep ankle bones. The four-sided anklebones were called *astaralagi* and the more modern six-sided spotted dice *tesserae* . Many dice games involve pairs of dice. Both of the Latin words are plural, as is the English word *dice* the plural of *die* for a singled spotted cube.

1.1. Example.

Suppose that two fair die are rolled and the player wins the game if the sum of the spots is "seven" or "eleven." What are the chances that the player wins?

Solution. To answer this question we need to consider what the possible outcomes, or rolls of the dice, might look like. Since there are two dice, we could think of them as being rolled first one, then the other. If the die are rolled sequentially then the rolls of the dice consist of thirty-six different possibilities:

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

The rolls that sum to seven are the six on the main diagonal. There are exactly two rolls that sum to eleven, so there are a total of eight of the 36 rolls that sum to either seven or eleven. Thus the probability of winning this game is eight out of thirty-six or $\frac{2}{9}$.

■

One might ask why we can think of the two die as being rolled sequentially? After all, in the context of our game the rolls $(3, 4)$ and $(4, 3)$ are the "same" since the sum is seven in both cases. In terms of winning or losing, the order in which the die are rolled does not matter; indeed, we would expect that the player would most likely roll the dice simultaneously. Following this reasoning, it would seem that an equally valid mental picture would be

1&1	1&2	1&3	1&4	1&5	1&6
-	2&2	2&3	2&4	2&5	2&6
-	-	3&3	3&4	3&5	3&6
-	-	-	4&4	4&5	4&6
-	-	-	-	5&5	5&6
-	-	-	-	-	6&6

This table lists the *distinguishable* outcomes of our game and the order of the rolls is not recorded. Is there some reason to prefer one table over the other for our mental picture of the outcomes?

The basic answer is that both are equally valid, but one is more *useful* in terms of calculating probabilities. In the first table, all of the listed outcomes are *equally likely* while in the second table some (those on the diagonal) are less likely than others. Certainly if the die are thrown sequentially – first one, then the other – then the first table of ordered pairs provides a correct thought-picture of 36 equally likely outcomes. Imagine, if you will, a game in which the player may either throw the dice simultaneously or sequentially, but does so in a locked chamber where the player cannot be observed rolling the dice. After the dice are thrown, you enter the room and observe the results. Intuitively, the chances of winning should be same regardless of how the dice are thrown. Further the results *look* the same in this imaginary scenario regardless of how the dice were thrown. Thus since the first table gives the correct thought-picture for sequential throws then it must also give the correct thought-picture for simultaneous throws. After all, the die don't "know" if they were thrown sequentially or not!

One could also carry out an experiment to see if the above reasoning can be empirically supported. For example, it would be easy to roll a pair of dice a large number of times (say 500 times). If the rolls are simultaneous, then we could write in the second table the number of times we observed each outcome. If the *model* from the first table is correct, one would expect to see around 14 observations ($500 \div 36$) in each of the diagonal cells and about 28 each the remaining cells. On the other hand, if the 21 outcomes in the second table were "equally likely," then one would expect to see about 23 observations ($500 \div 21$) in each of the 21 cells. Since this experiment is random, being based on the roll of dice, you of course won't get exactly the expected results. But your results will be far closer to the first model than to the second. Thus empirical evidence also supports the reasoning underlying "36 equally likely outcomes." (See the exercises.)

In the preceding example we found the *probability* of winning the game. Thus if the game were played repeatedly, say 900 times, we would expect to win about 200 of the games and to lose about 700 of the games. The *odds* of winning are then

$$\frac{200}{700}$$

or two sevenths, since we win the game 2 times for every seven times we lose game. More generally if the chances of winning a game are p then the *odds* of winning are

$$\frac{\text{chances of winning}}{\text{chances of losing}} = \frac{p}{1-p}.$$

Alternatively, if we know that the odds of winning are x then the probability p of winning is

$$p = \frac{x}{1+x}.$$

The payoffs for games of chance are sometimes phrased in terms of "odds." For example, in the above game if a one dollar bet pays off at odds of "two to seven" then the game is "fair" and the house and player will both break even over the long run. If, on the other hand, the game pays off at lower odds than break-even, then the house will always make a profit.

1.2. Example.

A simple dice game is called over and under seven. In this game the casino rolls a pair of dice and players bet on the outcome. The pay out is usually

sum of the spots is larger than seven	*casino pays off at even odds*
sum of the spots is less than seven	*casino pays off at even odds*
seven	*payoff is four times the bet*

If the sum of the spots is not seven this is called evens. Notice if a player bets on evens, then the best possible outcome is to lose most of the time while some of the time – when the sum of the spots is seven – the player will win four times the bet. If a player bets ”over seven” or ”under seven” then the player can at best break even and never win. On the other hand, the casino makes money from some betters regardless of the roll of the dice. If the sum of the spots is larger than seven, then the casino keeps all of the ”over seven” and ”evens” bets. If the sum of the spots is less than seven, then the casino keeps all of the ”under seven” and ”evens” bets. If the sum of the spots is exactly seven, then the casino keeps both the ”over seven” and ”under seven” bets.

The question here is whether or not the casino will make money on this game, regardless of the strategies followed by the players.

Solution. To answer this question we need to consider what the possible outcomes, or rolls of the dice, might look like. Since there are two dice, we can think of them as being rolled first one, then the other. Thus the rolls of the dice consist of thirty-six different possibilities:

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

It is exactly the six outcomes on the ascending diagonal that sum to seven:

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

Now this game has three possible outcomes: under seven, seven, and over seven.

Under seven is exactly the following rolls of the dice:

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

for a total of fifteen different rolls. On these rolls, the bettor retains the bet and the casino has no pay out. However, on the other twenty-one rolls the casino keeps the bet. Thus, while the casino pays off at even odds on an "under seven" bet, the casino actually "wins" an "under seven" bet 58% of the time:

$$\mathbf{0.5833 = \frac{21}{36}}$$

Similarly, the casino wins an "over seven" bet 58% of the time, again paying out at even odds. While both the "over seven" and "under seven" bets appear to be break-even, they in fact favor the casino.

What about the player who consistently bets "evens?" In this case, the casino wins the bet five times more often than it loses, while paying out at odds of only four to one. To see this, recall that there are exactly six ways that a roll of dice can total seven out of thirty-six possible rolls. Thus the casino wins a bet of "evens" five-sixths of the time and loses only one-sixth of the time, which implies that the casino has five-to-one odds (five-sixths being five times one-sixth) of winning the bet. Since the casino only pays at four-to-one odds, the game favors the casino even if the bet is "evens."

To approach the "evens" case another way, consider the player who bets on seven consistently. Such a player will win one sixth of the time. Thus in 600 one-dollar bets on seven, the pay outs and profits for the casino ought to look something like:

keeps the bet 500 times	makes \$500
pays on the winning seven 100 times	pays out \$400

Notice that the casino would expect to profit about \$100 in 600 bets of seven. If the casino paid sevens off at five to one rather than four to one, then the casino would exactly break even on this game.

1.3. Example.

Another simple dice game is sixes bet. In this game, the player has four tries to roll a six. Usually the bet is offered at even money, i.e., if a six is rolled in four tries the player wins \$1, otherwise the player loses \$1. In repeated plays of this game would you expect the player to win more often, lose more often, or have equal chances of winning and losing?

Solution. In analyzing this game it is useful to think about what the outcomes look like. Basically there are four sequential rolls of the dice so the outcome $(1, 2, 6, 6)$ is distinguishable from the outcome $(2, 1, 6, 6)$. The first question is 'how many outcomes' are there?

There are six possible outcomes for the first roll. For each of these six outcomes there are six outcomes on the second roll. Thus there are $6 \times 6 = 36$ different ways the first two rolls can turn out. Following this reasoning, there are $1296 = 6 \times 6 \times 6 \times 6$ rolls.

Next we need to calculate how many of these rolls are winners, i.e., include at least one six. This at first seems daunting since the event 'at least one six' can happen so many different ways. However a simple technique(*) involving a the "complementary" event makes this easy. The 'complementary' event in this case is 'no sixes' and is somewhat easier to analyze. The event 'no sixes in four rolls' can happen $625 = 5 \times 5 \times 5 \times 5$ ways. Thus slightly less than half the possible 1296 outcomes include no sixes and the remaining $671 (= 1296 - 625)$ events must include at least one six. In the long run the player will win slightly more often than not. More precisely, one would expect that the player would win this game 51.8% of the time since

$$0.518 = \frac{671}{1296}$$

∎

(*) The difference between a *trick* and a *technique* is that you use a technique more than once.

Notice that in any particular play of a game the outcome is unpredictable. Our intuition is that with *repeated plays over time* the rules of the game will produce a *pattern* in the outcomes. It is this pattern, or *statistical regularity* which is predictable and not the individual outcomes, which remain subject to chance. If we flip a fair coin one hundred times, our intuition is that about half of the flips will be heads and about half will be tails. This expectation of a pattern that is reasonably based on the fact that each outcome is equally likely.

In particular, if we have flipped a fair coin one hundred times and have observed one hundred heads – a pattern we intuitively believe to be most unlikely – the chances of a head on the *next* flip are still *exactly one half*. If the premise that coin is unbiased is correct, then each flip, regardless of what has happened before, is just like every other flip and has the same chance of turning up heads or tails.

A slightly different question is whether one hundred consecutive "heads" is sufficient evidence to cause one to doubt the premise that the coin was fair in the first place. Since this question deals with the *pattern* we expect to see from a fair coin, it is quite different from the question of "what happens on the next flip?" Most reasonable people would probably agree that a pattern of one hundred consecutive "heads" is ample evidence that the coin is biased and therefore far more likely in general to turn up "heads" than not on any flip, including the next one.

However, if one accepts the premise that the coin is unbiased, then the coin has an equally likely chance of turning up heads or tails regardless of the history or absence of history regarding prior flips. After all, the coin doesn't "know" what has happened in the past! The coin just comes up heads or tails. Nuances of this type often lead to misapplication of conclusions in probability and statistics and serve as a cautionary warning to both the beginner and the experienced practitioner.

1. Introduction: Problems.

1. The *coin-and-die* game involves flipping a coin and rolling a die. The player first flips a coin then rolls a die. If the coin is *heads* then the player wins and receives four times as many dollars as spots on the die. If the coin is *tails* then the player loses and pays to the casino four dollars plus as many dollars as there are spots on the die. Does the player or casino have the advantage in this game?

2. In the game *five rolls* the player rolls a single die five times. If the player rolls an even number at least three times, they win. If fewer than three even numbers are rolled, the player loses the amount bet otherwise the casino pays the player the amount bet. Does the player or casino have the advantage in this game or is each equally likely to win?

3. In the game *chuck-a-luck* three dice are rolled. The player picks a number between one and six and bets that at least one of the three dice will show that number. If none of the three dice show the selected number, then the player loses and the casino keeps the bet; otherwise the casino pays at 4-3 odds (i.e., pays $4 for every $3 bet). A player reasons that he has one chance in six that any one of the dice will show his number and so, since there are three dice, he has three chances in six of winning. Since this is even odds the player bets on the game thinking in the long run he will win more than he loses. Find the flaw in his reasoning.

4. Roll a pair of dice 105 times simultaneously. The table below lists the possible outcomes; record the number of times you observe each outcome in the cell where the outcome is listed.

1&1	1&2	1&3	1&4	1&5	1&6
-	2&2	2&3	2&4	2&5	2&6
-	-	3&3	3&4	3&5	3&6
-	-	-	4&4	4&5	4&6
-	-	-	-	5&5	5&6
-	-	-	-	-	6&6

If the outcomes as listed in this table are equally likely, you would expect to see about five outcomes $(105 \div 21)$ in each cell. On the other hand, if the ”sequential” model of the 36 equally likely outcomes is correct, then you would expect to see about 3 outcomes $(105 \div 36)$ in the cells on the diagonal and about 6 in the other cells. Which model is more consistent with your observations?

5. Verify that if p is the chance of winning and if

$$x = \frac{p}{1-p}$$

are the odds of winning, then

$$p = \frac{x}{1+x}.$$

2. Basic Probability

In the course of studying games of chance a set of more-or-less intuitive "rules" about probability were discovered (or invented, depending on your philosophical perspective). These fundamental rules provide the basic elements of probability theory. In this section we introduce these basic rules and reformulate them using the language of unions, intersections, and other operations on sets.

2.1. Example.

Suppose that an urn contains three red balls, four green balls, one white ball and twelve blue balls, for a total of twenty balls altogether. Suppose that except for color the balls are all the same. If one reaches into the urn and draws a ball at random what is the chance that ball is red? What is the chance that the ball is blue? What is the chance that it is red or blue? A ball that is not white?

Solution. Since there are three red balls out of twenty, clearly the chance of a red ball being drawn is three out of twenty or $\frac{3}{20}$. Similarly, the chance of a blue ball is twelve out of twenty or $\frac{3}{5}$.

The chance of a red or blue ball being drawn is equally simple. Since a total of fifteen balls are either red or blue, it follows that the chance of a red or blue ball being drawn is fifteen out of twenty or $\frac{3}{5}$.

Finally, the chance of drawing a ball that is not white is exactly the chance of drawing a ball that is red (3 balls), green (four balls) or blue (twelve balls, i.e., nineteen out of twenty or $\frac{19}{20}$.

∎

This very simple example illustrates a number of fundamental principles.

First, when we say we have selected a ball "at random" we mean that each ball is *equally likely* to be to be the one selected. We can imagine that the person selecting the ball is blindfolded and thus cannot see the color of the selected ball – the only distinguishing characteristic. If the balls were different in some additional way, such as size or texture, then this difference might interfere with the random character of the selection: balls with a rough texture or ones that are larger or smaller might be more likely to be selected. We also would need to assure that the balls were mixed in some random fashion such as by shaking the urn after inserting the balls – if all the blue balls were on top that would make them more likely to be selected. Making sure that a selection of this type is truly "random" can often be a difficult process and is beyond the scope of this course. The basic principle is no particular ball is more likely to be selected than any other.

Second, we have computed the chances by *counting* the number of ways a particular outcome can occur, then dividing by the total number of possible outcomes. Thus since there are three red balls, there are three ways we satisfy the requirement "a red ball is drawn." Since there are twenty balls altogether, there are twenty different ways of selecting a ball. Similarly there are a total of fifteen balls that are either red or blue, hence fifteen different ways a selecting a red or a blue ball out of a total of twenty possible ways.

Notice that the chance of drawing a white ball is one out of twenty or $\frac{1}{20} = 5\%$. If we have a 5% chance of drawing a white ball, then the chance of drawing a non-white ball must be

$$100\% - 5\% = 95\%.$$

We will almost always write probabilities as decimals rather than percentages, so we would normally write the above as

$$\begin{aligned}\mathfrak{Pr}\,(\text{non-white ball}) &= 1 - \mathfrak{Pr}\,(\text{white ball}) \\ &= 1 - .05 \\ &= .95\end{aligned}$$

complementary event

This actually illustrates a fundamental principle. If E represents an event (say, drawing a white ball) then we use E^C for the *complementary event* (say, drawing a non-white ball). Then

$$\mathfrak{Pr}\,(E) = 1 - \mathfrak{Pr}\,(E^C)$$

where the complementary event E^C is the event

$$E^C = \text{"E does not occur"}.$$

Counting is a fundamental part of many computations with probabilities. A slight change to our example will help illustrate some basic counting principles.

2.2. Example.

Suppose that an urn contains twenty balls numbered from one to twenty. The balls are also colored as follows:

red	*numbers 1-3*
green	*numbers 4-7*
white	*number 8*
blue	*numbers 9-20*

Suppose that the balls are indistinguishable except for the numbers and colors painted on the balls. A ball is selected at random. What are the chances that the ball is Green? That the number on the selected ball is Even? That the both the number on the selected ball is even and the ball is green. That the number on the selected ball is even or that the color is green?

Solution. Clearly the chance of a green ball is .2 and the chance of an even number is 0.5.

The event "even and green" consists of exactly two balls: number 4 and number 6. Thus the chance of "even and green" is 0.1. The event "even or green" consists of balls numbered

$$2, 4, 5, 6, 7, 8, 10, 12, 14, 16, 18, 20$$

or 0.6.

■

Using the language of sets we can rewrite the above is somewhat simpler form. First define events as follows:

$$\begin{aligned} E_1 &= \{\text{Green ball is selected}\} \\ E_2 &= \{\text{ball with even number is selected}\} \\ E_3 &= \{\text{selected ball is Green and has an even number}\} \\ E_4 &= \{\text{selected ball is either Green or has an even number}\} \end{aligned}$$

Using the language of sets,

$$E_1 \cap E_2 = E_3$$

and

$$E_2 \cup E_2 = E_4.$$

If we count the ways each of the first three events can occur we get

E_1	four ways
E_2	ten ways
E_3	two ways

Counting the ways that E_4 can occur is easy using the above:

$$\begin{aligned} E_1 &\quad \text{occurs 4 ways} \\ E_2 &\quad \text{occurs 10 ways} \\ E_3 = E_1 \cap E_2 &\quad \text{occurs 2 ways} \end{aligned}$$

To compute E_4, "green or even", suppose we just add together the ways that E_1 and E_2 can occur (14 ways). Then we have counted some outcomes twice, namely when the selected ball is both green and even. These "double-counted" outcomes are exactly $E_3 = E_1 \cap E_2$, so if we deduct the double-counted outcomes from fourteen we get twelve, exactly the number of ways that E_4 can occur.

In particular, we conclude – at least in this example – that

$$\mathfrak{Pr}\,(E_1 \cup E_2) = \mathfrak{Pr}\,(E_1) + \mathfrak{Pr}\,(E_2) - \mathfrak{Pr}\,(E_1 \cap E_2). \tag{2.1}$$

It seems reasonable that this would be true in general since the logic is based on counting and probability is – intuitively – based on counting. To calculate the number of ways $E_1 \cup E_2$ can occur, count the ways E_1 can occur, add the ways E_2 can occur, then deduct the ways that are double-counted, namely the number of ways $E_1 \cap E_2$ can occur.

2.3. Definition.

If two events E_1 and E_1 have nothing in common, i.e., if

$$E_1 \cap E_2 = \phi$$

then we say that E_1 and E_2 are mutually exclusive or, in the language of sets, disjoint.

In the above example, the events

$$E_1 = \text{ a red ball is selected}$$

and

$$E_2 = \text{ the number on the selected ball is at least ten}$$

are mutually exclusive. Since E_1 can happen three ways and E_2 can happen ten ways it makes sense that

$$\mathfrak{Pr}\left(E_1\right) = 0.15 \quad \text{and} \quad \mathfrak{Pr}\left(E_2\right) = 0.5.$$

Since the events are mutually exclusive, it further makes sense that the joint event

$$E_1 \cup E_2 = \text{ a red ball is selected or the number on the ball is at least ten}$$

can happen thirteen ways, so

$$\begin{aligned}\mathfrak{Pr}\left(E_1 \cup E_2\right) &= 0.65 \\ &= 0.15 + 0.5 \\ &= \mathfrak{Pr}\left(E_1\right) + \mathfrak{Pr}\left(E_2\right).\end{aligned}$$

In general we would expect that if E_1 and E_2 are mutually exclusive, then

$$\mathfrak{Pr}\left(E_1 \cup E_2\right) = \mathfrak{Pr}\left(E_1\right) + \mathfrak{Pr}\left(E_2\right).$$

Notice that this is a special case of (1) since $E_1 \cap E_2 = \phi$ and hence $\mathfrak{Pr}\left(E_1 \cap E_2\right) = 0$.

A slightly more complex example expands on the above.

2.4. Example.

A game consists of rolling a fair die repeatedly until a "two" or a "three" appears, at which time the game ends. What is the probability that game ends on the first roll? On the second roll? On the third roll? On the n^{th} roll? On an even-numbered roll?

Solution. Clearly the chances of a "two" or "three" on any given roll are one in three. Thus on the first roll, the chance of a "two" or "three" appearing is one in three.

As we have seen, there are thirty-six different ways that the two rolls of a single die can turn out. Of these eight satisfy the sequence "not "two" or "three" on the first roll, "two" or "three" on second roll:"

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

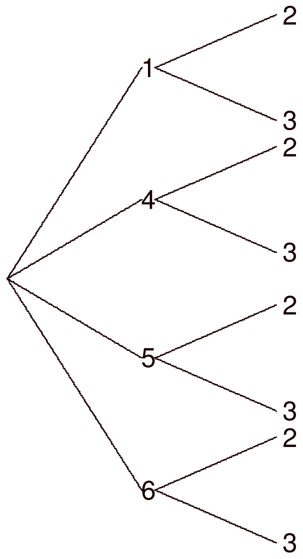
Thus the chance that the first time a "two" or "three" is rolled being is on the second roll is eight out of thirty-six or 2 out of nine.

stem diagram

Another way to think about this is the following. There are four ways that we can roll "one, four, five or six" on the first roll. For each of these four ways, there are two ways to roll a "two" or "three" on the second roll. So there are four times two or eight ways that the first time a "two" or "appears" is on the second roll. Since there are a total of thirty-six ways the roll of a single die can turn out, the chance that the first "two or "three" appears on the second roll is

$$\frac{\text{number of ways first two or three can be on second roll}}{\text{total number of outcomes}} = \mathbf{\frac{8}{36}} = \mathbf{\frac{2}{9}}$$

We can create a graphical representation of the above reasoning with a “stem diagram." The first branches or stems on the diagram represent the first step: rolling a "one, four five or six." Branching out from each of these initial four stems are two more stems, representing the ways that the second step – rolling a "two" or "three" – can occur. Thus the stem diagram starts with four stems, each of the initial four stems has two sub-branches growing of it for a total of eight endpoints.



Stem Diagram for game ending after two rolls

The second approach gives a way to find the chance that the first "two" or "three" occurs on the third roll. There are four ways of *not* rolling a "two" or "three" on the first roll. For each of these ways, there are four ways of *not* rolling a "two" or "three" on the second roll. Thus there are a total of sixteen ways of *not* rolling a "two" or "three" on the first two rolls. For each of these sixteen ways, there are two ways of rolling a "two" or "three" on the third roll. So the total number of outcomes that satisfy "first time a "two" or "three" is rolled is the third roll" is

$$\mathbf{4 \times 4 \times 2.}$$

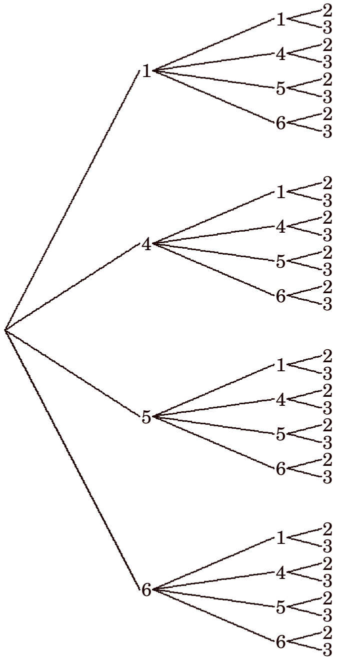
On the other hand, there are

$$\mathbf{6 \times 6 \times 6 = 196}$$

different outcomes when a single die is rolled three times in succession. Thus the chances that the first time a two or three is rolled happens on the third roll is

$$\frac{\mathbf{4^2 \times 2}}{\mathbf{6^3}}$$

The stem diagram is three branches long. The first roll has four stems; each of those initial four grows four new stems. The final row is represented by two more stems growing off of each of the initial sixteen.



Stem Diagram for game ending after three rolls geometric series

Extending this approach, then, it is reasonable that the chances that the first time a "two" or "three" is rolled happens on the n^{th} roll is

$$\frac{4^{n-1} \times 2}{6^n} = \left(\frac{2}{3}\right)^{n-1} \frac{1}{3}$$

To find the chances that we first roll a "two" or "three" on an even-numbered roll we need to compute an infinite sum:

$$\begin{aligned}
&\mathfrak{Pr}\,(\text{first "two" or "three" on even-numbered roll}) = \\
&= \sum_{k \text{ odd}} \frac{2}{3}^{k-1} \frac{1}{3} \\
&= \sum_{n=1}^{\infty} \frac{2}{3}^{2n-1} \frac{1}{3} \\
&= \left(\frac{2}{3}\right)^{-1} \frac{1}{3} \sum_{n=1}^{\infty} \left(\frac{4}{9}\right)^{n} \\
&= \frac{1}{2}\left(\frac{\frac{4}{9}}{1-\frac{4}{9}}\right) \\
&= \frac{2}{5}
\end{aligned}$$

∎

We have used the formula for summing a geometric series from Calculus

$$\sum_{n=1}^{\infty} r^n = \frac{r}{1-r}$$

where $-1 < r < 1$. Also, when we wrote

$$\mathfrak{Pr}\,(\text{first "two" or "three" occurs on even-numbered roll}) = \sum_{k \text{ odd}} \frac{2}{3}^{k-1} \frac{1}{3}$$

we made the assumption that we could calculate the probability of mutually exclusive events by summing the probabilities. More generally, we assumed that if $\{E_n\}$ is an infinite collection of pair-wise mutually exclusive events, then

$$\mathfrak{Pr}\,(\cup E_n) = \sum_n \mathfrak{Pr}\,(E_n)$$

This latter is an "infinite" version of our earlier intuitive observation that if $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$ are mutually exclusive then

$$\mathfrak{Pr}\left(\boldsymbol{E_1} \cup \boldsymbol{E_2}\right) = \mathfrak{Pr}\left(\boldsymbol{E_1}\right) + \mathfrak{Pr}\left(\boldsymbol{E_2}\right).$$

2. Basic Probability: Problems.

1.

3. The Multiplication Rule.

Cardano, Gerolamo
Pascal, Blaise
Fermat, Pierre
Huygens, Christian

With this section we begin a more systematic – and axiomatic – approach to the basic principles of probability. The examples we have carried out thus far all involved some aspect of *counting*. One thread that runs through many games of chance – and all of those we have thus far considered – involves counting how many ways equally likely outcomes can occur.

3.1. Axiom. Cardano Counting Axiom.

If there are n equally likely outcomes and if m of those satisfy a certain condition, then the probability of that condition is $\frac{m}{n}$.

Almost all of our examples thus far have made implicit use of this basic counting axiom. This was first formulated in 1525 in a book written by an Italian physician named Gerolamo Cardano. Cardano's interest arose from his obsession with gambling. Using this axiom Cardano became the first to compute a theoretical (as opposed to empirical) probability.

The actual impact of Cardano's ideas was minimal – his book was not even published until 1663! The mathematical community largely ignored the questions that posed, seeing them as an example of dissolute behavior.

Nearly a century later questions about events occuring by chance finally captured the imagination of the mathematical community. The ideas of Blaise Pascal and Pierre Fermat, worked out in an exchange of letters starting in 1654, laid the foundation of what we today call probability theory. These ideas were formalized in 1657 by Christian Huygens in *De Ratiociniis in Aleae Ludo* (*Calculations in Games of Chance*). The actual problem that led to the correspondence between Pascal and Fermat dealt with, unsurprisingly, a game of chance (see problem one at the end of this section).

Card games provide an almost endless supply of examples of games of chance and associated counting principles. A *poker deck* (sometimes also called a *bridge deck*) consists of 52 cards separated into four suits as follows:

Spades	Hearts	Diamonds	Clubs
♠	♡	♢	♣
King	King	King	King
Queen	Queen	Queen	Queen
Jack	Jack	Jack	Jack
10	10	10	10
9	9	9	9
8	8	8	8
7	7	7	7
6	6	6	6
5	5	5	5
4	4	4	4
3	3	3	3
2	2	2	2
Ace	Ace	Ace	Ace

The backs of all the cards are the same. The front of the card contains symbols that denote the suit (spade ♠, heart ♡, diamond ♢ or club ♣) and the denomination. Traditionally the "king," "queen" and "jack" are stylized pictures of the faces of medieval royalty and hence are called "face cards." The remaining cards have spades, hearts, diamonds or clubs in the same number as the denomination of the card. The card corresponding to "one" is called an "ace" and in many games is the highest rather than the lowest card. Finally Spades and Clubs have black symbols on the front of the card, Hearts and Diamonds have red symbols.

3.2. Example.

A game consists of selecting a card at random from a poker deck. What are the chances of selecting a face card?

Solution. Since the backs of the cards are indistinguishable, we can imagine that each card is equally likely to be selected. Thus there are 52 possible outcomes. Since there are 12 face cards, then the Cardano Counting Axiom tells us that the chances of selecting a face card are

$$\frac{12}{52} = \frac{1}{4}$$

∎

A slightly more complex game is the following.

3.3. Example.

A game consists of rolling a fair die and selecting a card random from a poker deck. What are the chances that outcome consists of a rolling a "2" or "3" and then selecting an Ace?

Solution. We first calculate the total number of possible outcomes when one rolls a die and selects a card at random. To do this we extend the notion of a stem diagram from the previous section. The roll of a die can turn out any one of six ways. For each of these six ways, there are 52 different ways of selecting a card from a poker deck. The resulting stem diagram has 6 initial branches, each of which has 52 sub-branches. This results in a total of $\mathbf{6 \times 52 = 312}$ possible outcomes.

The particular condition we need to satisfy consists of a roll of a "2" or "3" followed by an Ace. Again a stem diagram helps to calculate the number of ways this can happen. A roll of "2" or "3" can happen two ways. Since there are four aces in the poker deck, an ace can be selected in four ways. Thus the stem diagram for the particular condition starts out with two branches, each of which has four sub-branches for a total of 8 branches. Thus we can accomplish the particular condition of this game in 8 ways.

Then the Cardano Counting Axiom implies that the probability is

$$\mathbf{\frac{8}{312} = 0.256.}$$

∎

As games get more complex it becomes less and less practical to construct stem diagrams. For example, a poker hand consists of five cards selected at random from a poker deck. It turns out that there are 2,598,960 different such hands. Clearly we need a more systematic way of calculating than the visual approach afforded by a stem diagram. The answer to this is embodied in the Multiplication Rule.

3.4. Definition. Multiplication Rule.

Suppose that an outcome can be accomplished in k ordered steps. If step one can be accomplished in n_1 ways, step two in n_2 ways and so on, then the ordered sequence

$$\{\textit{step 1, step 2, }\ldots\textit{, step k}\}$$

can be accomplished in

$$n_1 \times n_2 \times \cdots n_k$$

ways.

The multiplication rule formalizes the notion of a stem diagram. The idea is that the first step gives rise to n_1 branches, each of which has n_2 sub-branches and so on.

3.5. Example.

In how many ways can five cards be selected at random in sequence and without replacement from a poker deck?

Solution. When we say "in sequence" we mean that the selection

(3 ♢, J ♣, A ♠, 5 ♡, Q ♠)

is different from the selection

(J ♣, 3 ♢, A ♠, 5 ♡, Q ♠)

in other words, the order in which the cards were selected matters.

The first card can be selected in $n_1 = 52$ ways. Once the first card is selected, there are only 51 cards left, so the second card can be selected in $n_2 = 51$ ways. Reasoning in a similar manner, the five cards can be selected in

$$52 \times 51 \times 50 \times 49 \times 48 = 311,875,200$$

ways.

∎

3.6. Example.

Suppose a password on a computer system must consist of 5 characters: three letters followed by two numbers. The computer system can't tell the difference between upper and lower case letters and letters may be repeated. How many different passwords are possible? Suppose that the numbers may repeat but the letters may not? How many passwords contain the letters RAY?

Solution. Each of the first three characters can be selected in any one of 26 ways, while the last two can be selected in any one of ten ways. Thus the total number of possible passwords is

$$26 \times 26 \times 26 \times 10 \times 10 = 1,757,600$$

If the letters may not repeat, then there are 26 ways to choose the first letter, 25 to choose the second and 24 to choose the third, so the number of different ways is

$$26 \times 25 \times 24 \times 10 \times 10 = 1,560,000$$

The number of passwords satisfying the particular condition that they contain the letters RAY can also be calculated using the multiplication rule. In this case, there are three ways of choosing the first letter from {R, A, Y}. Once the first letter is chosen, there are two ways of choosing the next letter and only one way of choosing the final letter. Thus the number of passwords that contain the letters RAY is

$$3 \times 2 \times 1 \times 10 \times = 600$$

∎

In the previous example suppose that we wanted to carry out the selection in a way that assured that each letter had an equal chance of being selected. To do this we might write each letter on a slip of paper, put the 26 slips in an urn, shake up the urn then reach in and select a letter. If we permit the letters to repeat, then we would replace the selected letter, re-agitate the urn, and select the second letter, and so on.

On the other hand, if the letters did not repeat, we would *not* replace the letter selected on the first drawing, would not replace the letter selected on the second drawing, and so on.

This methodology for randomly selecting the letters gives rise to the notion of selection *with replacement* and *without replacement*. If you have N distinguishable objects, such as $N = 26$ letters, then selecting k of the objects "with replacement" means that the objects may repeat, i.e., that a selected object is returned to pool before the next selection takes place. Selecting k objects "without replacement" means that the pool of possible objects is diminished by one after each choice, i.e., that no object may repeat in the selection.

The multiplication rule enables one to count the number of ordered selections, both with and without replacement. In the latter case, factorial notation is useful.

3.7. Definition. Factorials.

If n is a positive integer then the factorial of n is the number

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

For example,

$$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$$

With this notation in hand, the following two counting principles are special cases of the multiplication rule.

3.8. Definition. Selection with replacement.

The number of ordered selections of size k from N objects where objects may be selected more than once is

$$N^k$$

3.9. Definition. Selection without replacement (permutations).

The number of ordered selections of size k from N objects where no object may be selected more than once is

$$N(N-1)(N-1)\cdots(N-k+1) = \frac{N!}{(n-k)!}$$

Note that if one selects without replacement, then the number of selections k cannot be larger than the number of objects N. However, if one selects with replacement then there is no theoretical limit on the number of objects k that could be selected.

3. The Multiplication Rule.: Problems.

1. Two people, A and B, agree to play a series of fair games until one person has won six games. They have each wagered the same amount of money and whoever is the first to win six games will collect the entire amount wagered. Suppose, however, that they are unable to complete their wager and must terminate early, at a point when person A has won five games and person B has won three games. How should the wager be fairly divided?

In a poker hand what matters is the five cards you are dealt, not the order in which they were dealt to you. This is true of many games of chances and of many applications of probability: what matters is not the order of the outcomes but which outcomes occured. As usual, an example can help illuminate more basic principles.

4.1. Example.

A poker player has been dealt five cards:

$$A\spadesuit, A\heartsuit, Q\clubsuit, 4\diamondsuit, 2\diamondsuit.$$

In how many different ways can he arrange these five cards from left to right in his hand?

Solution. This is a simple application of permutations. There are $5! = 120$ ways of selecting five cards, in order and without replacement, from the five cards in his hand. This is exactly the number of ways that the player can arrange the five cards in order, from left to right, in his hand.

■

In the last section we saw that there are

$$\frac{52!}{(52-5)!} = 311,875,200$$

ways that five cards can be selected, in order, from a poker deck. When a poker player is dealt five cards the deal does occur without replacement. However, even if the cards are dealt in a particular order the player doesn't care what order the cards are dealt, just which cards are dealt. for example, the hands

$$A\spadesuit, A\heartsuit, Q\clubsuit, 4\diamondsuit, 2\diamondsuit$$

and

$$A\heartsuit, A\spadesuit, Q\clubsuit, 4\diamondsuit, 2\diamondsuit$$

are the "same" even though the first and second cards are reversed in order. Both hands constitute the poker hand "a pair of Aces." Thus many of the 311 million ordered hands counted above are actually the "same" poker hand since the order in which the cards are dealt is irrelevant to the poker player.

The poker player wants to know, for example, how likely are various hands such as those listed below?

one pair	two of the same denomination and three not matching
three-of-a-kind	three matching cards and two non-matching
full house	one pair and one three-of-a-kind
flush	five in the same suit

There are, of course, other possible poker hands; these are just listed as examples.

As a preliminary step, we will calculate the number of possible poker hands.

4.2. Example. Example Poker Hands.

A poker hand consists of five cards selected at random and without replacement from a deck of 52 cards. How many hands are there if the order matters? How many if the order does not matter?

Solution. As we have seen, if the order in which the cards are selected matters then there are

$$\frac{52!}{(52-5)!} = 311,875,200$$

possible hands. Next we will use the multiplication rule and example one to calculate this number in a different way.

Let x be the number of possible hands if the order does not matter. Now each of these x hands can be ordered in $5! = 120$ different ways by the first example in this section. Thus we can think of the task of constructing all of the "ordered" five card hands as taking two sequential steps:

Step 1. Select an *unordered* hand of five cards;

Step 2. Order the selected hand of five cards.

The first step can happen in x ways, the second step can happen in $5!$ ways. Thus by the multiplication rule the total number of ordered hands must be

$$x \times 5!$$

This is a second way of finding the number of ordered hands and so must equal the first way, i.e.,

$$x \times 5! = \frac{52!}{(52-5)!}.$$

Thus, solving for x,

$$x = \frac{52!}{(52-5)!5!} = 2,598,960.$$

This latter number is the number of five-card poker hands where the order in which the cards were dealt does not matter.

∎

The preceding example generalizes in a natural way.

4.3. Definition. Combinations

The number combinations of n things taken k at a time is

$$\binom{n}{k} \equiv \frac{n!}{(n-k)!k!}$$

4.4. Theorem. Combination Rule.

The number of combinations of n things taken k at a time is exactly the number of un-ordered selections of k objects from n distinct objects where no object can be selected more than once.

Proof. The proof follows exactly the argument in the preceding example. The number of *ordered* selections of k objects from n distinct objects where no object can be selected more than once is

$$\frac{n!}{(n-k)!}$$

Alternatively, let x be the number of *un-ordered* selections. Each of these un-ordered selections can be arranged in $k!$ different ways. Thus all of the ordered selections can be constructed in a two step process, first selected an un-ordered group of k objects, then order it. The multiplication rule says that these two tasks together can be done in

$$x \times k!$$

ways.

These two different ways of calculating the number of ordered samples must be equal, so

$$\frac{n!}{(n-k)!} = x \times k!$$

which gives on solving for x

$$x = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

∎

4.5. Example. Full House in Poker.

A **full house** *in poker consists of a pair and three-of-a-kind in one hand. The order in which the cards are dealt does not matter. How many different poker hands result in a full house?*

Solution. The multiplication rule can be used to solve this problem. Dividing the problem of "constructing" a full house up into tasks we might consider:
Task 1. Select the denomination for the pair.
Task 2. Select two cards of that denomination to be in the hand.
Task 3. Select the denomination for the three-of-a-kind.
Task 4. Select the three cards of that denomination to be in the hand.

Then Task 1 can be accomplished in $\binom{13}{1}$ ways and Task 2 can be accomplished in $\binom{4}{2}$ ways. Task 3 can be accomplished in $\binom{12}{1}$ (there are only 12 denominations left to choose from since we have already selected one for the pair). Finally task for can be accomplished $\binom{4}{3}$ ways. Thus the total number of full houses is:

$$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{3}.$$

∎

4.6. Example. Arranging symbols.

Suppose that we have $n+1$ red balls and r blue balls.

(a) How many different arrangements are there of the $n+r+1$ balls?

(b) How many different arrangements are of the $n+r+1$ balls where the first and last balls blue?

Solution. In both parts we need select the r positions in which we will place the blue balls. For part (a) there are $n+r+1$ positions that are possibilities for the blue balls so we can do this in

$$\binom{n+r+1}{r} \quad \text{ways.}$$

For (b), the first and last positions are fixed with blue balls, so we have $n+r-1$ positions in which to place the r red balls. Thus the number of ways in which we can do this is

$$\binom{n+r-1}{r} \quad \text{ways.}$$

∎

A slightly different problem involves distributing indistinguishable balls between distinguishable urns. This kind of problem is quite important in particle physics, for example, where the “balls" represent elementary particles and the “urns" represent partitions of the three-dimensional space. Different particles are distributed in different ways, depending on the physical properties of the particle, but in all cases it is necessary to solve problems like the following example.

4.7. Example.

Suppose that you have twenty balls to distribute among five urns. After the distribution, the k^{th} *urn will have* m_k *balls in it. The numbers* $(m_1, m_2, m_3, m_4, m_5)$ *are called the* occupancy numbers *since* m_k *counts how balls occupy the* k^{th} *urn. Since there are twenty balls, these numbers must satisfy* $0 \leq m_k \leq 20$ *and*

$$\sum_{k=1}^{5} m_k = 20.$$

One possible set of occupancy numbers is pictured below.

In how many ways can a particular set of occupancy numbers $(m_1, m_2, m_3, m_4, m_5)$ *occur?*

Solution. There are five tasks to accomplish.

Task 1.	Choose m_1 balls from the original 20 and place them in Urn 1.
Task 2.	Choose m_2 balls from the remaining $20 - m_1$ balls and place them in Urn 2.
Task 3.	Choose m_3 balls from the remaining $20 - m_1 - m_2$ balls and place them in Urn 3.
Task 4.	Choose m_4 balls from the remaining $20 - m_1 - m_2 - m_3$ balls and place them in Urn 4.
Task 5.	Choose m_5 balls from the remaining $20 - m_1 - m_2 - m_3 - m_4$ balls and place them in Urn 5.

From this we can calculate the number of ways each task can be done.

Task 1.	$\binom{20}{m_1} = \frac{20!}{(20-m_1)!(m_1)!}$
Task 2.	$\binom{20-m_1}{m_2} = \frac{(20-m_1)!}{(20-m_1-m_2)!(m_2)!}$
Task 3.	$\binom{20-m_1-m_2}{m_3} = \frac{(20-m_1-m_2)!}{(20-m_1-m_2-m_3)!(m_3)!}$
Task 4.	$\binom{20-m_1-m_2-m_3}{m_4} = \frac{(20-m_1-m_2-m_3)!}{(20-m_1-m_2-m_3-m_4)!(m_4)!}$
Task 5.	$\binom{20-m_1-m_2-m_3-m_4}{m_5} = \frac{(20-m_1-m_2-m_3-m_4)!}{(20-m_1-m_2-m_3-m_4-m_5)!(m_5)!}$

Multiplying these together, canceling out like terms and using the fact that

$$\boldsymbol{\sum_{k=1}^{5} m_k = 20.}$$

gives that the number of ways is

$$\boldsymbol{\frac{20!}{(m_1)!(m_2)!(m_3)!(m_4)!(m_5)!}}$$

∎

The general statement of the above is the following example.

Maxwell-Boltzmann statistics

4.8. Example.

Suppose that one has r balls and distributes them among n urns. If the m_i is the "occupancy number" for the i^{th} urn then

$$m_1 + m_2 + \cdots + m_n = r$$

and $0 \leq m_i \leq r$ for each i. A particular set of occupancy numbers

$$m_1, m_2, \cdots, m_n$$

represents

$$\frac{r!}{m_1!m_2!m_3!\cdots m_n!}$$

of the total n^r outcomes.

While "balls in urns" are a useful way of phrasing this problem, similar problems arise in statistical mechanics. In this setting a region is subdivided into a large number (n) of smaller sub-regions and the occupancy number corresponds to the distribution of r elementary particles into these regions. There are various assumptions that one might make about the liklihood of different distributions of the particles, each with different physical consequences. For example if all n^r outcomes are equally likely then the probability of any particular set of occupancy numbers is

$$\frac{r!}{m_1!m_2!m_3!\cdots m_n!}n^{-r}$$

and physicists speak of the *Maxwell-Boltzmann statistics*. While this seems intuitively attractive, it turns out that this assumption does not apply to any known collection of particles. Thus, for elementary particles, not all of the n^r outcomes are equally likely!

The occupancy numbers

$$m_1, m_2, \cdots, m_n$$

must solve the equation

$$m_1 + m_2 + \cdots + m_n = r. \tag{4.1}$$

and satisfy $0 \leq m_i \leq r$ for each i. It is possible to show (see the problems) that number of different solutions

$$m_1, m_2, \cdots, m_n$$

to equation (4.1) satisfying $0 \leq m_i \leq r$ for each i is

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

Bose-Einstein statistics

Fermi-Dirac statistics

If one assumes that each of the *occupancy numbers* are equally likely (rather than each of the individual n^r distributions of particles among the regions), then it follows that the chance of any particular occupancy number is

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

Physicists would call this assumption the *Bose-Einstein statistics*. It turns out that this assumption can be used to model certain elementary particles such as photons.

If one assumes that the occupancy numbers m_i can be at most one (no two particles in the same region), and that the corresponding solutions to (O) are equally likely, then physicists refer to the resulting physical model as the *Fermi-Dirac statistics*. This model applies to, for example, protons, electrons and neutrons. In this case any particular set of occupancy numbers has probability

$$\binom{n}{r}^{-1}$$

since this corresponds to the problem of selecting the r positions for the particles from the n partitions.

From a mathematical perspective, there is no reason to prefer one model over another. In particular, absent physical observations, there is no mathematical reason to suspect that photons and neutrons, for example, would follow different probabilistic models. From the standpoint of probability, the point is that combinations have applications ranging from simple card games to statistical mechanics.

4. Combinations.: Problems.

1. Show that the number of distinct solutions

$$m_1, m_2, \cdots, m_n$$

to equation (O) satisfying $0 \leq m_i \leq r$ for each i is

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

Hint. Represent the urns as containers with bars separating each container; for example six bars represent five containers:

$$|\cdot|\cdot|\cdot|\cdot|\cdot|$$

Then $n+1$ bars represent the n containers. Now represent the balls as stars and distribute the balls among the containers. For example

$$|**||***|*||$$

might represent $n = 5$ urns, $r = 6$ balls and occupancy numbers of

$$2, 0, 3, 1, 0.$$

2. A student attempts to calculate the number of full houses with the following set of tasks:
Task 1. Select the denominations that will appear in the hand.
Task 2. Select one of those two to be the pair.
Task 3. Select two cards of that denomination to be in the hand.
Task 4. Select the three cards of the remaining denomination to be in the hand.

This results in an answer of

$$\binom{13}{2}\binom{2}{1}\binom{4}{2}\binom{4}{1}$$

which appears to be different from the answer in the text. Comment.

5. Probability Spaces

In the previous sections we calculated how likely the outcomes of various games might be and examined some consequences of counting with respect to probability. In each case we were able to define or list all possible outcomes – the *universe* – and then calculate the liklihood of any individual outcome. From a mathematical perspective, the universe of all possible outcomes, Ω, is a set. Individual outcomes ω are elements of the universe, $\omega \in \Omega$. A subset of $E \subseteq \Omega$ is called an *event*. In this context, then, individual outcomes ω are singleton events $\{\omega\}$.

In previous sections we were satisfied with an intuitive understanding of probabilities and events based on counting. In this section we give this intuitive approach a more abstract and axiomatic – and hence more mathematical – foundation.

5.1. Definition.

Let Ω be a set and let $\mathcal{E}$ be a non-empty collection of subsets of Ω; the pair $(\Omega, \mathcal{E})$ is said to be a **σ-algebra** *if*
(i) $E \in \mathcal{E}$ implies $E^c \in \mathcal{E}$;
(ii) whenever $\{E_1, E_2, \cdots\} \subset \mathcal{E}$ then both

$$\bigcap_n E_n \in \mathcal{E} \quad \text{and} \quad \bigcup_n E_n \in \mathcal{E}.$$

We can think of the elements $\omega \in \Omega$ as being the *simple outcomes* of a random experiment and the subsets $\mathcal{E}$ as being the collection of *events* whose probabilities we can compute.

5.2. Proposition.

If $(\Omega, \mathcal{E})$ is a σ-algebra, then
(i) $\phi \in \mathcal{E}$
(ii) $\Omega \in \mathcal{E}$
(iii) if E_1 and E_2 are in Ω then both $E_1 \cup E_2 \in \Omega$ and $E_1 \cap E_2 \in \Omega$.

Proof. Since E is a non-empty collection, there is at least one set $E \in \mathcal{E}$. Set $E_1 = E$

and, for $n \geq 2$, set $E_n = E^c$. Then

$$\Omega = \bigcup_n E_n \in \mathcal{E}$$

by *(ii)*. Thus $\Omega \in \mathcal{E}$. It then follows from *(i)* that

$$\phi = \Omega^c \in \mathcal{E}.$$

For *(iii)*, set $E_n = E_2$ for $n \geq 3$. Then

$$E_1 \cap E_2 = \bigcap_n E_n \quad \text{and} \quad E_1 \cup E_2 = \bigcup_n E_n$$

showing *(iii)*.

$\blacksquare$

5.3. Definition.

Let $(\Omega, \mathcal{E})$ *be a* σ*-algebra and let*

$$\mathfrak{Pr} : \mathcal{E} \to \mathbb{R}$$

be a real-valued function defined on $\mathcal{E}$*. Then* $\mathfrak{Pr}$ *is a probability function and the triple* $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *is a probability space provided that*
(i) $0 \leq \mathfrak{Pr}\,(E) \leq 1$ *for all* $E \in \mathcal{E}$*;*
(ii) $\mathfrak{Pr}\,(\Omega) = 1$*; and*
(iii) if $\{E_n\}$ *is a pair-wise disjoint collection of sets in E then*

$$\mathfrak{Pr}\,(\cup_n E_n) = \sum_n \mathfrak{Pr}\,(E_n).$$

5.4. Proposition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *be a probability space. Then*
(i) $\mathfrak{Pr}(\phi) = 0$
(ii) If E_1 *and* E_2 *are disjoint sets in* $\mathcal{E}$ *then*

$$\mathfrak{Pr}(E_1 \cup E_2) = \mathfrak{Pr}(E_1) + \mathfrak{Pr}(E_2)$$

(iii) If E_1 *and* E_2 *are any sets in* $\mathcal{E}$ *then*

$$\mathfrak{Pr}(E_1 \cup E_2) = \mathfrak{Pr}(E_1) + \mathfrak{Pr}(E_2) - \mathfrak{Pr}(E_1 \cap E_2)$$

Proof. Parts (i) and (ii) follow in obvious ways from Proposition 3.2 and the definitions. For part (iii) first observe that if

$$E_1 \setminus E_2 \equiv \{\omega \in \Omega | \omega \in E_1 \text{ and } \omega \notin E_2\}$$

then $E_1 \setminus E_2$ and E_2 are disjoint. Thus from (ii)

$$\mathfrak{Pr}\ ((E_1 \setminus E_2) \cup E_2) = \mathfrak{Pr}(E_1 \setminus E_2) + \mathfrak{Pr}(E_2).$$

Note that

$$(E_1 \setminus E_2) \cup E_2 = E_1 \cup E_2$$

and hence

$$\mathfrak{Pr}(E_1 \cup E_2) = \mathfrak{Pr}(E_1 \setminus E_2) + \mathfrak{Pr}(E_2)$$

or

$$\mathfrak{Pr}(E_1 \cup E_2) - \mathfrak{Pr}(E_2) = \mathfrak{Pr}(E_1 \setminus E_2)$$

In an exactly similarly fashion

$$\mathfrak{Pr}(E_1 \cup E_2) - \mathfrak{Pr}(E_1) = \mathfrak{Pr}(E_2 \setminus E_1).$$

Now the three sets

$$E_1 \setminus E_2, \quad E_2 \setminus E_1 \quad \text{and} \quad E_1 \cap E_2$$

are disjoint and

$$E_1 \cup E_2 = (E_1 \setminus E_2) \cup (E_2 \setminus E_1) \cup (E_1 \cap E_2).$$

From this and the above

$$\begin{aligned}\mathfrak{Pr}\,(E_1 \cup E_2) &= \mathfrak{Pr}\,(E_1 \setminus E_2) + \mathfrak{Pr}\,(E_2 \setminus E_1) + \mathfrak{Pr}\,(E_1 \cap E_2) \\ &= \mathfrak{Pr}\,(E_1 \cup E_2) - \mathfrak{Pr}\,(E_2) + \mathfrak{Pr}\,(E_1 \cup E_2) - \mathfrak{Pr}\,(E_1) + \mathfrak{Pr}\,(E_1 \cap E_2)\end{aligned}$$

Rearranging gives (iii).

∎

Thus from this minimal abstract structure we can deduce the intuitive characteristics of probability that we inferred from examples on counting. Even more striking, this minimal structure provides a framework for deducing results about limits of sequences of events. The theorem below is a kind of ‘continuity’ theorem expressed in terms of events and will useful later as we consider various real-valued functions defined on probability spaces.

5.5. Theorem.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space.

(i) If $\{E_n\}$ are events in $\mathcal{E}$ and if $E_1 \subseteq E_2 \subseteq E_3 \subseteq \cdots$ then

$$\mathfrak{Pr}\left(\bigcup_n^\infty E_n)\right) = \lim_{N\to\infty} \mathfrak{Pr}\,(E_N)$$

(ii) If $\{E_n\}$ are events in $\mathcal{E}$ and if $E_1 \supseteq E_2 \supseteq E_3 \supseteq \cdots$ then

$$\mathfrak{Pr}\left(\bigcap_n^\infty E_n)\right) = \lim_{N\to\infty} \mathfrak{Pr}\,(E_N)$$

Proof. For *(i)* set $A_1 = E_1$ and, for $n > 1$ set

$$\begin{aligned}A_n &= \{\omega \in \Omega : \omega \in E_n \quad \text{and} \quad \omega \notin E_{n-1}\} \\ &= E_n \setminus E_{n-1}.\end{aligned}$$

The sets $\{A_n\}$ are pair-wise disjoint and

$$\begin{aligned} \bigcup_{n=1}^{N} A_n &= \bigcup_{n=1}^{N} E_n \\ &= E_N. \end{aligned}$$

note that this further implies that

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} E_n.$$

Thus

$$\begin{aligned} \mathfrak{Pr}\left(\bigcup_{n}^{\infty} E_n)\right) &= \mathfrak{Pr}\left(\bigcup_{n}^{\infty} A_n)\right) \\ &= \sum_{1}^{\infty} \mathfrak{Pr}\left(A_n\right) \\ &= \lim_{N\to\infty} \sum_{n=1}^{N} \mathfrak{Pr}\left(A_n\right) \\ &= \lim_{N\to\infty} \mathfrak{Pr}\left(\bigcup_{n=1}^{N} A_n\right) \\ &= \lim_{N\to\infty} \mathfrak{Pr}\left(E_N\right) \end{aligned}$$

as desired.

For *(ii)* note that

$$E_1^c \subseteq E_2^c \subseteq E_3^c \cdots$$

and that

$$\left(\bigcap_{n} E_n\right)^c = \bigcup_{n} E_n^c.$$

Thus

$$\begin{aligned} \lim_{N\to\infty} \mathfrak{Pr}\left(E_N\right) &= \lim_{N\to\infty} (1 - \mathfrak{Pr}\left(E_N^c\right) \\ &= 1 - \lim_{N\to\infty} \mathfrak{Pr}\left(\bigcup_{n=1}^{N} E_n^c\right) \\ &= 1 - \mathfrak{Pr}\left((\cap_{n=1}^{\infty} E_n)^c\right) \\ &= \mathfrak{Pr}\left(\cap_{n=1}^{\infty} E_n\right) \end{aligned}$$

as desired. ∎

5. Probability Spaces: Problems.

1. A Game consists of rolling a fair die until the second three appears. Calculate the probability that it takes more than 8 rolls for the second three to appear.

6. Independence and Conditional Probability

The multiplication rule provides a formalized way to think through certain kinds of calculations. The notions of independence and conditional probability are, in some ways, formal generalizations of the multiplication rule. A couple of simple examples will help to illustrate the basic concepts.

6.1. Example.

Suppose that an urn contains r crimson balls numbered $1, 2, \cdots, r$ and n white balls numbered $1, 2, \cdots, n$. A ball is selected at random from the urn and both the number and the color are noted.
(a) What are the chances that the color is crimson?
(b) What are the chances that the number is a "1?"
(c) What are the chances that the color is crimson and the number is a "1?"
(d) If we observe the ball is crimson, what are the chances the number is a "1?"

Solution. For (a)-(c) the total number of balls is $r + n$ and the chance of drawing any single ball is therefore $\frac{1}{r+n}$. Thus we can apply the Cardano Counting Principal to each of (a)-(c). For (a) there are r crimson balls each of which are equally likely so the probability of a crimson ball is $\frac{r}{r+n}$. Similarly, there are two balls with the number "1" so the probability of drawing a ball with the number "1" is $\frac{2}{r+n}$. There is exactly one crimson ball numbered one, so the probability of drawing a crimson ball with the number "1" is $\frac{1}{r+n}$.

For part (d) the number of possible balls is reduced to r since we know we have selected a crimson ball. Of the r crimson balls, exactly one has the number "1" so the probability of drawing a "1" given that we have selected a red ball is $\frac{1}{r}$. Note that this answer is the same as that in (b) *only* in the case $r = n$. In particular, knowing the color of the ball helps us predict the number on the ball.

∎

Our next example is somewhat more complex in that involves successive draws from the urn without replacement.

6.2. Example.

An urn contains 10 balls, of which four are red, four are blue and two are white. Two balls are selected from the urn in sequence and without replacement.

(a)) What are the chances that both balls are red?

(b)) What are the chances that the second ball selected is red and the first one is not red?

(c)) What are the chances that the second ball is red?

(d)) If we know that the first ball selected is red, what are the chances that the second ball selected is red?

Solution. This problem consists of two tasks: selecting the first ball, then selecting the second ball. Depending on the outcomes and what we are given about the conditions, we can compute the probability of each task under the varying circumstances.

For *(a)*, the number of ways of selecting the first ball – with no restrictions on color – is one in ten. After the first ball is selected, there are only nine balls left, so the number of ways of selecting the second ball is nine. Thus, by the multiplication rule, the total number of ways we can select two balls without replacement is $\mathbf{10 \times 9 = 90}$ (the number of permutations of ten things taken two at a time).

First, consider the desired outcome of "both balls red." In this case, the first task is to select a red ball. There are four red balls initially, so there are four ways that this can be done. The second task is to select the second ball and have the second ball also be red. After the first selection, there are only three red balls left, so there are three ways of accomplishing the second task. Thus the total ways of selecting two red balls is $\mathbf{4 \times 3 = 12}$.

From this, applying the Cardano Counting Axiom, the chances of selecting two red balls in sequence and without replacement is

$$\frac{\mathbf{4 \times 3}}{\mathbf{10 \times 9}} = \frac{\mathbf{2}}{\mathbf{15}}.$$

In exactly a similar way we can calculate *(b)* and so find that the chances that the first is not red and the second is red are

$$\frac{\mathbf{6 \times 4}}{\mathbf{10 \times 9}} = \frac{\mathbf{4}}{\mathbf{15}}.$$

The previous two parts describe exactly the way in which *(c)* can occur. If

$$\begin{aligned} \boldsymbol{E_1} &= \{\text{ draw a red ball, followed by a red ball}\} \\ \boldsymbol{E_2} &= \{\text{ draw a non-red ball, followed by a red ball}\} \\ \boldsymbol{E} &= \{\text{ draw a red ball on the second draw}\} \end{aligned}$$

Then

$$E_1 \cup E_2 = E \quad \text{and} \quad E_1 \cap E_2 = \phi$$

and so

$$\begin{aligned} \mathfrak{Pr}\,(E) &= \mathfrak{Pr}\,(E_1 \cup E_2) \\ &= \mathfrak{Pr}\,(E_1) + \mathfrak{Pr}\,(E_2) \\ &= \frac{2}{15} + \frac{4}{15} \\ &= \frac{2}{5} \end{aligned}$$

For *(d)*,we are given that the first ball selected is red so the first selection can be done in four ways. Once the first ball is selected, there are nine ways to select the second ball, so there are a total of $4 \times 9 = 36$ was of selecting a red ball on the first draw (with the color of the second draw not specified). As we saw in *(a)* there are twelve ways that two red balls can be selected. Thus the chances that the second ball is red given that the first ball is red are

$$\frac{12}{36} = \frac{1}{3}.$$

In calculating *(d)* we counted the number of ways of selecting two red balls in succession (12) and the number of ways of selecting a red ball followed by a ball of any color (36). In particular if

$$\begin{aligned} A &= \{\text{two red balls}\} \qquad \text{and} \\ B &= \{\text{red ball on first draw, any color on the second draw}\} \end{aligned}$$

then with *no* fore-knowledge of how the first draw turned out

$$\mathfrak{Pr}\,(A) = \frac{12}{90} \qquad \text{and} \quad \mathfrak{Pr}\,(B) = \frac{36}{90}.$$

Notice that

$$\begin{aligned} \frac{\left(\frac{12}{90}\right)}{\left(\frac{36}{90}\right)} &= \frac{12}{36} \\ &= \mathfrak{Pr}\left(\begin{matrix}\text{red on second draw given that} \\ \text{a red was selected on the first draw}\end{matrix}\right) \end{aligned}$$

In particular since $A \cap B$ = A in this particular example,

$$\mathfrak{Pr}\ (A \quad \text{given} \quad B) = \frac{\mathfrak{Pr}\,(A \cap B)}{\mathfrak{Pr}\,(B)}$$

The above formula, which makes intuitive sense in terms of counting principles, is the basis for the general definition of conditional probability.

6.3. Definition.

If A and B are events with $\mathfrak{Pr}\left(B\right) \neq 0$, then the probability of A given B, denoted by $\mathfrak{Pr}\left(A|B\right)$ is the numberconditional probability

$$\mathfrak{Pr}\left(A|B\right) = \frac{\mathfrak{Pr}\left(A \cap B\right)}{\mathfrak{Pr}\left(B\right)}$$

In our first example, if A is the even "draw a ball numbered '1'" and B is the even "draw a crimson ball" then we calculated

$$\mathfrak{Pr}\left(A\right) = \frac{2}{r+n}, \quad \mathfrak{Pr}\left(B\right) = \frac{r}{r+n} \quad \text{and} \quad \mathfrak{Pr}\left(A \cap B\right) = \frac{1}{r+n}.$$

From this, applying the definition,

$$\begin{aligned}\mathfrak{Pr}\left(A|B\right) &= \frac{\mathfrak{Pr}\left(B \cap A\right)}{\mathfrak{Pr}\left(B\right)} \\ &= \frac{\frac{1}{r+n}}{\frac{r}{r+n}} \\ &= \frac{1}{r}\end{aligned}$$

which agrees with the answer we deduced directly in the example.

In our second example, if A is the event "draw a red ball on the second draw" and B is the event "draw a red ball of the first draw" then we have shown that

$$\mathfrak{Pr}\left(A\right) = \frac{2}{5}, \quad \mathfrak{Pr}\left(B\right) = \frac{2}{5} \quad \text{and} \quad \mathfrak{Pr}\left(A \cap B\right) = \frac{2}{15}$$

and so

$$\begin{aligned}\mathfrak{Pr}\left(A|B\right) &= \frac{\mathfrak{Pr}\left(A \cap B\right)}{\mathfrak{Pr}\left(B\right)} \\ &= \frac{\frac{2}{15}}{\frac{2}{5}} \\ &= \frac{1}{3}\end{aligned}$$

In particular,

$$\mathfrak{Pr}\left(A|B\right) \neq \mathfrak{Pr}\left(A\right)$$

i.e., knowing what happened on the first draw helps to predict what might happen on the second draw. This makes sense since we are drawing without replacement: if we know that we drew a red ball on the first draw, then there are fewer red balls to choose from on the second draw, hence the chances of a red ball on the second draw are diminished.

6.4. Example.

Suppose an unbiased coin is flipped three times. What is the probability that exactly two flips are heads given that at least one is heads?

Solution. Let A be the event "two heads in three flips" and B be the event "at least one head in three flips." Then A can happen exactly three ways:

$$\begin{gathered} HHT \\ HTH \\ THH \end{gathered}$$

each of which have probability $\left(\frac{1}{2}\right)^3$. Thus

$$\mathfrak{Pr}\left(A\right) = \frac{3}{8}.$$

For B it is easier to calculate the complementary event:

$$B^C = \{\text{no heads in three flips}\}.$$

Since B^C is exactly "three tails in three flips"

$$\mathfrak{Pr}\left(B\right) = 1 - \mathfrak{Pr}\left(B^C\right) = 1 - \left(\frac{1}{2}\right)^3 = \frac{7}{8}$$

Finally, $A \cap B = A$ and so

$$\begin{aligned} \mathfrak{Pr}\left(A|B\right) &= \frac{\mathfrak{Pr}\left(A \cap B\right)}{\mathfrak{Pr}\left(B\right)} \\ &= \frac{\frac{3}{8}}{\frac{7}{8}} \\ &= \frac{3}{7} \end{aligned}$$

∎

An important consequence of the definition of conditional probability is the following.

6.5. Theorem. Bayes' Rule.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ *be a probability space and let* $A, B \in \mathcal{E}$*. If* $\mathfrak{Pr}\,(A) \neq 0 \neq \mathfrak{Pr}\,(B)$ *then*

$$\mathfrak{Pr}\,(A|B) = \frac{\mathfrak{Pr}\,(A)\,\mathfrak{Pr}\,(B|A)}{\mathfrak{Pr}\,(B)}.$$

Proof. It follows from the definition that

$$\mathfrak{Pr}\,(A \cap B) = \mathfrak{Pr}\,(A|B)\,\mathfrak{Pr}\,(B)$$

and

$$\mathfrak{Pr}\,(A \cap B) = \mathfrak{Pr}\,(B|A)\,\mathfrak{Pr}\,(A).$$

Thus

$$\mathfrak{Pr}\,(A|B)\,\mathfrak{Pr}\,(B) = \mathfrak{Pr}\,(B|A)\,\mathfrak{Pr}\,(A)$$

or

$$\mathfrak{Pr}\,(A|B) = \frac{\mathfrak{Pr}\,(A)\,\mathfrak{Pr}\,(B|A)}{\mathfrak{Pr}\,(B)}.$$

■

6.6. Corollary.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ *be a probability space and let* $A, B \in \mathcal{E}$*. If* $\mathfrak{Pr}\,(A) \neq 0 \neq \mathfrak{Pr}\,(B)$ *then*

$$\mathfrak{Pr}\,(A|B) = \frac{\mathfrak{Pr}\,(A)\,\mathfrak{Pr}\,(B|A)}{\mathfrak{Pr}\,(A)\,\mathfrak{Pr}\,(B|A) + \mathfrak{Pr}\,(A^C)\,\mathfrak{Pr}\,(B|A^C)}.$$

Proof. Note that

$$\begin{aligned} B &= B \cap \left(A \cup A^C\right)) \\ &= (B \cap A)) \cup \left(B \cap A^C\right). \end{aligned}$$

Since $B \cap A$ and $B \cap A^C$ are disjoint,

$$\begin{aligned} \mathfrak{Pr}\,(B) &= \mathfrak{Pr}\,(B \cap A) + \mathfrak{Pr}\,(B \cap A^C) \\ &= \mathfrak{Pr}\,(A)\,\mathfrak{Pr}\,(B|A) + \mathfrak{Pr}\,(A^C)\,\mathfrak{Pr}\,(B|A^C). \end{aligned}$$

The result now follows from Bayes' Rule.

■

In exactly the same manner we can deduce a slightly more general version of the above.

6.7. Corollary.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space and let $B \in \mathcal{E}$. Suppose that $\{A_1, A_2, \cdots, A_n\} \subseteq \mathcal{E}$ are disjoint sets for which

$$\mathfrak{Pr}\,(A_i) \neq 0 \quad i = 1, \cdots, n$$

and

$$\cup_{i=1}^n A_i = \Omega.$$

Then for each $i = 1, \cdots, n$

$$\mathfrak{Pr}\,(A_i|B) = \frac{\mathfrak{Pr}\,(A_i)\,\mathfrak{Pr}\,(B|A_i)}{\sum_{k=1}^n \mathfrak{Pr}\,(A_k)\,\mathfrak{Pr}\,(B|A_k)}.$$

Bayes' rule is most often used when the "given" and the "unknown" probabilities are reversed.

6.8. Example.

An urn contains 10 balls, of which four are red, four are blue and two are white. Two balls are selected from the urn in sequence and without replacement. What is the probability that the first ball selected is red given that the second ball is blue?

Solution. Let A_1 be the event "first ball is red", let A_2 be the event "first ball is blue", let A_3 be the event "first ball is white," and let B be the event "second ball is blue." Then

$$\mathfrak{Pr}\left(A_1|B\right) = \frac{\mathfrak{Pr}\left(A_1\right)\mathfrak{Pr}\left(B|A_1\right)}{\mathfrak{Pr}\left(A_1\right)\mathfrak{Pr}\left(B|A_1\right) + \mathfrak{Pr}\left(A_2\right)\mathfrak{Pr}\left(B|A_2\right) + \mathfrak{Pr}\left(A_3\right)\mathfrak{Pr}\left(B|A_3\right)}.$$

Now

$$\begin{aligned} \mathfrak{Pr}\left(B|A_1\right) &= \frac{4}{9} \\ \mathfrak{Pr}\left(B|A_2\right) &= \frac{3}{9} \\ \mathfrak{Pr}\left(B|A_3\right) &= \frac{4}{9} \end{aligned}$$

while

$$\begin{aligned} \mathfrak{Pr}\left(A_1\right) &= \frac{4}{10} \\ \mathfrak{Pr}\left(A_2\right) &= \frac{4}{10} \\ \mathfrak{Pr}\left(A_3\right) &= \frac{2}{10} \end{aligned}$$

Thus

$$\begin{aligned} \mathfrak{Pr}\left(A_1|B\right) &= \frac{\mathfrak{Pr}\left(A_1\right)\mathfrak{Pr}\left(B|A_1\right)}{\mathfrak{Pr}\left(A_1\right)\mathfrak{Pr}\left(B|A_1\right) + \mathfrak{Pr}\left(A_2\right)\mathfrak{Pr}\left(B|A_2\right) + \mathfrak{Pr}\left(A_3\right)\mathfrak{Pr}\left(B|A_3\right)} \\ &= \frac{\frac{4}{10}\frac{4}{9}}{\frac{4}{10}\frac{4}{9} + \frac{4}{10}\frac{3}{9} + \frac{2}{10}\frac{4}{9}} \\ &= \frac{4}{9} \end{aligned}$$

∎

In contrast, if we had drawn with replacement in the above example, then the outcome of the first draw does not influence the number of ways the second draw can occur so we would obtain

$$\mathfrak{Pr}\left(B|A_i\right) = \mathfrak{Pr}\left(B\right) = \frac{4}{10}$$

for $i = 1, 2, 3$. Extending this

$$\mathfrak{Pr}\left(B \cap A_i\right) = \mathfrak{Pr}\left(B|A_i\right)\mathfrak{Pr}\left(A_i\right) = \mathfrak{Pr}\left(B\right)\mathfrak{Pr}\left(A_i\right)$$

for $i = 1, 2, 3$, again assuming the selections are made with replacement. In this case we say that A_i and B are **independent** since knowledge about how A_i turned out does not provide additional information about how B turns out. More formally,

6.9. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let $A, B \in \mathcal{E}$. Then we say that A and B are independent if

$$\mathfrak{Pr}\left(A \cap B\right) = \mathfrak{Pr}\left(A\right)\mathfrak{Pr}\left(B\right).$$

6.10. Example.

Let $\Omega = \{1, 2, 3, 4\}$ and suppose that

$$\mathfrak{Pr}\left(k\right) = \frac{1}{4}$$

for $k = 1, 2, 3, 4$. Set $A = \{1, 2\}$, $B = \{1, 3\}$ and $C = \{1, 4\}$. Then A, B and C are pair-wise independent but C and $A \cap B$ are not independent.

Solution. It is routine to show that the three events are pair-wise independent. Note that $\mathfrak{Pr}\left(C\right) = \frac{1}{2}$ while

$$\mathfrak{Pr}\left(C|A \cap B\right) = 1$$

and so C and $A \cap B$ are not independent.

∎

In most applications, when we consider a collection of events $\{E_1, E_2, \cdots, E_n\}$ we will need to know not only that the events are pair-wise independent but that any subset of the events are also independent. More formally,

6.11. Definition.

Let $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{E}}, \mathfrak{Pr})$ *be a probability space and let* $\{\boldsymbol{E_1}, \boldsymbol{E_2}, \cdots, \boldsymbol{E_n}\} \subseteq \boldsymbol{\mathcal{E}}$*. Then we say that* $\{\boldsymbol{E_1}, \boldsymbol{E_2}, \cdots, \boldsymbol{E_n}\}$ *are mutually independent if*

$$\mathfrak{Pr}\left(\cap_{i_k} \boldsymbol{E}_{i_k}\right) = \boldsymbol{\Pi}_{i_k} \mathfrak{Pr}\left(\boldsymbol{E}_{i_k}\right)$$

for any set of indices $\{\boldsymbol{i_k}\} \subseteq \{\boldsymbol{1}, \boldsymbol{2}, \cdots, \boldsymbol{n}\}$*.*

Let's Make a Deal
Hall, Monty

6. Independence and Conditional Probability: Problems.

1. Suppose that there are three chests, each with two drawers. In one chest, both drawers contain a gold coin. In another chest, one drawer contains a silver coin and the other contains a gold coin. In the final chest, both drawers contain a silver coin. A chest is selected at random and then a drawer in that chest is selected at random and opened. If the open drawer contains a gold coin, what are the chances that the other drawer contains a gold coin?

2. The game show *Let's Make a Deal* was hosted by Monty Hall. On the game show, contestants were confronted with three closed doors. Behind one of the doors there was a desirable prize (say, a new car) while behind the other two doors there was a gag prize (say, a goat). The contestant would select a door. Then the host Monty Hall would open one of the other two doors. Of course, there would always be at least one of the two remaining doors that had a gag prize behind it and *that* door was always the one that Monty opened.

After exposing that one of the unselected doors had a gag prize, the contestant was then always offered the opportunity to stick with their original choice or to change their original choice to the other, as yet unexposed, door. Can the contestant improve their chance of winning by changing their choice of doors? Justify your answer mathematically.

3. A particular disease is very rare, infecting only one person in one thousand in a population. There is a test for the disease that is very accurate. The probability that the test is positive given that someone has the disease is 99%, while the chance that the test is positive given that someone does *not* have the disease is only 2%. A person is selected at random from the population and tests positive for the disease. What are the chances that the person is really infected?

Anti-spam algorithms are often based on Bayes' rule. What, if anything, does this problem suggest about these algorithms?

7. Random Variables

Generally instead of considering the outcomes $\omega \in \Omega$ directly we will apply a systematic *measurement* to each outcome. The process of assigning a measurement to an outcome can be thought of as a function

$$X : \omega \mapsto X(\omega)$$

assigning a real number $X(\omega)$ to the outcome ω. In order for our measurements $X(\omega)$ to be useful, we generally need to be able to compute probabilities associated with outcomes of the form

$$\{\omega \in \Omega : X(\omega) \leq x\}$$

for any real number x. Thus, we only consider functions $X : \Omega \longrightarrow \mathbb{R}$ for which the above set is an event, i.e., for which

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{E}$$

7.1. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space. A random variable is a function $X : \Omega \longrightarrow \mathbb{R}$ for which

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{E}$$

for each real number x. The state space for X is the range of X in $\mathbb{R}$, i.e., the state space is $X(\Omega)$.

7.2. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let $X : \Omega \longrightarrow \mathbb{R}$ be a random variable defined on Ω. If the state space of X is either finite

$$X(\Omega) = \{x_1, x_2, \cdots x_N\}$$

or if the state space is countably infinite

$$X(\Omega) = \{x_1, x_2, \cdots x_N, \cdots\}$$

then the random variable X is said to be discrete.

If X is a random variable, then for each real number x we will write $\mathfrak{Pr}\,(X \le x)$ as short-hand for

$$\mathfrak{Pr}\,(X \le x) = \mathfrak{Pr}\,(\{\omega \in \Omega : X(\omega) \le x\})\,.$$

7.3. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let X be a random variable defined on $(\Omega, \mathcal{E}, \mathfrak{Pr})$. Then the distribution function of X is the function

$$F_X(x) = \mathfrak{Pr}\,(X \le x)$$

defined for $x \in \mathbb{R}$.

Let $F : \mathbb{R} \longrightarrow \mathbb{R}$ be a real-valued function defined on $\mathbb{R}$. For each $x \in \mathbb{R}$ define

$$F(x+) = \lim_{h \downarrow 0} F(x+h)$$

and

$$F(x-) = \lim_{h \uparrow \ 0} F(x+h)$$

provided that the limits exist.

7.4. Corollary.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space and let X be a random variable defined on $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ having distribution function $F_X(x)$. Then
(i) $0 \leq F_X(x) \leq 1$ for all $x \in \mathbb{R}$;
(ii) F_x is a non-decreasing function;
(iii) $\lim_{x \to +\infty} F_X(x) = 1$;
(iv) $\lim_{x \to -\infty} F_X(x) = 0$; and
(v) both $F_X(x+)$ and $F_X(x-)$ exist for all $x \in \mathbb{R}$ and for all $x \in \mathbb{R}$

$$\mathfrak{Pr}\,(X \leq x) = F_X(x) = F_X(x+) \quad \text{and} \quad \mathfrak{Pr}\,(X < x) = F_X(x-).$$

Proof. We will prove only that $F_X(x+)$ exists and equals $F_X(x)$, the other proofs being similar.

Let $h > 0$ be arbitrary. Note that for the first conclusion it suffices to show that

$$\lim_{n \to \infty} F_X\left(x + \frac{h}{n}\right) = F_X(x).$$

For any natural number n set

$$E_n = \{\omega \in \Omega : X(\omega) \leq x + \frac{h}{n}\}.$$

Then $\{E_n\}$ satisfies $E_1 \supseteq E_2 \supseteq E_3 \supseteq \cdots$. Thus

$$\begin{aligned} F_X(x) &= \mathfrak{Pr}\,(X \leq x) \\ &= \mathfrak{Pr}\,(\cap_n E_n) \\ &= \lim_{n \to \infty} \mathfrak{Pr}\,(E_n) \\ &= \lim_{n \to \infty} F_X\left(x + \frac{h}{n}\right) \end{aligned}$$

which proves the desired conclusion.

∎

Of course F_X is continuous at x precisely when $F_X(x+) = F_X(x-)$. If X is a discrete random variable then F_X will be discontinuous at every x_i in the state space. Indeed,

$$\mathfrak{Pr}\,(X = x) = F_X(x+) - F_X(x-)$$

for all $x \in \mathbb{R}$.

The conclusions of the corollary describe the basic properties of distribution functions. Indeed, it is possible to establish the following theorem.

7.5. Theorem.

Suppose that $F : \mathbb{R} \longrightarrow \mathbb{R}$ satisfies
(i) $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$;
(ii) F is a non-decreasing function;
(iii) $\lim_{x\to+\infty} F(x) = 1$;
(iv) $\lim_{x\to-\infty} F(x) = 0$; and
(v) both $F(x+)$ and $F(x-)$ exist for all $x \in \mathbb{R}$ and $F(x) = F(x+)$ for all $x \in \mathbb{R}$.
Then there is a probability space $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ and a random variable X defined on $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ with $F_X(x) = F$.

7.6. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space and let X be a random variable defined on $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$. The random variable X is said to be **continuous** *if $F_X(x)$ is continuous.*

7.7. Corollary.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space and let X be a random variable defined on $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$. Then X is continuous if and only if $\mathfrak{Pr}\,(X = x) = 0$ for all $x \in \mathbb{R}$.

With the above definition the notions of ‘continuous’ and ‘discrete’ random variables are not collectively exhaustive. For example, suppose that F(x) is defined by

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 0.5 \\ 1 & 0.5 \leq x \end{cases}$$

Then X satisfies properties *(i)-(v)* above and hence is a distribution function for some random variable X defined on a probability space $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$. Since the state space for

X is the set

$$[0, 0.5) \cup 1$$

this random variable is not discrete. However, neither is X continuous since $F_X(x)$ is discontinuous at $x = 0.5$.

7. Random Variables: Problems.

1.

8. Density Functions

This section will deal primarily with discrete random variables, although we will briefly discuss the concept of a density function for a continuous random variable at the end of the section.

8.1. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *be a probability space and let* $X : \Omega \longrightarrow \mathbb{R}$ *be a discrete random variable defined on* Ω*. The probability density function for* X *is*

$$f_X(x) = F_X(x+) - F_X(x-).$$

Since
$$f_X(x) = \mathfrak{Pr}(X \leq x) - \mathfrak{Pr}(X < x) = \mathfrak{Pr}(X = x)$$
we can think of $f_X(x)$ as describing the "infinitesimal" probability at state $x \in \mathbb{R}$. Notice that if the state space of X is finite
$$\mathcal{S} = \{x_1, x_2, \cdots, x_n\}$$
or infinite
$$\mathcal{S} = \{x_1, x_2, \cdots, x_n, \cdots\}$$
then
$$f_X(x) = \begin{cases} \mathfrak{Pr}(X = x_k) & \text{if } x = x_k \\ 0 & \text{otherwise} \end{cases}$$
Further notice that
$$\sum_{x_k \in \mathcal{S}} f_X(x_k) = 1$$
Density functions can be particularly useful in calculating probabilities.

8.2. Example.

Suppose that an unbiased coin is flipped until the first head appears. Let X *be the random variable that counts when the first head appears. Find the probability density function for* X *and use it to find the probability that* X *is odd.*

Solution. Note that the state space of X is

$$\mathcal{S} = \{1, 2, 3, \cdots\}.$$

The probability of a head on the first coin flip is $\frac{1}{2}$, so

$$f_X(1) = \frac{1}{2}.$$

If the first head occurs on the second flip, then the first must have been a tail, so

$$f_X(2) = \frac{1}{2}^{2}.$$

Similarly,

$$f_X(n) = \frac{1}{2}^{n-1}.$$

Now the probability that X is odd is

$$\begin{aligned}
\mathfrak{Pr}\left(X \text{ is odd}\right) &= \sum_{k \text{ odd}} \mathfrak{Pr}\left(X = k\right) \\
&= \sum_{k=0}^{\infty} \mathfrak{Pr}\left(X = 2k+1\right) \\
&= \sum_{k=0}^{\infty} \frac{1}{2}^{2k+1} \\
&= \frac{1}{2} \sum_{k=0}^{\infty} \frac{1}{4}^{k} \\
&= \frac{1}{2} \frac{1}{1 - \frac{1}{4}} \\
&= \frac{2}{3}
\end{aligned}$$

∎

Knowing the density function for a discrete random variable reduces the problem of finding probabilities to one of calculating sums. Since sums of real numbers are more concrete (and generalizable) than particular probability spaces, this is an enormously powerful tool.

The first conclusion in the theorem below follows readily from the definitions. The converse conclusion highlights the fundamental connection between density functions and probability theory.

8.3. Theorem.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let $X : \Omega \longrightarrow \mathbb{R}$ be a discrete random variable with probability density function f and state space $\mathcal{S}$. Then
(i) $f(x) \geq 0$ for all x;
(ii) $\sum_{x \in \mathcal{S}} f(x) = 1$;
(iii) $f(x) > 0$ if and only if $x \in \mathcal{S}$.
Conversely, if $\mathcal{S} \subset \mathbb{R}$ is a finite or countably infinite set and if $f : \mathbb{R} \longrightarrow [0, 1]$ is a function satisfying (i)-(iii) above, then there is a probability space $(\Omega, \mathcal{E}, \mathfrak{Pr})$ and a discrete random variable X defined on Ω having state space $\mathcal{S}$ and density function f.

For the first conclusion see the exercises. While not especially difficult to prove, we only outline the proof of the converse conclusion and omit the details. We take

$$\Omega = \mathcal{S}$$

and $\mathcal{E}$ to be the collection of all subsets of $\mathcal{S}$. For $A \subseteq \mathcal{E}$ we define

$$\mathfrak{Pr}(A) = \sum_{x \in A} f(x).$$

It is routine to show that $(\Omega, \mathcal{E}, \mathfrak{Pr})$ is a probability space. If $X : \Omega \longrightarrow \mathbb{R}$ is defined by

$$X(\omega) = \omega$$

then the rest of the conclusions follow.

8.4. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let $X : \Omega \longrightarrow \mathbb{R}$ and $Y : \Omega \longrightarrow \mathbb{R}$ be a discrete random variables defined on Ω. Then the joint density function *of X and Y is*

$$f_{XY}(x, y) = \mathfrak{Pr}(X = x \quad \text{and} \quad Y = y)$$

Notice that if X has state space $\mathcal{S}_X$ and Y has state space $\mathcal{S}_Y$ then

joint density function!discrete random variables

independent!discrete random variables

$$f_{XY}(x,y) = \begin{cases} \mathfrak{Pr}\,(X = x \quad \text{and} \quad Y = y) & \text{if } x \in \mathcal{S}_X \text{ and } y \in \mathcal{S}_Y \\ 0 & \text{otherwise} \end{cases}$$

8.5. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space and let $X : \Omega \longrightarrow \mathbb{R}$ and $Y : \Omega \longrightarrow \mathbb{R}$ be a discrete random variables defined on Ω. We say that X and Y are independent if

$$\mathfrak{Pr}\,(X = x \quad \textit{and} \quad Y = y) = \mathfrak{Pr}\,(X = x)\,\mathfrak{Pr}\,(Y = y)$$

for all $x, y \in \mathbb{R}$.

The proof of the following proposition is immediate from the definitions.

8.6. Proposition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space and let $X : \Omega \longrightarrow \mathbb{R}$ and $Y : \Omega \longrightarrow \mathbb{R}$ be a discrete random variables defined on Ω. Then X and Y are independent if and only if

$$f_{XY}(x,y) = f_X(x) f_Y(y)$$

for all $x, y \in \mathbb{R}$.

Thus discrete random variables X and Y are independent if and only if

$$f_{XY}(x,y) = \begin{cases} f_X(x) f_Y(y) & \text{if } x \in \mathcal{S}_X \text{ and } y \in \mathcal{S}_Y \\ 0 & \text{otherwise} \end{cases}$$

For a continuous random variable

$$F_X(x+) - F_X(x-) = 0$$

for all x. Thus notion of a density function is somewhat more complex, replacing summations with integrals. Worse, it turns out that not every continuous random variable has a density function.

Suppose that X is a continuous random variable and suppose that the distribution function $F_X(x)$ is continuously differentiable, i.e., suppose that $F'_X(x)$ exists and is continuous. Then by the fundamental theorem of calculus

Lebesgue, Henri

$$F_X(b) - F_X(a) = \int_a^b F'_X(x)\, dx$$

and, in view of *(iv)* above,

$$F_X(x) = \int_{-\infty}^x F'_X(t)\, dt.$$

If F has a continuously differentiable distribution function, then the probability of any event associated with X can be calculated in terms of the above integral and without reference to the underlying probability space. This is an extraordinarily useful observation in that it reduces the abstract problem of finding probabilities to the more mundane one of calculating integrals. For this reason we generally only consider random variables X for which F'_X exists.

8.7. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let X be a continuous random variable defined on $(\Omega, \mathcal{E}, \mathfrak{Pr})$. Suppose that the distribution function $F_X(x)$ for X is everywhere differentiable with a continuous derivative $F'_X(x)$. Then the density function of X is

$$f_X(x) = F'_X(x).$$

In this case we say that X is an absolutely continuous random variable.

Clearly the above definition only applies to a subset of the continuous random variables X. It is possible, with some ingenuity, to construct continuous distribution functions that are not differentiable. Interestingly, since F_X is non-decreasing, F_X must be differentiable (but not necessarily continuously differentiable) everywhere except on a set E that has the property that

$$\int_E 1\, dt = 0.$$

This important result is due to the French mathematician Henri Lebesgue. Sets E of the above type are said to have (Lebesgue) measure zero and are of particular importance in understanding integrals and probability. It is possible – but difficult – to show that the absolutely continuous random variables are those that map events of probability zero to sets of measure zero.

We defer a discussion of joint densities and independence for absolutely continuous random variables to a later section.

8. Density Functions: Problems.

1.

9. Examples of Discrete Random Variables

If X is a discrete random variable, then the range of X is either a finite

$$x_1, x_2, \cdots, x_n$$

or countably infinite

$$x_1, x_2, \cdots, x_n, \cdots$$

set of real numbers. For each x_j in the range of X we could take

$$p_j = \mathfrak{Pr}\,(X = x_j) = \mathfrak{Pr}\,(\{\omega \in \Omega | X(\omega) = x_j\}).$$

Then clearly each $p_j \geq 0$ and

$$\sum_j p_j = 1.$$

The density function for X can then be realized as

$$f_X(x) = = \mathfrak{Pr}\,(X = x)$$
$$\begin{cases} p_j & \text{if } x = x_j \\ 0 & \text{otherwise} \end{cases}$$

Almost all random variables that arise in applications either have range contained in the non-negative integers or are constructed from random variables having range contained in the non-negative integers. Thus in this section almost every example will have range

$$1, 2, 3, \cdots, n$$

or

$$1, 2, 3, \cdots, n, \cdots$$

If one knows the density function for a discrete random variable then one also knows the distribution function and vice versa. Thus we will refer to random variables that have the same density function as having a common *distribution*. Since density functions are the basic building blocks in actually calculating probabilities associated with random variables, when discussing specific examples of random variables the focus is almost always on the density function rather than the distribution function even though the word "distribution" is used to describe classes of random variables having the same density/distribution functions.

9.1. Example. Uniform Distribution.

Suppose that a number is randomly selected from the $\{1, 2, \cdots, n\}$*. If each number is equally likely to be selected and if* X *is the number selected, then the density function for* X *is*

$$f_X(x) = \begin{cases} \frac{1}{n} & \text{if } x = 1, 2, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$

This is the simplest distribution, corresponding to the Cardano Counting Axiom.

9.2. Example.

Suppose that X *is a discrete random variable that assumes exactly two values,* 0 *and* 1*, i.e, suppose that* $X(\Omega) = \{0, 1\}$*. If*

$$p = \mathfrak{Pr}(X = 1) \quad \text{and} \quad q = 1 - p = \mathfrak{Pr}(X = 0)$$

Then X *is said to be a* $Bernoulli$ *random variable. Generally* p *is said to be the probability of* $success$ *and* $q = 1 - p$ *is the probability of* $failure$*.*

The most familiar example of a Bernoulli random variable would be flipping an unbiased coin, with

$$X = \begin{cases} 1 & \text{coin flip is heads} \\ 0 & \text{otherwise} \end{cases}$$

In this case $p = q = 0.5$. For an example when $p \neq q$, suppose that we roll a pair of unbiased dice and

$$X = \begin{cases} 1 & \text{if we roll doubles} \\ 0 & \text{otherwise} \end{cases}$$

Then X is a Bernoulli random variable with $p = \frac{1}{6}$ and $q = \frac{5}{6}$.

In many situations we will perform repeated and independent Bernoulli trials, such as flipping an unbiased coin repeatedly.

9.3. Definition.

A finite or infinite sequence of random variables $\{X_i\}$ is said to be a sequence of independent Bernoulli trials if

(i) Each X_i is a Bernoulli Random Variable having common probability of success p;

(i) For each k The family of random variables $\{X_1, X_2, \cdots, X_k\}$ is independent.

9.4. Example.

Suppose that we flip an unbiased coin n times and let X be the random variable that counts the number of heads. Find the density function for X.

Solution. Since we have flipped the coin n times and each flip has two outcomes, H and T, there are 2^n possible outcomes. Now if $X = k$ then we must have k heads and $n - k$ tails. Thus to calculate how many of the 2^n outcomes satisfies $X = k$ we must calculate how many different ways we can assign k "heads" (and hence $n - k$ tails) to the n flips. This amounts to selecting k numbers from $\{1, 2, \cdots, n\}$ to be "heads." Since the numbers are selected without replacement and since the order in which they are selected does not matter, this is exactly the number of combinations of n things taken k at a time

$$\binom{n}{k}.$$

Thus applying the counting principle,

$$\mathfrak{Pr}\,(X = k) = \binom{n}{k}\left(\frac{1}{2}\right)^k.$$

■

More generally, we have the following.

9.5. Example.

Suppose that we have n independent Bernoulli trials $\{X_1, X_2, \cdots, X_n\}$ having common probability of success p. If X is the random variable that counts the number of successes in n trials, i.e., if

$$X = \sum_{i=1}^{n} X_i$$

then the density function for X is

$$f_X(x) = \begin{cases} \binom{n}{k} p^k q^{n-k} & \text{if } k = 0, 1, \cdots n \\ 0 & \text{otherwise} \end{cases}$$

In this case X is called a $binomial\ random\ variable$.

The above can be deduced in a manner similar to the previous example – see the problems. In order to be a density function, we must of course have

$$\sum_{k=0}^{n} f_X(k) = 1.$$

We can see that this is indeed the case by applying the binomial theorem:

$$\begin{aligned} \sum_{k=0}^{n} f_X(k) &= (p+q)^n \\ &= 1^n \end{aligned}$$

motivating the name of the random variable.

9.6. Example.

Let $\{X_i\}$ be an infinite sequence of independent Bernoulli trials having common probability of success p. Let X be the random variable that counts the number of trials until the first success occurs. Find the density function for X. The random variable X is said to have the $geometric$ distribution.

Solution. Since we are counting the number of trials *until* the first success, the state space for X is $\{0, 1, 2, 3, \cdots\}$. (In similar examples up until now, we have counted the trial on which the first success occured; for technical reasons that will become obvious later it is preferable to count the number of trials until the first success.)

If $X = 0$ then the very first trial must have been a success, so

$$f_X(0) = \mathfrak{Pr}\,(X = 0) = p.$$

If $X = k$ and $k > 0$ then the first k trials must have been failures followed by a success, so

$$f_X(k) = pq^k.$$

Thus in general

$$f_X(k) = \begin{cases} pq^k & k = 1, 2, \cdots \\ 0 & \text{otherwise} \end{cases}$$

∎

9.7. Example. Rolls of a die.

Suppose we roll a fair die n times. Then there are 6^n possible distinct outcomes.

Generally we are not interested in finding the probability of any particular one of the 6^n outcomes but are interested instead in summary outcomes such as counting the number of times each of the six possible outcomes occurs. For example if

$$m_1, m_2, m_3, m_5, m_5, m_6$$

represent the number of times each roll occurs then for each $i = 1, 2, 3, 4, 5, 6$

$$0 \leq m_i \leq n$$

and

$$m_1 + m_2 + m_3 + m_4 + m_5 + m_6 = n.$$

If we have observed a particular set rolls that result in counts

$$m_1, m_2, m_3, m_5, m_5, m_6$$

then we have observed m_1 rolls of a "1," m_2 rolls of a "2" and so on. Since the die is fair, this happens with probability

$$\frac{1}{6}^{m_k}.$$

Thus if we have a particular set of n rolls that result in counts of

$$m_1, m_2, m_3, m_5, m_5, m_6$$

the chances of this particular set of rolls is

$$\frac{1}{6}^{m_1}\frac{1}{6}^{m_2}\frac{1}{6}^{m_3}\frac{1}{6}^{m_4}\frac{1}{5}^{m_1}\frac{1}{6}^{m_1} = \frac{1}{6}^{n}. \tag{9.1}$$

This is of course consistent with notion that there are 6^n outcomes and that each is equally likely.

A more challenging question is *how many of the 6^n outcomes result in counts of*

$$m_1, m_2, m_3, m_5, m_5, m_6?$$

However we have already discussed this problem: this is exactly the question that was answered in example 4.6. The answer is

$$\frac{n!}{m_1!m_2!\cdots m_6!}.$$

Thus the probability of observing a particular set of counts

$$m_1, m_2, m_3, m_5, m_5, m_6$$

is

$$\frac{n!}{m_1!m_2!\cdots m_6!}\frac{1}{6}^{n}$$

This is a particular case of the multinomial distribution.

9.8. Example. Multinomial Distribution.

Suppose that an experiment has r possible outcomes $\{\omega_1, \omega_2, \cdots \omega_r\}$. An example might be rolling a die which has six possible outcomes. The probability of the j^{th} outcome is assumed to be p_j, i.,e.,

$$\mathfrak{Pr}\left(\{\omega_j\}\right) = p_j.$$

Of course

$$\sum_{j=1}^{r} p_j = 1.$$

Now suppose that we repeat this experiment n times with the repetitions being independent. In the language of random variables, this means that we have n independent random variables

$$Y_1, Y_2, \cdots, Y_n$$

with

$$Y_i(\omega_j) = j$$

and

$$\mathfrak{Pr}\left(Y_i = j\right) = p_j.$$

As in the roll of dice, we are interested in the number of times each of the r outcomes occurs, i.e., in the random variables $X_1, X_2, \cdots, X_r$ where

$$X_k = \textit{number of times } \{Y_i\} \textit{ assumes the value } k.$$

so that the range of each X_k is $\{0, 1, \cdots, n\}$ and

$$X_1 + X_2 + \cdots X_r = n.$$

The multinomial distribution describes the joint density of the random vector $(X_1, X_2, \cdots, X_r)$. Equivalently, for a possible set of observed counts

$$m_1, m_2, \cdots, m_r$$

the multinomial distribution describes

$$\mathfrak{Pr}\left(X_1 = m_1, X_2 = m_2, \cdots, X_r = m_r\right)$$

Our goal is to develop a closed form for the above probability.

Solution. Notice that $0 \leq m_i \leq n$ for each i and that

$$m_1 + m_2 + \cdots m_r = n.$$

Now if $Y_1, Y_2, \cdots, Y_n$ is a particular set of outcomes that result in counts of

$$m_1, m_2, \cdots, m_r$$

then by independence we see that the joint probability of the $\{Y_i\}$ is

$$p_1^{m_1} \cdot p_2^{m_2} \cdots p_r^{m_r}.$$

This is equivalent to (1) above. Applying example 4.6, among the n^r possible outcomes, those that result in counts of

$$m_1, m_2, \cdots, m_r$$

must number

$$\frac{n!}{m_1! m_2! \cdots m_r!}$$

and hence occur with probability

$$\frac{n!}{m_1! m_2! \cdots m_r!} p_1^{m_1} \cdot p_2^{m_2} \cdots p_r^{m_r}.$$

This last represents exactly

$$\mathfrak{Pr}\left(X_1 = m_1, X_2 = m_2, \cdots, X_r = m_r\right).$$

A random vector $(X_1, X_2, \cdots, X_r)$ having the above distribution is said to have the *multinomial distribution.*

■

9.9. Example. Poisson Distribution.

Suppose that $\lambda \in \mathbb{R}$ *and that*

$$f_X(k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & k = 0, 1, 2, \cdots \\ 0 & \textit{otherwise} \end{cases}$$

Then f_X *satisfies (i)-(iii) of Theorem 8.3 with* $\mathcal{S} = \mathbb{N}$.

A random variable X having the above density function is said to have the Poisson distribution. There are many settings in which the distribution of X can be empirically demonstrated to approximate a Poisson distribution. Some of these include the number of misprints on the pages of a book, the number of calls arriving per unit time in a telephone switch and the number of atoms of a radioactive substance that disintegrate per unit time. The Poisson distribution plays an important role in random processes that evolve over time (stochastic processes) that will be discussed in detail later in this text.

The Poisson distribution also has important connections with the binomial distribution. Suppose, for example, we are studying the arrival of phone calls at a phone switch. We might divide each hour up in n equal subdivisions where n is chosen so large that we are confident two calls will not arrive in any subdivision. If we assume that the calls are equally likely to arrive in any interval then we might set

$$p_n = \mathfrak{Pr}\,(\text{call arrives in one of the } n \text{ intervals})\,.$$

Clearly as n gets larger p_n would get smaller. However, empirical studies of phone networks suggest that there is a number λ such that

$$\lim_n np_n = \lambda.$$

Thus for large values of n we may assume that

$$p_n \approx \frac{\lambda}{n}.$$

Now if S_n is the random variable that counts the number of successes (phone calls arriving) in n trials, then S_n is a binomial random variable with

$$\begin{aligned}
\mathfrak{Pr}\,(S_n = k) &= \binom{n}{k}(p_n)^k(1-p_n)^{n-k} \\
&\approx \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}.
\end{aligned}$$

Letting n go to infinity results in

$$\mathfrak{Pr}\,(S_n = k) \approx \frac{\lambda^k}{k!}e^{-\lambda}$$

i.e., that S_n is approximately a Poisson random variable for large values of n (which correspond to small slices of the time interval). There is considerable data available for phone networks validating this model, although the applicability to data networks is less clear.

Poisson approximation to binomial

In any case, the above arguments show that the Poisson can be used to approximate a binomial distribution.

9.10. Theorem.

Suppose that $0 \leq p_n \leq 1$ for each n and in addition $\{p_n\}$ satisfies

$$\lim_{n\to\infty} np_n = \lambda$$

for some $\lambda > 0$. Then for each $k = 0, 1, 2, \cdots$

$$\lim_{n\to\infty} \binom{n}{k} (p_n)^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

This result enables one to approximate binomial calculations with Poisson sums.

9.11. Example. Poisson Approximations.

Suppose that 1% of all emails arriving in my mailbox are not spam. If I currently have 200 emails in my inbox, what is the probability that they are all spam?

Solution. In this case we would have $n = 200$ Bernoulli trials with probability of success $p = 0.01$. The probability that all of the email is spam is therefore

$$(1 - .01)^{200} = 0.1340.$$

The Poisson approximation is given by

$$e^{-200(0.01)} = e^{-2} = 0.1353.$$

∎

As the above example shows, the Poisson approximation to the binomial is reasonably accurate. However, since the terms from the binomial density can be readily calculated directly, using the approximation as a labor saving device is less significant than the fact that certain Poisson distributions – such as those for phone networks – arise as the limit binomial random variables.

Another way in which Bernoulli trials arise in waiting times to the r^{th} success.

9.12. Example. Waiting times.

Suppose that we consider a sequence of Bernoulli trials with probability of success p. While a geometric random variable counts when the first success occurs, suppose instead we are interested in when the r^{th} success occurs. One approach might be to use the random variable Y which counts the total number of trials until the r^{th} success. Then Y can assume values

$$r, r+1, r+2, \cdots$$

since there must be at least r trials to observe r successes. Algebraically it will be slightly easier to deal with the random variable $X = Y - r$ that can assume the values

$$0, 1, 2, 3, \cdots.$$

The random variable $X = Y - r$ counts the number of trials $k \geq r$ needed for the r^{th} success to occur. Our problem is to find the density function for X.

Solution. In order for the r^{th} success to occur on the $(r+k)^{th}$ trial, we must have observed

(a) exactly $r-1$ successes and k failures in the first $r+k-1$ trials, followed by

(a) a success on the $(r+k)^{th}$ trial.

The probability of (a) is exactly (see example 4.5)

$$\binom{r+k-1}{k} p^{r-1}(1-p)^k$$

while the probability of (b) is p. Hence

$$\mathfrak{Pr}\,(X = k) = \binom{r+k-1}{k} p^r (1-p)^k. \tag{9.2}$$

∎

While the reasoning leading to the above seems clear, it is not necessarily obvious that

$$\sum_{k=0}^{\infty} \mathfrak{Pr}\,(X = k) = \sum_{k=0}^{\infty} \binom{r+k-1}{k} p^r (1-p)^k = 1$$

which is of course required for a density function. However, if one expands

$$(1-t)^{-r}$$

in a Taylor's series, then it follows that density function does indeed sum to one (see the problems at the end of this section).

The preceding example is a special case of the *negative binomial* distribution. The negative binomial usually involves a generalization of the notion of combinations: if α is any positive real number and if k is any non-negative integer then we define

$$\binom{-\alpha}{k} = \frac{(-\alpha)(-\alpha-1)(-\alpha-2)\cdots(-\alpha-k+1)}{k!}.$$

9.13. Example. Negative Binomial Distribution.

Suppose that α is a positive real number and $0 < p < 1$. A random variable X having density function

$$f_x(k) = \begin{cases} \binom{-\alpha}{k} p^{\alpha} (-1)^k (1-p)^k & k = 0, 1, \cdots \\ 0 & \textit{otherwise} \end{cases}$$

is said to have a negative binomial distribution. (The reason for the name "negative binomial" is because of the similarity to the coefficients of a binomial random variable.)

To see the connection with the preceding example, note that

$$\begin{aligned} \binom{-\alpha}{k} &= \frac{(-\alpha)(-\alpha-1)\cdots(-\alpha-k+1)}{k!} \\ &= (-1)^k \frac{(\alpha)(\alpha+1)\cdots(\alpha+k-1)}{k!} \\ &= (-1)^k \binom{\alpha+k-1}{k} \end{aligned}$$

Thus if $k = 0, 1, \cdots$ then

$$
\begin{aligned}
f_x(k) &= \binom{-\alpha}{k} p^\alpha (-1)^k (1-p)^k \\
&= \binom{\alpha + k - 1}{k} p^\alpha (1-p)^k
\end{aligned}
$$

which agrees with † when $\alpha = r$. Similarly, using the Taylor's series for $(1-t)^{-\alpha}$ one can show that the above is indeed a density function.

If X is a geometric random variable, then $Y = X - 1$ is a negative binomial random variable with $\alpha = 1$ and p the probability of success for any trial.

9. Examples of Discrete Random Variables: Problems.

1. Suppose that a box contains ten balls numbered $1, 2, \cdots, 10$. Suppose that there are two independent repetitions of the experiment of selecting a ball from the box and recording the number (i.e., the ball is replaced after each selection and the balls are randomized before the next selection). Let X be the random variable that records the larger of the two numbers. What is the density function for X?

2. Repeat the previous problem, but suppose instead that the balls are selected without replacement.

3. Show that

$$\sum_{k=0}^{\infty} \binom{r+k-1}{k} p^r (1-p)^k = 1$$

where $0 < p < 1$ and $r \geq 0$ is an integer.

4. Let X be a geometric random variable with parameter p. Let N be a fixed integer and define the random variable Y by

$$Y = \min\{X, N\}.$$

Find the density function for Y.

5. Let X and Y be independent random variables having the uniform distribution on the set $\{1, 2, \cdots, n\}$.

(a) Find the density for $W = \min\{X, Y\}$.
(b) Find the density function for $W = X + Y$.
(c) Find the density function for $W = \max\{X, Y\}$.
(d) Find the density function for $W = |X - Y|$.

6. Let X_1 and X_2 be independent Poisson random variables with parameters λ_1 and λ_2 respectively. For a fixed $z \geq 0$ Find

$$\mathfrak{Pr}\left(Y = y | X + Y = z\right)$$

for $y = 0, 1, \cdots$.

7. Suppose that 2% of all memory chips from a manufacturer will be defective. Use the Poisson approximation to estimate the probability that a shipment of 1000 chips will have at most 30 defectives.

10. Examples of Continuous Random Variables

Continuous random variables that arise in applications will most frequently be absolutely continuous, i.e., have a probability density functions. While this is not uniformly the case, in this section we will consider only those continuous random variables that have density functions. Given a random variable X having density function f_x, then the distribution function

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

can be calculated from the density function and since

$$F_X'(x) = f_X(x)$$

the density can be calculated from the distribution function. As with discrete random variables, we will refer to random variables that have the same density (equivalently the same distribution) as having a common distribution, even though we almost always refer to the density function rather than the distribution function.

10.1. Example. Uniform Distribution.

Let $a < b$ be two real numbers. Let X be the random variable that records the outcome when a real number is randomly selected from the interval $[a, b]$. Then X has a uniform distribution with density function

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \textit{if } a \le x \le b \\ 0 & \textit{otherwise} \end{cases}$$

random variables, continuous!sum of exponential

10.2. Example. Exponential Distribution.

Let $\lambda > 0$ be a real number and define

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \textit{if } x \geq 0 \\ 0 & \textit{otherwise} \end{cases}$$

If X is a random variable having density function f then X is said to have an exponential distribution.

Exponential random variables arise in numerous applications, including the lifespan of electrical components, the decay of radioactive isotopes, the biomass of bacteria in a culture and the "service times" for certain queues such as those arising in telephony networks.

When we construct new random variables from continuous random variables, elementary computations using integrals often reveal properties of the density of the new random variable. By way of example, consider the following.

10.3. Example.

Suppose that X has an exponential distribution with parameter λ. If $Y = X^2$ what is the density function of Y?

Solution. We can start by finding the distribution function for Y; for $y > 0$

$$\begin{aligned} F_Y(y) &= \mathfrak{Pr}\,(Y \leq y) \\ &= \mathfrak{Pr}\,(X^2 \leq y) \\ &= \mathfrak{Pr}\,(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \mathfrak{Pr}\,(0 \leq X \leq \sqrt{y}) \\ &= \int_0^{\sqrt{y}} \lambda e^{-\lambda t}\,dt \\ &= e^{-\lambda t}\Big|_{t=0}^{\sqrt{y}} \\ &= \left(1 - e^{-\lambda\sqrt{y}}\right) \end{aligned}$$

Then the density function for Y is

$$\begin{aligned} f_Y(y) &= F_Y'(y) \\ &= \frac{\lambda}{2\sqrt{y}} e^{-\lambda\sqrt{y}} \end{aligned}$$

∎

10.4. Example.

Suppose that X is an absolutely continuous random variable having density function $f_X(x)$ and set

$$Y = aX + b$$

where a and b are real numbers. Then Y is an absolutely continuous random variable having density function

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$$

Proof. Note that

$$\begin{aligned} \mathfrak{Pr}\,(Y \le y) &= \mathfrak{Pr}\,(aX + b \le y) \\ &= \mathfrak{Pr}\left(X \le \frac{y-b}{a}\right) \end{aligned}$$

and hence

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\mathfrak{Pr}\,(Y \le y) \\ &= \frac{d}{dy}\mathfrak{Pr}\left(X \le \frac{y-b}{a}\right) \\ &= \frac{1}{a} f_X\left(\frac{y-b}{a}\right) \end{aligned}$$

using the chain rule and the fact that

$$\frac{d}{dx}\mathfrak{Pr}\,(X \le x) = f_X(x).$$

∎

More generally we can deduce the following. A function is *monotone* if it is either non-decreasing or non-increasing. A function is *strictly monotone* if it is either strictly increasing or strictly decreasing. Note that strictly monotone function φ must have an inverse function, i.e., there is a function φ^{-1} so that

$$x = \varphi^{-1}(\varphi(x))$$

Recall also that from the inverse function theorem φ is differentiable if and only if φ^{-1} is differentiable.

10.5. Theorem.

Let φ be a continuous strictly monotone function having derivative φ'. Suppose that X is an absolutely continuous random variable having density function f_X. If $Y = \varphi(X)$ is absolutely continuous, then Y has density function

$$g_Y(y) = f_X(\varphi^{-1}(y))(\varphi^{-1})'(y).$$

10.6. Example. Standard Normal Distribution

A random variable Z is said to have a standard normal distribution if Z has density function

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

It may not be immediately apparent that the above defines a density function, i.e., that

$$\int_{-\infty}^{\infty} f_Z(z)\, dz = 1.$$

To see why this is the case, we set $\gamma = \int_{-\infty}^{\infty} f_Z(z)\, dz$ introduce polar coordinates as follows:

$$\begin{aligned}
\gamma^2 &= \left(\int_{-\infty}^{\infty} f_Z(x)\,dx\right)\left(\int_{-\infty}^{\infty} f_Z(y)\,dy\right)\\
&= \int_{-\infty}^{\infty} f_Z(x)\int_{-\infty}^{\infty} f_Z(y)\,dy\,dx\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{\frac{-x^2}{2}} e^{\frac{-y^2}{2}}\,dy\,dx\\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-(x^2+y^2)}{2}}\,dy\,dx\\
&= \frac{1}{2\pi}\int_{0}^{\infty}\int_{-\pi}^{\pi} e^{\frac{-r^2}{2}} r\,d\theta\,dr\\
&= \int_{0}^{\infty} re^{\frac{-r^2}{2}}\,dr\\
&= e^{\frac{-r^2}{2}}|_{r=0}^{\infty}\\
&= 1.
\end{aligned}$$

Thomson, William

Kelvin, Lord!see Thomson, William

Thus

$$\int_{-\infty}^{\infty} e^{\frac{-t^2}{2}}\,dt = \sqrt{2\pi}$$

as desired.

∎

William Thomson – Lord Kelvin – would routinely pepper his lectures with complex mathematics, often leaving out crucial steps. His students requested that he include more mathematical details in his lectures and he promised to try to do better. As it happened, the next day the above formula was needed. Lord Kelvin wrote the formula on the board, commenting that it was as simple to derive as “two plus to equals four." Then, recalling his promise to his students, he turned to blackboard and wrote:

$$2 + 2 = 4.$$

Of course, Lord Kelvin could also be wrong. Using the thermodynamics of chemical processes he calculated the age of the earth to be little more than twenty million years, and hence claimed to have disproved Darwin’s theory of evolution. Lord Kelvin was also responsible for the now infamous 1895 observation that “Heavier than air flying machines are impossible." Lord Kelvin assertions notwithstanding, the fact that

$$\int_{-\infty}^{\infty} e^{\frac{-t^2}{2}}\,dt = \sqrt{2\pi}$$

is one of the less obvious ways in which the transcendental numbers e and π are related.

gamma function

Applying 10.4, if Z is a normal random variable with parameters $\mu = 0$ and $\sigma = 1$ and if $X = \sigma Z + \mu$ then X has density function

$$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

This leads to the following definition.

10.7. Example.

More generally, a random variable X is said to be normally distributed with parameters μ and $\sigma > 0$ if

$$X = \sigma Z + \mu$$

where Z is a standard normal random variable. In particular then X has density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

The parameters μ and σ have a particular meaning that will be discussed in a later section.

10.8. Example.

Suppose that X is normally distributed with parameters $\mu = 0$ and $\sigma > 0$. Let Y be the random variable $Y = X^2$; then the density function of Y is

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

In the case that $\sigma = 1$ Y is said to have the χ^2 distribution.

Proof. This is just an application of previous theorems. ∎

A more general version of the above comes from the gamma function.

10.9. Definition.

If $\alpha > 0$ then

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx.$$

The gamma function $\Gamma(\alpha)$ has many useful properties. For example, a simple integration by parts shows that

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

so that the gamma function can be thought of as a generalized factorial. Two other useful formulae are

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

and if $\alpha > 0$ and λ is an real number

$$\int_0^\infty x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\Gamma(\alpha)}{\lambda^\alpha}$$

(see the exercises).

10.10. Definition.

A random variable X is said to have the gamma distribution with parameters α and λ if X has density function

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \qquad 0 \le x$$

The exponential random variables are a special case of the gamma random variables (with $\alpha = 1$). Similarly, the square of a normal random variable X having $\mu = 0$ and $\sigma > 0$ corresponds to a gamma random variable with $\alpha = 0.5$ and

$$\lambda = \frac{1}{2\sigma^2}.$$

10. Examples of Continuous Random Variables: Problems.

1. Show that

$$f(x) = \frac{1}{\pi(1+x^2)}$$

satisfies

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

and hence that f defines a density function. Find the cumulative distribution function that corresponds to f, i.e., find

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

A random variable X having density function f is said to have the *Cauchy* distribution. This distribution will have important theoretical implications in later problems.

2. Verify

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

3. Verify

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

4. Let X be an exponential random variable having parameter λ. Show that if $a \geq 0$ and $b \geq 0$ then

$$\mathfrak{Pr}\,(X > a+b) = \mathfrak{Pr}\,(X > a)\,\mathfrak{Pr}\,(X > b).$$

5. Let X be a random variable and suppose that for any $a \geq 0$ and $b \geq 0$

$$\mathfrak{Pr}\,(X > a+b) = \mathfrak{Pr}\,(X > a)\,\mathfrak{Pr}\,(X > b).$$

Then either $\mathfrak{Pr}\,(X > 0) = 0$ or else X is exponentially distributed.

11. Jointly Distributed Discrete Random Variables

11.1. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ *be a probability space and let* X *and* Y *be random variables defined on* $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$*. Then the joint distribution function of* X *and* Y *is*

$$F_{XY}(x, y) = \mathfrak{Pr}\,(X \le x \quad \textit{and} \quad Y \le y).$$

11.2. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ *be a probability space and let* X *and* Y *be discrete random variables defined on* $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$*. If*

$$f_{XY}(x, y) \equiv \mathfrak{Pr}\,(X = x \quad \textit{and} \quad Y = y)$$

then we say that the joint density function for X *and* Y *is* $f_{XY}(x, y)$*.*

11.3. Example.

Suppose that one flips a coin and rolls a single die and the the coin flip and the roll are indendepnt. Let X *be the random variable*

$$X = \begin{cases} 0 & \textit{if the coin flip is heads} \\ 1 & \textit{if the coin flip is tails} \end{cases}$$

and Y *be the random variable*

$$Y = \quad \textit{value showing on the die}.$$

Find the joint distribution of X *and* Y*.*

Solution. Clearly for $(x, y) \in \mathbb{Z} \times \mathbb{Z}$,

$$f_{XY}(x, y) = \begin{cases} 1/12 & \text{if } (x, y) \in \{0, 1\} \times \{0, 1, \cdots\} \\ 0 & \text{otherwise} \end{cases}$$

and so

$$F_{XY}(x, y) = \begin{cases} 0 & x < 0 \\ y/12 & x = 0 \text{ and } y = 1, 2, \cdots, 6 \\ (y+6)/12 & x = 1 \text{ and } y = 1, 2, \cdots, 6 \\ 1 & x > 1 \text{ and } y > 6 \end{cases}$$

∎

The proof of the next theorem is similar to that of Corollary 7.4 and is omitted.

11.4. Theorem.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let X and Y be discrete random variables defined on $(\Omega, \mathcal{E}, \mathfrak{Pr})$ having distribution functions $F_X(x)$ and $F_Y(y)$ respectively. If $F_{XY}(x, y)$ is the joint distribution of X and Y then

(i) For each fixed $x \in \mathbb{R}$

$$\lim_{y \to -\infty} F_{XY}(x, y) = 0 \quad \text{and} \quad \lim_{y \to \infty} F_{XY}(x, y) = F_X(x)$$

(ii) For each fixed $y \in \mathbb{R}$

$$\lim_{x \to -\infty} F_{XY}(x, y) = 0 \quad \text{and} \quad \lim_{x \to \infty} F_{XY}(x, y) = F_Y(y)$$

(iii) If

$$f_{XY}(x, y) \equiv \mathfrak{Pr}\,(X = x \quad \text{and} \quad Y = y)$$

then X and Y have density functions given by

$$f_X(x) = \sum_y f_{XY}(x, y) \quad \text{and} \quad f_Y(y) = \sum_X f_{XY}(x, y)$$

the sums being finite or countably infinite since X and Y are discrete.

(iv) If $E \subseteq \mathbb{R} \times \mathbb{R}$ then

$$\mathfrak{Pr}\,((X, Y) \in E) = \sum\sum\nolimits_{(x,y) \in E} f_{XY}(x, y).$$

11.5. Theorem.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let X and Y be discrete random variables defined on $(\Omega, \mathcal{E}, \mathfrak{Pr})$ having joint density function $f_{XY}(x, y)$. If $Z = X + Y$ then the density function of Z is

$$\mathfrak{Pr}\,(Z = z) = \sum_x f_{XY}(x, z - x).$$

Proof. For fixed z

$$\begin{aligned}\mathfrak{Pr}\,(Z = z) &= \mathfrak{Pr}\,(X + Y = z)\\ &= \sum_{x} \mathfrak{Pr}\,(X = x \quad \text{and} \quad Y = z - x)\\ &= \sum_{x} f_{XY}(x, z - x)\end{aligned}$$

the sums being finite or countably infinite since both X and Y are discrete. ∎

The above result is most frequently used when X and Y are independent.

11.6. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space and let X and Y be discrete random variables defined on $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$. Then X and Y are independent *if for each $x, y \in \mathbb{R} \times \mathbb{R}$*

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

11.7. Theorem.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space and let X and Y be discrete, independent random variables defined on $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ having density functions $f_X(x)$ and $f_Y(y)$ respectively. If $Z = X + Y$ then the density function for Z is

$$f_Z(z) = \sum_{x} f_X(x) f_Y(z - x)$$

11. Jointly Distributed Discrete Random Variables: Problems.

1.

12. Expectations: Discrete Random Variables

Historically expectations arose in connection with the study of gambling, casinos, and the construction of games of chance in which the casino is assured, in the long run, of making a profit. A simple example can illustrate the concepts.

12.1. Example.

Suppose that a casino offers the following game of chance to its patrons. The patron rolls a single fair die and the casino pays the patron \$D, where D is the number showing on the face of the die. Thus on any play a patron can win any of $\{\$1, \$2, \$3, \$4, \$5, \$6\}$. For each opportunity to play this game, the casino charges \$P. The problem is to find the least entry fee \$P so that the casino does not lose money.

Solution. For this example, $\Omega = \{\$1, \$2, \$3, \$4, \$5\$6\}$ and each simple event is equally likely, having probability one sixth. It is reasonable to assume that every play of the game is independent of every other play. Thus if n patrons play the game, we have n independent outcomes. If we let

$$X_i = \text{dollar amount paid on the } i^{th} \text{ game}$$

then we have a sequence

$$X_1, X_2, \cdots, X_n$$

of n independent and identically distributed discrete random variables having common state space

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$

with

$$\mathfrak{Pr}\left(X_i = x\right) = \begin{cases} \frac{1}{6} & \text{if } x \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$$

Clearly the *average* amount that the casino pays out for these n plays of the game is

$$\frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right).$$

On the other hand, if we define

$$N(i) = \text{number of times the casino pays \$i}, \quad i = 1, 2, \cdots, 6$$

then

$$\frac{1}{n}\left(1 \times N(1) + 2 \times N(2) + \cdots + 6 \times N(6)\right)$$

is also a representation for the average amount paid by the casino on these n plays of the game. But in the long run we would expect that each outcome occurs roughly one-sixth of the time, i.e., that

$$\lim_{n\to\infty} \frac{N(i)}{n} = \frac{1}{6}.$$

From this,

$$\lim_{n\to\infty} \frac{1}{n}\left(1 \times N(1) + 2 \times N(2) + \cdots + 6 \times N(6)\right) = \sum_{i=1}^{6} i\frac{1}{6}$$

Thus in the long run we would expect that

$$\lim_{n\to\infty} \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right) = \sum_{i=1}^{6} i\frac{1}{6}$$

Notice that the right-hand side of this equation is just

$$\sum_x x f_X(x).$$

The answer to the casino's question is that the break-even charge for playing the game should be

$$\sum_{i=1}^{6} i\frac{1}{6} = \frac{7}{2}$$

or $3.50.

■

In this simple case, we deduced that a reasonable calculation for the 'expected' pay outs by the casino were

$$E(X) = \sum_x x f_X(x).$$

In fact, the reasoning in this simple case applies for any *discrete* random variable, leading to the following definition.

12.2. Definition.

Let X be a discrete random variable having state space $\mathcal{S}$. Then the expected value *of X is*

$$E(X) = \sum_{x \in \mathcal{S}} x f_X(x)$$

provided that the sum converges absolutely, i.e., provided that

$$\sum_{x \in \mathcal{S}} |x| f_X(x) < \infty$$

12.3. Example.

If X has a Poisson distribution with parameter λ then the expected value of X is λ.

Solution. This can be calculated directly:

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} \\ &= \sum_{x=1}^{\infty} \frac{x\lambda^x}{x!} e^{-\lambda} \\ &\quad \text{make the change of variables } y = x - 1 \\ &= \sum_{y=0}^{\infty} \frac{(y+1)\lambda^{y+1}}{(y+1)!} e^{-\lambda} \\ &= \lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{(y)!} e^{-\lambda} \\ &= \lambda \end{aligned}$$

so $E(X) = \lambda$.

∎

Since discrete random variables always have density functions in this section we will concentrate on properties of the expectation of discrete random variables. We begin with a simple change of variables formula.

12.4. Theorem.

Let X be a discrete random variable having density function $f_X(x)$ and state space $\mathcal{S}$. Let φ be a real-valued function. Then the random variable $Y = \varphi(X)$ has finite expectation if and only if

$$\sum_{x\in\mathcal{S}} |\varphi(x)| f_X(x) < \infty$$

and in this case

$$E(\varphi(X)) = \sum_{x\in\mathcal{S}} \varphi(x) f_X(x)$$

Proof. Let $\mathcal{S}_Y$ be the state space for Y and let $f_Y(y)$ be the density function for Y, so that Y has finite expectation if and only if

$$\sum_{y\in\mathcal{S}_Y} |y| f_Y(y) < \infty.$$

For each $y \in \mathcal{S}_Y$ there is at least one $x \in \mathcal{S}$ so that $\varphi(x) = y$; thus if we let

$$E_y = \{x \in \mathcal{S} : \varphi(x) = y\}$$

then each E_y is non-empty and, if $y_1 \neq y_2$, then E_{y_1} and E_{y_2} are disjoint. Further, the events $\{Y = y\}$ and $\{X \in E_y\}$ are the same events, so

$$\begin{aligned} f_Y(y) &= \mathfrak{Pr}\,(Y = y) \\ &= \sum_{x\in E_y} \mathfrak{Pr}\,(X = x) \\ &= \sum_{x\in E_y} f_X(x) \end{aligned}$$

From this

$$\begin{aligned}
\sum_{y\in\boldsymbol{\mathcal{S}}_Y} |\boldsymbol{y}|\boldsymbol{f}_Y(\boldsymbol{y}) &= \sum_{y\in\boldsymbol{\mathcal{S}}_Y} |\boldsymbol{y}| \sum_{x\in \boldsymbol{E}_y} \mathfrak{Pr}\,(\boldsymbol{X}=\boldsymbol{x}) \\
&= \sum_{y\in\boldsymbol{\mathcal{S}}_Y} \sum_{x\in \boldsymbol{E}_y} |\boldsymbol{y}|\boldsymbol{f}_X(\boldsymbol{x}) \\
&\quad \text{for } \boldsymbol{x}\in \boldsymbol{E}_y,\ \varphi(\boldsymbol{x})=\boldsymbol{y}, \text{ so...} \\
&= \sum_{y\in\boldsymbol{\mathcal{S}}_Y} \sum_{x\in \boldsymbol{E}_y} |\varphi(\boldsymbol{x})|\boldsymbol{f}_X(\boldsymbol{x})
\end{aligned}$$

Now the events $\{\boldsymbol{E}_y\}$ are disjoint and their union must be $\boldsymbol{\mathcal{S}}$, the state space for $\boldsymbol{X}$. Thus the last line in the above is equivalent to summing over $\boldsymbol{x}\in\boldsymbol{\mathcal{S}}$, i.e.,

$$\sum_{y\in\boldsymbol{\mathcal{S}}_Y} |\boldsymbol{y}|\boldsymbol{f}_Y(\boldsymbol{y}) = \sum_{x\in\boldsymbol{\mathcal{S}}} |\varphi(\boldsymbol{x})|\boldsymbol{f}_X(\boldsymbol{x})$$

showing the first conclusion. The second conclusion follows upon repeating the above arguments without the absolute values.

∎

Expectation is in fact a $linear$ operator acting on random variables, as the following theorem shows.

12.5. Theorem.

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be continuous random variables having density functions $\boldsymbol{f}_Y(\boldsymbol{x})$ and $\boldsymbol{f}_Y(\boldsymbol{y})$ respectively. Then
(i) if $\boldsymbol{\lambda}\in\mathbb{R}$ then $\boldsymbol{E}(\boldsymbol{\lambda X})=\boldsymbol{\lambda E}(\boldsymbol{X})$
(ii) if $\mathfrak{Pr}\,(\boldsymbol{X}=\boldsymbol{\lambda})=\boldsymbol{1}$ then $\boldsymbol{E}(\boldsymbol{X})=\boldsymbol{\lambda}$
(iii) If $\boldsymbol{X}$ and $\boldsymbol{Y}$ have finite expectation, then $\boldsymbol{X}+\boldsymbol{Y}$ has finite expectation and

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

(iv) if $\mathfrak{Pr}\,(\boldsymbol{X}\geq\boldsymbol{Y})=\boldsymbol{1})$ then $\boldsymbol{E}(\boldsymbol{X})\geq\boldsymbol{E}(\boldsymbol{Y})$
(v) $|\boldsymbol{E}(\boldsymbol{X})|\leq\boldsymbol{E}(|\boldsymbol{X}|)$

Proof. Part *(i)* follows from the preceding theorem with $\varphi(x) \equiv \lambda$:

$$\begin{aligned} E(\lambda X) &= \sum_{x \in \mathcal{E}_X} \lambda f_X(x) \\ &= \lambda \sum_{x \in \mathcal{E}_X} f_X(x) \\ &= \lambda E(X) \end{aligned}$$

For *(ii)*, if $\mathfrak{Pr}\,(X = \lambda) = 1$ then the density function for X must be

$$f_X(x) = \begin{cases} 1 & \text{if } x = \lambda \\ 0 & \text{otherwise} \end{cases}$$

From this, $E(X) = \sum_x x f_X(x) = \lambda$.

Part *(iii)* is slightly more complex. Suppose that X has state space $\mathcal{S}_X$ and Y has state space $\mathcal{S}_Y$. Set $Z = X + Y$ and let $\mathcal{S}_Z$ be the state space for Z; then

$$\begin{aligned} \mathfrak{Pr}\,(Z = z) &= \mathfrak{Pr}\,(X + Y = z) \\ &\sum_{y \in \mathcal{S}_Y} \mathfrak{Pr}\,(X + Y = z \quad \text{and} \quad Y = y) \\ &\sum_{y \in \mathcal{S}_Y} \mathfrak{Pr}\,(X = z - y \quad \text{and} \quad Y = y) \end{aligned}$$

Now Z as finite expectation if and only if $\sum_z |z| f_Z(z) < \infty$, i.e., if and only if

$$\begin{aligned} \sum_{z \in \mathcal{S}_Z} |z| f_Z(z) &= \sum_{z \in \mathcal{S}_Z} \sum_{y \in \mathcal{S}_Y} |z|\, \mathfrak{Pr}\,(X = z - y \quad \text{and} \quad Y = y) \\ &= \sum_{y \in \mathcal{S}_Y} \sum_{z \in \mathcal{S}_Z} |z|\, \mathfrak{Pr}\,(X = z - y \quad \text{and} \quad Y = y) \end{aligned}$$

Now make the change of variables $u = z - y$. Then u ranges over exactly $\mathcal{S}_X$ and so

$$\begin{aligned}
\sum_{z\in\mathcal{S}_Z} |z| f_Z(z) &= \sum_{y\in\mathcal{S}_Y}\sum_{u\in\mathcal{S}_X} |u+y|\,\mathfrak{Pr}\left(X = u \quad \text{and} \quad Y = y\right)\\
&\le \sum_{y\in\mathcal{S}_Y}\sum_{u\in\mathcal{S}_X} (|u|+|y|)\,\mathfrak{Pr}\left(X = u \quad \text{and} \quad Y = y\right)\\
&=\le \sum_{y\in\mathcal{S}_Y}\sum_{u\in\mathcal{S}_X} |u|\,\mathfrak{Pr}\left(X = u \quad \text{and} \quad Y = y\right)+\\
&\qquad + \sum_{y\in\mathcal{S}_Y}\sum_{u\in\mathcal{S}_X} |y|\,\mathfrak{Pr}\left(X = u \quad \text{and} \quad Y = y\right)\\
&= \sum_{u\in\mathcal{S}_X} \le \sum_{y\in\mathcal{S}_Y} |u|\,\mathfrak{Pr}\left(X = u \quad \text{and} \quad Y = y\right)+\\
&\qquad + \sum_{y\in\mathcal{S}_Y} |y|\,\mathfrak{Pr}\left(Y = y\right)\\
&= \sum_{u\in\mathcal{S}_X} |u|\,\mathfrak{Pr}\left(X = u\right) + \sum_{y\in\mathcal{S}_Y} |y|\,\mathfrak{Pr}\left(Y = y\right)\\
&< \infty
\end{aligned}$$

since both X and Y have finite expectation. This shows that $Z = X + Y$ has finite expectation. Repeating the above argument without the absolute values proves that $E(X+Y) = E(X) + E(Y)$.

For *(iv)* set

$$U = X - Y.$$

Then if $u < 0$ it follows that

$$\mathfrak{Pr}\left(U = u\right) = \mathfrak{Pr}\left(X - Y = u\right) = 0$$

and hence that

$$\begin{aligned}
E(U) &= \sum_{u\in\mathcal{S}_U} u f_U(u)\\
&= \sum_{u>0,\, u\in\mathcal{S}_U} u f_U(u)\\
&\ge 0
\end{aligned}$$

Then with $\lambda = -1$ in *(ii)* it follows that

$$\begin{aligned}
0 &\le E(U)\\
&= E(X - Y)\\
&= E(X) - E(Y)
\end{aligned}$$

from which $E(Y) \leq E(X)$.

Now *(v)* follows from *(iv)* and the observation that

$$-|X| \leq X \leq |X|$$

with probability one.

$\blacksquare$

The expected value of a random variable X is sometimes also referred to as the **mean** of the random variable. The letter μ is usually reserved for the expectation or mean, so we will sometimes write

$$E(X) \equiv \mu_X.$$

While the mean measures the average, or central tendency, of the distribution, it does not tell the whole story. For example, consider the two random variables X_1 and X_2 having respective densities f_1 and f_2 given by

$$\begin{array}{rcl}
f_1(2) = \dfrac{1}{9} & \qquad & f_2(11) = \dfrac{1}{9} \\
f_1(2) = \dfrac{1}{9} & & f_2(15) = \dfrac{1}{9} \\
f_1(2) = \dfrac{1}{9} & & f_2(16) = \dfrac{1}{9} \\
f_1(10) = \dfrac{1}{9} & & f_2(17) = \dfrac{1}{9} \\
f_1(17) = \dfrac{1}{9} & & f_2(18) = \dfrac{1}{9} \\
f_1(24) = \dfrac{1}{9} & & f_2(20) = \dfrac{1}{9} \\
f_1(26) = \dfrac{1}{9} & & f_2(20) = \dfrac{1}{9} \\
f_1(34) = \dfrac{1}{9} & & f_2(20) = \dfrac{1}{9} \\
f_1(45) = \dfrac{1}{9} & & f_2(25) = \dfrac{1}{9}
\end{array}$$

One can readily check that both X_1 and X_2 have the same expectation, namely 18. However it is clear from inspection that there is a qualitative difference between the two random variables: there is more variability in the range of X_1 than there is in the range of X_2. In order to formalize this notion of variability we introduce the concept of variance.

12.6. Definition.

Let X be a discrete random variable having finite expectation μ and suppose that

$$E\left((X-\mu)^2\right) < \infty.$$

Then the **variance** *of X is the number*

$$\sigma_X^2 \equiv E\left((X-\mu)^2\right).$$

The **standard deviation** *of X is the square root of the variance, i.e.*

$$\sigma_X = \sqrt{(E\left((X-\mu)^2\right)}.$$

The following gives a computational formula for the variance that is often useful.

12.7. Theorem.

Let X be a discrete random variable having finite mean μ and finite variance σ^2. Then

$$\sigma_X^2 = E(X^2) - (E(X))^2.$$

Proof. This follows on expanding the definition of the variance:

$$\begin{aligned}
\sigma_X^2 &= E\left((X-\mu)^2\right) \\
&= E\left(X^2 - 2\mu X + \mu^2\right) \\
&= E(X^2) - 2\mu E(X) + \mu^2 \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - (E(X))^2.
\end{aligned}$$

■

12. Expectations: Discrete Random Variables: Problems.

1. Let X be a random variable with density function defined by

$$f_X(x) = \begin{cases} \frac{1}{x(x+1)} & \text{if } x = 1, 2, \cdots \\ 0 & \text{elsewhere} \end{cases}$$

(a) Show that f_X satisfies

$$\sum_x f_X(x) = 1.$$

(Hint: Note that

$$\frac{1}{x(x+1)} = \frac{1}{x} - \frac{1}{x+1}$$

and compute the partial sums.)

(b) Show that X does not have finite expectation.

2. Find the mean and variance for a Poison random variable.

13. Probability Generating Functions

The Probability Generating Function has many important applications, only a few of which we will be able to discuss here. It is sometimes also called the z-Transform, a term introduced by E.I. Jury in 1958 in connection with sampled data control systems.

13.1. Definition.

Let X be a discrete random variable defined on the probability space $(\Omega, \mathcal{E}, \mathfrak{Pr})$ and let t be a real number. If

$$\sum_x \mathfrak{Pr}\,(X = x)t^x$$

converges absolutely then we define the probability generating function for X to be

$$\Phi_X(t) = \sum_x \mathfrak{Pr}\,(X = x)t^X.$$

Because of Theorem 13.4, the following proposition is immediate.

13.2. Proposition.

Let X be a discrete random variable having probability generating function $\Phi_X(t)$. Then

$$\Phi_X(t) = E(t^X).$$

Proof. Taking $\phi(u) = t^u$ in Theorem 13.4, this result follows immediately. ∎

13.3. Example.

Let X be a Poisson random variable having parameter λ. Then

$$\Phi_X(t) = e^{\lambda(t-1)}.$$

Proof.

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

∎

Because the series that defines the Probability Generating Function converges absolutely, it is well-behaved with respect to differentiation.

13.4. Proposition.

Let X be a discrete random variable having probability generating function $\Phi_X(t)$ defined on some open interval I. Then for any $t \in I$ $\Phi_X(t)$ is differentiable,

$$\Phi'_X(t) = E\left(Xt^{X-1}\right)$$

and

$$\Phi''_X(t) = E\left(X(X-1)t^{X-2}\right)$$

Proof. Since the series converges absolutely, we can interchange the infinite sum and the differentiation operator:

$$\begin{aligned}\frac{d}{dt}\Phi_X(t) &= \frac{d}{dt}E\left(t^X\right)\\ &= E\left(\frac{d}{dt}t^X\right)\\ &= E\left(Xt^{X-1}\right).\end{aligned}$$

The second conclusion is deduced in exactly the same way.

∎

13.5. Definition.

Let X be a discrete random variable. The n^{th} **moment of X** *is the number $E(X^n)$ provided that the expectation is finite.*

Because of the above proposition, probability generating functions give a link between the moments of a random variable and the derivatives of the moment generating function.

13.6. Proposition.

Let X be a discrete random variable having finite first and second moments. Then

$$E(X) = \Phi'_X(t)\big|_{t=0}$$

and

$$E\left(X(X-1)\right) = \Phi''_X(t)\big|_{t=0}.$$

Proof. This follows readily from Proposition 14.4

∎

The above proposition makes it possible to compute the mean and variance for a random variable X using the moment generating function Φ_X. The following example does this for a Poisson random variable; the problems at the end of this chapter extend this to other discrete random variables.

13.7. Example.

If X is a Poisson random variable with parameter λ then X has mean $\mu_X = \lambda$ and variance $\sigma_X^2 = \lambda$.

Proof. Since $\Phi_X(t) = e^{\lambda(t-1)}$ it follows that

$$\begin{aligned}\mu_X &= \frac{d}{dt}\Phi_X(t)\big|_{t=0}\\ &= \frac{d}{dt}e^{\lambda(t-1)}\big|_{t=0}\\ &= \lambda e^{\lambda(t-1)}\big|_{t=0}\\ &= \lambda.\end{aligned}$$

Similarly,

$$\begin{aligned}E(X^2) - E(X) &= E\left(X(X-1)\right)\\ &= \frac{d^2}{dt^2}\Phi_X(t)\big|_{t=0}\\ &= \frac{d^2}{dt^2}e^{\lambda(t-1)}\big|_{t=0}\\ &= \lambda^2 e^{\lambda(t-1)}\big|_{t=0}\\ &= \lambda^2.\end{aligned}$$

From this

$$\begin{aligned}E(X^2) &= \lambda^2 + E(X)\\ &= \lambda^2 + \lambda.\end{aligned}$$

Thus

$$\begin{aligned}\sigma_X^2 &= E(X^2) - (E(X))^2\\ &= \lambda^2 + \lambda - (\lambda)^2\\ &= \lambda.\end{aligned}$$

∎

While the probability generating functions are one way to calculate the mean and variance of a random variable, by far the more important application of probability generating functions is in the next theorem.

13.8. Theorem.

Let X and Y be independent discrete random variables having probability generating functions Φ_X and Φ_Y respectively. Set $Z = X + Y$. Then Z has probability generating function

$$\Phi_Z(t) = \Phi_X(t)\Phi_Y(t)$$

i.e., the probability generating function of the sum of two independent random variables is the product of the probability generating functions:

$$\Phi_{X+Y}(t) = \Phi_X(t)\Phi_Y(t)$$

Proof. If Z has density function $f_Z(z)$ then

$$\begin{aligned}
\Phi_{X+Y}(t) &= \sum_z f_Z(z)t^z \\
&= \sum_z t^z\, \mathfrak{Pr}\left(X + Y = z\right) \\
&= \sum_z t^z \sum_x \mathfrak{Pr}\left(X = x \quad \text{and} \quad Y = z - x\right) \\
&= \sum_z t^z \sum_x \mathfrak{Pr}\left(X = x\right) \mathfrak{Pr}\left(Y = z - x\right) \\
&= \sum_z t^z \sum_x^z f_X(x) f_Y(z - x) \\
&= \sum_x f_X(x)t^x \sum_z t^{z-x} f_Y(z - x) \\
&\quad \text{(change of variables } y = z - x\text{)} \\
&= \sum_x f_X(x)t^x \sum_y t^y f_Y(y) \\
&= \Phi_X(t)\Phi_Y(t).
\end{aligned}$$

∎

13.9. Example.

Let X and Y be independent Poisson random variables with parameters λ_X and λ_Y respectively. Then $Z = X + Y$ is a Poisson random variable with parameter $\lambda_Z = \lambda_X + \lambda_Y$.

Proof. This follows from the preceding theorem:

$$\begin{aligned}\Phi_{X+Y}(t) &= \Phi_X(t)\Phi_Y(t)\\ &= e^{\lambda_X(t-1)}e^{\lambda_Y(t-1)}\\ &= e^{(\lambda_X+\lambda_Y)(t-1)}\end{aligned}$$

which is the probability generating function for a Poisson random variable having parameter $\lambda_X + \lambda_Y$.

∎

Readers who are familiar with the Laplace transform from differential equations will notice the obvious similarities to the probability generating function. For example, Theorem 14.8 is analogous to the fact that the Laplace transform of the convolution is the product of the Laplace transforms.

The z-transformation mentioned at the start of this section is actually defined to be $E(z^X)$ where z is a complex number. The z-transform has many properties beyond those discussed in this section. It is a transformation from the *time domain* to the *frequency domain* of the random variable. Roughly speaking, the time domain graph shows how the signal changes over time whereas the frequency domain graph shows how much of the signal lies within each frequency band over a range of frequencies.

The resulting function is also called the *frequency spectrum* of the signal. The frequency spectrum has two components: magnitude and phase. In many applications only the magnitude is important. When the phase information is discarded the result is the *power spectrum* of the variable. A device that displays the power spectrum is a spectrum analyzer.

The inner ear is an example of a biological spectrum analyzer. The sounds arriving in the ear are transformed by the basilar membrane of the inner ear, which acts in effect as a spectrum analyzer of the incoming sound waves. This results in the brain perceiving the incoming sound as a collection of distinct notes rather than as disorganized noise.

13. Probability Generating Functions: Problems.

1. Let X be a Bernoulli random variable with parameter p, i.e., suppose that X has density function

$$f_X(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Show that $\Phi_X(t) = pt + 1 - p$.

2. Let X be a binomial random variable with parameters n and p. Show that

$$\Phi_X(t) = (pt + 1 - p)^n.$$

3. Let X be a negative binomial random variable with parameters α and p. Show that

$$\Phi_X(t) = \left(\frac{p}{1 - t(1 - p)} \right)^{\alpha}.$$

4. Let X_1 be a binomial random variable with parameters n_1 and p and let X_2 be a binomial random variable with parameters n_2 and p. If X_1 and X_2 are independent show that $X_1 + X_2$ is a binomial random variable with parameters p and $n_1 + n_2$.

5. Let X_1 be a negative binomial random variable with parameters α_1 and p and let X_2 be a negative binomial random variable with parameters α_2 and p. If X_1 and X_2 are independent show that $X_1 + X_2$ is a negative binomial random variable with parameters p and $\alpha_1 + \alpha_2$.

14. Jointly Distributed Continuous Random Variables

Jointly distributed continous random variables exhibit many of the same properties as jointly distributed discrete random variables, with integrals replacing sums.

14.1. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *be a probability space and let* X *and* Y *be random variables defined on* $(\Omega, \mathcal{E}, \mathfrak{Pr})$. *Then the joint distribution function of* X *and* Y *is*

$$F_{XY}(x, y) = \mathfrak{Pr}\,(X \leq x \quad \text{and} \quad Y \leq y).$$

While discrete random variables must always have a density function, and hence jointly distributed discrete random variables must always have a joint density, the same is not true for continuous random variables. However, in the case that the joint distribtution is differentiable, then we can find the joint density function.

14.2. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *be a probability space and let* X *and* Y *be continuous random variables defined on* $(\Omega, \mathcal{E}, \mathfrak{Pr})$. *If*

$$f_{XY}(x, y) \equiv \left. \frac{\partial^2 F_{XY}}{\partial x \partial y} \right|_{(x,y)}$$

exists then we say that the joint density function for X *and* Y *is* $f_{XY}(x, y)$.

The following analog to Theorem 11.3 is immediate.

14.3. Theorem.

Let $(\boldsymbol{\Omega}, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be random variables defined on $(\boldsymbol{\Omega}, \mathcal{E}, \mathfrak{Pr})$ having distribution functions $\boldsymbol{F_X(x)}$ and $\boldsymbol{F_Y(y)}$ respectively. If $\boldsymbol{F_{XY}(x,y)}$ is the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ then

(i) For each fixed $\boldsymbol{x} \in \mathbb{R}$

$$\lim_{y \to -\infty} F_{XY}(x,y) = 0 \quad \textit{and} \quad \lim_{y \to \infty} F_{XY}(x,y) = F_X(x)$$

(ii) For each fixed $\boldsymbol{y} \in \mathbb{R}$

$$\lim_{x \to -\infty} F_{XY}(x,y) = 0 \quad \textit{and} \quad \lim_{x \to \infty} F_{XY}(x,y) = F_Y(y)$$

(iii) If

$$f_{XY}(x,y) \equiv \left. \frac{\partial^2 F_{XY}}{\partial x \partial y} \right|_{(x,y)}$$

exists then $\boldsymbol{X}$ and $\boldsymbol{Y}$ have density functions given by

$$f_X(x) = \int_{\mathbb{R}} f_{XY}(x,y)\, dy \quad \textit{and} \quad f_Y(y) = \int_{\mathbb{R}} f_{XY}(x,y)\, dx$$

(iv) If $\boldsymbol{E} \subseteq \mathbb{R} \times \mathbb{R}$ then

$$\mathfrak{Pr}\,((X,Y) \in E) = \int\!\!\int_E f_{XY}(x,y)\, dx\, dy$$

provided that the integral exists.

14.4. Example.

Let X and Y have the bivariate density given by

$$f_{XY}(x,y) = \frac{\sqrt{3}}{4\pi}\exp\left(-\frac{x^2-xy+y^2}{2}\right) \qquad -\infty < x,y < \infty.$$

Then X is normally distributed with $\mu = 0$ and $\sigma^2 = 4/3$.

Solution. Using (iii) above

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy \\
&= \frac{\sqrt{3}}{4\pi}\int_{-\infty}^{\infty}\exp\left(-\frac{x^2-xy+y^2}{2}\right)dy \\
&= \frac{\sqrt{3}}{4\pi}\int_{-\infty}^{\infty}\exp\left[-\left(\frac{\left(y-\frac{x}{2}\right)^2+\frac{3x^2}{4}}{2}\right)\right]dy \\
&= \frac{\sqrt{3}}{4\pi}\exp\left(-\frac{3x^2}{8}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{\left(y-\frac{x}{2}\right)^2}{2}\right)dy \\
&= \frac{\sqrt{3}}{4\pi}\exp\left(-\frac{3x^2}{8}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{u^2}{2}\right)du \\
&= \frac{\sqrt{3}}{4\pi}\exp\left(-\frac{3x^2}{8}\right)\sqrt{2\pi} \\
&= \frac{\sqrt{3}}{2\sqrt{2\pi}}\exp\left(-\frac{3x^2}{8}\right)
\end{aligned}$$

which is the density function for a normally distributed random variable having parameters $\mu = 0$ and $\sigma^2 = 4/3$ as desired.

∎

14.5. Theorem.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let X and Y be random variables defined on $(\Omega, \mathcal{E}, \mathfrak{Pr})$ having joint density function $f_{XY}(x, y)$. If $Z = X + Y$ then the density function of Z is

$$\mathfrak{Pr}\,(Z \leq z) = \int_{\mathbb{R}} f_{XY}(x, z - x)\; dx.$$

Proof. Fix z and let A_z be the set

$$A_z = \{(x, y) \in \mathbb{R} \times \mathbb{R} : x + y \leq z\}.$$

Then

$$\begin{aligned}
\mathfrak{Pr}\,(Z \leq z) &= \mathfrak{Pr}\,(X + Y \leq z) \\
&= \int\!\!\int_{A_z} f_{XY}(x, y)\; dy\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{XY}(x, y)\; dy\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{z} f_{XY}(x, u - x)\; du\, dx \\
&= \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_{XY}(x, u - x) dx\, du
\end{aligned}$$

From this the density function for Z is

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(x, z - x) dx$$

∎

The above result is most frequently used when X and Y are independent.

independent!random variables

14.6. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *be a probability space and let* X *and* Y *be random variables defined on* $(\Omega, \mathcal{E}, \mathfrak{Pr})$. *Then* X *and* Y *are independent if for each* $x, y \in \mathbb{R} \times \mathbb{R}$

$$F_{XY}(x, y) = F_X(x)F_Y(y)$$

or equivalently if and only if

$$\mathfrak{Pr}(X \leq x \quad \textit{and} \quad Y \leq y) = \mathfrak{Pr}(X \leq x)\,\mathfrak{Pr}(Y \leq y).$$

Note that X and Y are independent if and only if whenever

$$a < b \quad \text{and} \quad c < d$$

then it follows that

$$\mathfrak{Pr}(a < X < b \quad \text{and} \quad c < Y < d) = F_X(b) - F_X(a) + F_Y(d) - F_Y(c).$$

More generally, if A and B are subsets of $\mathbb{R}$ that can be decomposed into the union of a finite or countably infinite set of intervals, then

$$\mathfrak{Pr}(X \in A \quad \text{and} \quad Y \in B) = \mathfrak{Pr}(X \in A)\,\mathfrak{Pr}(Y \in B)$$

or that the events

$$\{\omega \in \Omega : X(\omega) \in A\}$$

and

$$\{\omega \in \Omega : Y(\omega) \in B\}$$

are independent.

14.7. Definition.

Let X and Y be random variables and suppose that there is a density function f and associated distribution function $F(t) = \int_{-\infty}^{t} f(s)\,ds$ so that both

$$f_X(x) = f(x) \qquad -\infty < x < \infty$$

and

$$f_Y(y) = f(y) \qquad -\infty < y < \infty.$$

Then we say that X and Y are **identically distributed** *with common density f and common distribution F.*

Often we will simply say "X and Y are identically distributed" with the existence of the common density f and the common distribution function F implied. Of course if X and Y are identically distributed it follows immediately that

$$F_X(x) = \int_{-\infty}^{x} f(s)\,ds$$

and

$$F_Y(y) = \int_{-\infty}^{y} f(s)\,ds$$

The particular case where X and Y are also independent arises most frequently, especially in sampling theory.

The following proposition is immediate from the definitions.

14.8. Theorem.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let X and Y be random variables defined on $(\Omega, \mathcal{E}, \mathfrak{Pr})$ having joint density function $f_{XY}(x,y)$. Then X and Y are independent if and only if

$$f_{XY}(x,y) = f_X(x) f_Y(y)$$

Proof. This is immediate from the definitions and the formula

$$F_X(x)F_Y(y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_X(x) f_Y(y)\,dy\,dx$$

∎

14.9. Theorem.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *be a probability space and let* X *and* Y *be continuous, independent random variables defined on* $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *having density functions* $f_X(x)$ *and* $f_Y(y)$ *respectively. If* $Z = X + Y$ *then the density function for* Z *is*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx$$

14.10. Corollary.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *be a probability space and let* X *and* Y *be continuous, independent, non-negative random variables defined on* $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *having density functions* $f_X(x)$ *and* $f_Y(y)$ *respectively. If* $Z = X + Y$ *then the density function for* Z *is*

$$f_Z(z) = \int_{0}^{z} f_X(x) f_Y(z-x)dx$$

Proof. Since $X \geq 0$, $f_X(x) = 0$ if $x \leq 0$. Hence the integral vanishes on $[-\infty, 0]$. Since $Y \geq 0$, $f_Y(z-x) = 0$ if $z - x \leq 0$. Hence the integral vanishes on $[z, \infty]$. ∎

If you have had a course in differential equations will recognize the integral

$$\int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx$$

as the *convolution* of the density functions. In differential equations convolutions arise in connection with applying the Laplace transform to second order (or higher) linear differential equations.

14.11. Example.

Suppose that X_1 and X_2 are independent random variables having the gamma distribution with parameters $\lambda = \lambda_1 = \lambda_2$ and α_1 and α_2. Then $Z = X_1 + X_2$ has a gamma distribution with parameters λ and $\alpha_1 + \alpha_2$.

Proof. From the definitions, it follows that

$$f_{X_1}(x) = \frac{\lambda^{\alpha_1} x^{\alpha_1 - 1} e^{-\lambda x}}{\Gamma(\alpha_1)} \qquad x > 0$$

and

$$f_{X_2}(y) = \frac{\lambda^{\alpha_2} y^{\alpha_2 - 1} e^{-\lambda y}}{\Gamma(\alpha_2)} \qquad y > 0.$$

Thus $f_Z(z) = 0$ for $z \leq 0$ and for $z > 0$

$$\begin{aligned} f_Z(z) &= \int_0^z f_{X_1}(x) f_{X_2}(z - x)\, dx \\ &= \frac{\lambda^{\alpha_1 + \alpha_2} e^{-\lambda x}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z x^{\alpha_1 - 1} (z - x)^{\alpha_2 - 1}\, dx. \end{aligned}$$

Now make the change of variables $x = zu$ in the integral so, with $dx = zdu$, we obtain

$$f_Z(z) = \frac{\lambda^{\alpha_1 + \alpha_2} e^{-\lambda}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1 + \alpha_2 - 1} \int_0^1 u^{\alpha_1 - 1} (1 - u)^{\alpha_2 - 1}\, du. \qquad (14.1)$$

A calculation completes the argument:

$$\begin{aligned}
\Gamma(\alpha_1)\Gamma(\alpha_2) &= \left(\int_0^\infty x^{\alpha_1-1}e^{-x}\,dx\right)\left(\int_0^\infty y^{\alpha_2-1}e^{-y}\,dy\right)\\
&= \int_0^\infty \int_0^\infty x^{\alpha_1-1}y^{\alpha_2-1}e^{-(x+y)}\,dx\,dy\\
&\qquad \text{make the change of variables } u = x+y \text{ so } du = dx)\\
&= \int_0^\infty \int_y^\infty (u-y)^{\alpha_1-1}y^{\alpha_2-1}e^{-u}\,du\,dy\\
&= \int_0^\infty e^{-u}\int_0^u (u-y)^{\alpha_1-1}y^{\alpha_2-1}\,dy\,du\\
&\qquad \text{make the change of variables } uv = y \text{ so } u\,dv = dy)\\
&= \int_0^\infty e^{-u}\int_0^1 (1-v)^{\alpha_1-1}u^{\alpha_1-1}u^{\alpha_2-1}v^{\alpha_2-1}u\,dv\,du\\
&= \int_0^\infty u^{\alpha_1+\alpha_2-1}e^{-u}\int_0^1 (1-v)^{\alpha_1-1}v^{\alpha_2-1}\,dv\,du\\
&= \Gamma(\alpha_1+\alpha_2)\int_0^1 (1-v)^{\alpha_1-1}v^{\alpha_2-1}\,dv.
\end{aligned}$$

Substituting into equation (14.1) gives the result.

■

The following corollary is immediate.

14.12. Corollary.

If $\{X_1, \cdots, X_n\}$ are independently distributed random variables and if X_i has the gamma distribution with parameters λ and α_i, then

$$Z = X_1 + \cdots + X_n$$

has a gamma distribution with parameters λ and $\alpha_1 + \cdots + \alpha_n$.

14.13. Corollary.

Suppose that $X_1, X_2, \cdots, X_n$ are independent and identically distributed exponential random variables having common parameter λ. Then $Z = X_1 + X_2 + \cdots + X_n$ has a gamma distribution with parameters $\alpha = n$ and λ.

Proof. This follows from the fact that an exponential random variable is a special case of the gamma distribution with parameter $\alpha = 1$.

∎

As we have already noted, if X is normally distributed with parameters $\mu = 0$ and σ then $\frac{X^2}{\sigma^2}$ has the gamma distribution with parameters $\alpha = \frac{1}{2}$ and $\lambda = \frac{1}{2}$. Then the following corollary is also a consequence of the above.

14.14. Corollary.

Let $\{X_1, \cdots, X_n\}$ be independent, identically distributed normal random variables having $\mu = 0$ and $\sigma^2 = 1$. Then the random variable $Z = X_1^2 + \cdots X_n^2$ is a gamma random variable with parameters $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$.

The random variable Z in the preceding corollary is important in estimation theory and is said to have the **chi-squared distribution with n degrees of freedom.**

14.15. Theorem.

Let X and Y be random variables having a joint density function $f_{XY}(x,y)$ and set

$$Z = \frac{Y}{X}.$$

Then the density function of Z is

$$f_Z(z) = \int_{-\infty}^{\infty} |x| f_{XY}(x, xz)\, dx \qquad -\infty < z < \infty.$$

In particular, if $X \geq 0$ and $Y \geq 0$ then $f_Z(z) = 0$ if $z \leq 0$ and

$$f_Z(z) = \int_0^{\infty} x f_{XY}(x, xz)\, dx \qquad 0 < z < \infty.$$

Proof. Begin by setting

$$A_z = \{(x,y) : \frac{y}{x} \leq z\}.$$

We can decompose A_z into two sets depending on the algebraic sign of x:

$$A_z = \{(x,y) : x \leq 0 \quad \text{and} \quad y \geq xz\} \bigcup \{(x,y) : x \geq 0 \quad \text{and} \quad y \geq xz\}.$$

Thus

$$\begin{aligned} F_Z(z) &= \mathfrak{Pr}\,(Z \leq z) \\ &= \mathfrak{Pr}\left(\frac{Y}{X} \leq z\right) \\ &= \int\int_{A_z} f_{XY}(x,y)\, dy\, dx \\ &= \int_{-\infty}^{0} \int_{xz}^{\infty} f_{XY}(x,y)\, dy\, dx + \int_0^{\infty} \int_{-\infty}^{xz} f_{XY}(x,y)\, dy\, dx \end{aligned}$$

Now make the change of variables $y = xv$ in the inner integrals. Since, with this change of variables, $dy = x\,dv$,

$$\begin{aligned}
F_Z(z) &= \int_{-\infty}^{0}\int_{xz}^{\infty} f_{XY}(x,y)\,dy\,dx + \int_{0}^{\infty}\int_{-\infty}^{xz} f_{XY}(x,y)\,dy\,dx \\
&= \int_{-\infty}^{0}\int_{z}^{\infty} xf_{XY}(x,xv)\,dv\,dx + \int_{0}^{\infty}\int_{-\infty}^{z} xf_{XY}(x,xv)\,dv\,dx \\
&\qquad \text{now change the direction of integration in the inner intergral of the first term} \\
&= \int_{-\infty}^{0}\int_{-\infty}^{z} (-x)f_{XY}(x,xv)\,dv\,dx + \int_{0}^{\infty}\int_{-\infty}^{z} xf_{XY}(x,xv)\,dv\,dx \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z} |x|f_{XY}(x,xv)\,dv\right)dx
\end{aligned}$$

Now change the order of integration to obtain

$$F_Z(z) = \int_{-\infty}^{z}\int_{-\infty}^{\infty} |x|f_{XY}(x,xv)\,dx\,dv.$$

Differentiating with respect to z then gives

$$f_Z(z) = \int_{-\infty}^{\infty} |x|f_{XY}(x,xz)\,dx \qquad -\infty < z < \infty.$$

If X and Y are non-negative, then f_{XY} vanishes on $(-\infty, 0]$ and so $f_Z(z) = 0$ if $z < 0$ and

$$f_Z(z) = \int_{0}^{\infty} xf_{XY}(x,xz)\,dx \qquad 0 < z < \infty.$$

∎

14.16. Theorem.

Let X and Y be independent random variables and suppose that X has a gamma distribution with parameters λ and α_1 and that Y has a gamma distribution with parameters λ and α_2. Then the random variable

$$Z = \frac{Y}{X}$$

has density function $f_Z(z)$ given by

$$f_Z(z) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{z^{\alpha_2 - 1}}{(z+1)^{\alpha_1+\alpha_2}} \qquad 0 < z < \infty$$

and $f_Z(z) = 0$ if $z \le 0$.

Proof. Since

$$f_X(x) = \frac{\lambda^{\alpha_1} x^{\alpha_1 - 1} e^{-\lambda x}}{\Gamma(\alpha_1)} \qquad x > 0$$

and

$$f_X(x) = \frac{\lambda^{\alpha_2} y^{\alpha_2 - 1} e^{-\lambda y}}{\Gamma(\alpha_2)} \qquad y > 0$$

it follows that

$$\begin{aligned} f_Z(z) &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty x x^{\alpha_1 - 1} e^{-\lambda x} (xz)^{\alpha_2 - 1} e^{-\lambda x z}\, dx \\ &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^\infty x^{\alpha_1 + \alpha_2 - 1} e^{-\lambda x (z+1)}\, dx. \end{aligned} \qquad (14.2.)$$

Since

$$\begin{aligned} \int_0^\infty x^{\alpha-1} e^{-\lambda x}\, dx &= \int_0^\infty \left(\frac{u}{\lambda}\right)^{\alpha-1} e^{-\lambda u} \frac{1}{\lambda}\, du \\ &= \frac{1}{\lambda^\alpha} \int_0^\infty u^{\alpha-1} e^{-u}\, du \\ &= \frac{\Gamma(\alpha)}{\lambda^\alpha} \end{aligned}$$

it follows that

$$\int_0^\infty x^{\alpha_1+\alpha_2-1} e^{-\lambda x(z+1)}\, dx = \frac{\Gamma(\alpha_1+\alpha_2)}{(\lambda(z+1))^{\alpha_1+\alpha_2}}.$$

Substituting into (14.2) above yields the result.

∎

14.17. Example.

Suppose that X and Y are independent, normally distributed random variables having $\mu = 0$ and $\sigma^2 = 1$. If

$$T = \frac{Y^2}{X^2}$$

then T has distribution given by

$$\begin{aligned} f_T(t) &= \frac{\Gamma(1)}{\Gamma(1/2)\Gamma(1/2)} \frac{t^{-(1/2)}}{t+1} \\ &= \frac{1}{\pi(t+1)\sqrt{t}}. \end{aligned}$$

Proof. This follows immediately from the above theorem, that $\Gamma(1/2) = \sqrt{\pi}$ and that both X^2 and Y^2 have the gamma distribution with parameters $\alpha = 1/2$ and $\lambda = 1/2$.

∎

14. Jointly Distributed Continuous Random Variables: Problems.

1. Let X and Y be independent random variables each uniformly distributed on $(0,1)$. Find
(a) $\mathfrak{Pr}\ (|X-Y|) \leq 0.5$
(b) $\mathfrak{Pr}\ (Y \geq X)$.

2. Let X and Y be normally distributed random variables with $\mu = 0$ and the same parameter σ^2. Find $\mathfrak{Pr}\,(X^2 + y^2 \leq 1)$.

3. Suppose that the times it takes two workers to complete a task are independently and exponentially distributed random variables with a parameter of λ. What are the chances that it takes the first worker at least twice as long as the second worker to complete the task?

4. Let X and Y be continuous random variables having joint density given by

$$f_{XY}(x,y) = \begin{cases} \lambda^2 e^{-\lambda y} & \text{if } 0 \leq x \leq y \\ 0 & \text{elsewhere.} \end{cases}$$

Find $f_X(x)$ and $f_Y(y)$.

5. Let R and Θ be independent random variables and suppose that R has the Rayleigh density:

$$f_R(r) = \begin{cases} \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) & r \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and that Θ is uniformly distributed over $(-\pi, \pi)$. Show that $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$ are independent random variables and that each has a normal distribution with parameters $\mu = 0$ and σ^2.

6. Let X and Y be independent, normally distributed random variables having parameters $\mu = 0$ and $\sigma^2 = 1$. Show that the random variable

$$Z = \frac{Y}{X}$$

has the Cauchy distribution, i.e., that

$$f_Z(z) = \frac{1}{\pi(1+z^2)} \qquad -\infty < z < \infty.$$

conditional density!discrete random variables

15. Conditional Densities

Suppose that X and Y are discrete random variables. We can compute the conditional probability

$$\mathfrak{Pr}\left(Y = Y \middle| X = x\right)$$

using the conventional definition of conditional probability provided, of course, that $\mathfrak{Pr}\left(X = x\right) \neq 0$:

$$\mathfrak{Pr}\left(Y = Y \middle| X = x\right) = \frac{\mathfrak{Pr}\left(Y = y \quad \text{and} \quad X = y\right)}{\mathfrak{Pr}\left(X = x\right)} = \frac{f_{XY}(x, y)}{f_X(x)} \tag{15.1}$$

Thus we can define a conditional density function $f_{Y|X}(y|x)$ for each value of x for which $f_X(x) \neq 0$:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}. \tag{15.2}$$

Conditional densities and the related calculations for discrete random variables reduce readily to cases considered in section six since the probabilities can all be reduced to sums involving $f_{XY}(x, y)$ and $f_X(x)$. This reduction is not possible for continuous random variables since $\mathfrak{Pr}\left(X = x\right) = 0$ for all x and hence the calculations in (15.1) are undefined.

One possible approach would be to consider conditional probabilities for continuous random variables by using a limiting process. For example, we might approximate $\mathfrak{Pr}\left(a \leq Y \leq b \middle| X = x\right)$ with

$$\begin{aligned}
\mathfrak{Pr}\left(a \leq Y \leq b \middle| X = x\right) &= \lim_{h \to 0} \mathfrak{Pr}\left(a \leq Y \leq b \middle| x - h \leq X \leq x + h\right) \\
&= \lim_{h \to 0} \frac{\mathfrak{Pr}\left(a \leq Y \leq b \text{ and } x - h \leq X \leq x + h\right)}{\mathfrak{Pr}\left(x - h \leq X \leq x + h\right)}.
\end{aligned}$$

Now if $f_X(x) > 0$ then for $h > 0$ sufficiently small $\mathfrak{Pr}\left(x - h < X, x + h\right) > 0$, and so using this approximation and supposing fairly mild continuity conditions one obtains

$$\begin{aligned}
\mathfrak{P}\mathfrak{r}\left(a \leq Y \leq b \middle| X = x\right) &= \lim_{h \to 0} \frac{\mathfrak{P}\mathfrak{r}\left(a \leq Y \leq b \text{ and } x - h \leq X \leq x + h\right)}{\mathfrak{P}\mathfrak{r}\left(x - h \leq X \leq x + h\right)} \\
&= \lim_{h \to 0} \frac{\int_{x-h}^{x+h} \int_a^b f_{XY}(u, y)\, dy\, du}{\int_{x-h}^{x+h} f_X(u)\, du} \\
&= \lim_{h \to 0} \frac{\frac{1}{2h} \int_{x-h}^{x+h} \int_a^b f_{XY}(u, y)\, dy\, du}{\frac{1}{2h} \int_{x-h}^{x+h} f_X(u)\, du} \\
&= \frac{\int_a^b f_{XY}(xy)\, dy}{f_X(x)} \\
&\quad \text{(applying the fundamental theorem of calculus)} \\
&= \int_a^b \frac{f_{XY}(xy)}{f_X(x)}\, dy
\end{aligned}$$

i.e.,

$$\mathfrak{P}\mathfrak{r}\left(a \leq Y \leq b \middle| X = x\right) = \int_a^b \frac{f_{XY}(xy)}{f_X(x)}\, dy$$

which would then yield that the conditional random variable $Y|_{X=x}$ would have density function

$$f_{Y|X}(y|x) = \frac{f_{XY}(xy)}{f_X(x)}$$

which agrees exactly with discrete case (15.2). Here we are using the version of the fundamental theorem of calculus which asserts

$$\begin{aligned}
\xi(x) &= \frac{d}{dx} \int_a^x \xi(u)\, du \\
&= \lim_{h \to 0} \frac{1}{2h} \int_{x-h}^{x+h} \xi(u)\, du
\end{aligned}$$

where the integrand ξ is continuous in an interval about $u = x$. Thus would need to know that the integrands

$$\int_a^b f_{XY}(u, y)\, dy \quad \text{and} \quad f_X(u)$$

were continuous in an interval around $u = x$. Of course we also need to know that $f_X(x) \neq 0$. These fairly modest continuity requirements then yield exactly the same definition for the conditional density $f_{Y|X}(y|x)$ as in the discrete case. Thus we will *define* this to be the conditional density in the continuous case as well and dispense with taking limits as above.

15.1. Definition.

Let X and Y be random variables having joint density function $f_X(x, y)$. Then we define the conditional density of Y given X to be

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

whenever $f_X(x) \neq 0$.

Note that the definition does not distinguish between discrete and continuous random variables because of our arguments above.

15.2. Example.

Suppose that X and Y have the joint density function given by

$$f_{XY}(x, y) = \frac{\sqrt{3}}{4\pi} \exp\left(-\frac{(x^2 - xy + y^2)}{2}\right).$$

Then

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(y - \frac{x}{2}\right)^2}{2}\right).$$

Proof. We have previously shown that X has the normal density with parameters $\mu = 0$ and $\sigma^2 = 4/3$. Thus for $-\infty < y < \infty$

$$\begin{aligned} f_{Y|X} &= \frac{\frac{\sqrt{3}}{4\pi} \exp\left(-\frac{(x^2 - xy + y^2)}{2}\right)}{\frac{\sqrt{3}}{2\sqrt{2\pi}} \exp\left(-\frac{3x^2}{8}\right)} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - (x/2))^2}{2}\right). \end{aligned}$$

Thus the conditional distribution of Y given $X = x$ is normal with parameters $\mu = x/2$ and $\sigma^1 = 1$.

∎

In the above example we started with the joint density and deduced the marginal and conditional densities. In many applications we need to reverse this process, i.e., we will know the marginal and conditional densities and need to deduce the joint distribution. The following is an example of this.

15.3. Example.

Suppose that X is uniformly distributed on $(0,1)$ and Y is uniformly distributed on $(0,X)$. Find the joint density f_{XY} of X and Y and find the marginal density $f_Y(y)$ of Y.

Solution. Clearly the marginal density of X is given by

$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Further the conditional density $f_{Y|X}(y|x)$ is

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{x} & \text{if } 0 < y < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Thus the joint density of X and Y is

$$f_{XY}(x,y) = \begin{cases} \frac{1}{x} & \text{if } 0 < y < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

The marginal density of Y is then given by

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx \\ &= \int_y^1 \frac{1}{x}\,dx \\ &= -\ln(y) \end{aligned}$$

if $0 < y < 1$ and $f_Y(y) = 0$ otherwise.

∎

In the following example, one of the random variables is discrete while the other is continuous. While the example below deals with the distribution of accidents in a human population, similar examples arise network design.

15.4. Example.

Suppose that the number of automobile accidents a driver will have in a given year is a random variable Y having the Poisson distribution with parameter λ where the value of λ depends upon the driver. The the random variable Λ that assigns λ to each member of the population will have a distribution $f_\Lambda(\lambda)$. Under certain circumstances it is reasonable to suppose that $f_\Lambda(\lambda)$ has a gamma distribution with parameters α and β, i.e., that

$$f_\Lambda(\lambda) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\lambda\beta}}{\Gamma(\alpha)}.$$

Find $f_{\Lambda Y}(\lambda, y)$, $f_Y(y)$ and $f_{\Lambda|Y}(\lambda|y)$.

Solution. From the narrative,

$$f_{Y|\Lambda}(y|\lambda) = \begin{cases} \frac{\lambda^y e^{-y}}{y!} & \text{for } y = 0, 1, 2, \cdots \\ 0 & \text{elsewhere} \end{cases}$$

Thus the joint density of X and Y is

$$\begin{aligned} f_{\Lambda Y}(\lambda, y) &= f_\Lambda(\lambda) f_{Y|\Lambda}(y|\lambda) \\ &= \begin{cases} \frac{f_\Lambda(\lambda)\lambda^y e^{-\lambda}}{y!} & \text{if } y = 0, 1, 2, \cdots \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

From this it follows that

$$f_{\Lambda Y}(\lambda, y) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\lambda\beta}}{\Gamma(\alpha)} \frac{\lambda^y e^{-\lambda}}{y!}$$

for $\lambda >$ and for $y = 0, 1, 2, \cdots$ and is zero elsewhere.

Hence

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{\Lambda Y}(\lambda, y)\, d\lambda \\ &\int_0^{\infty} \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\lambda\beta}}{\Gamma(\alpha)} \frac{\lambda^y e^{-\lambda}}{y!}\, d\lambda \\ &= \frac{\beta^\alpha}{y!\Gamma(\alpha)} \int_0^{\infty} \lambda^{\alpha+y-1} e^{-\lambda(\beta+1)}\, d\lambda \\ &= \frac{\Gamma(\alpha+y)\beta^\alpha}{y!\Gamma(\alpha)(\beta+1)^{\alpha+y}}. \end{aligned}$$

We have used

$$\int_0^{\infty} x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\Gamma(\alpha)}{\lambda^\alpha}.$$

We leave it to the exercises to verify that Y has a negative binomial distribution with parameters α and $p = \beta/(1+\beta)$.

Finally for $\lambda > 0$

$$\begin{aligned} f_{\Lambda|Y}(\lambda|y) &= f_{\Lambda Y}(\lambda, y) \\ &= \frac{\beta^\alpha \lambda^{\alpha+y-1} e^{-\lambda(\beta+1)} y!\Gamma(\alpha)(\beta+1)^{\alpha+y}}{\Gamma(\alpha) y! \Gamma(\alpha+y)\beta^\alpha} \\ &= \frac{(\beta+1)^{\alpha+y} \lambda^{\alpha+y-1} e^{-\lambda(\beta+1)}}{\Gamma(\alpha+y)} \end{aligned}$$

which implies that the conditional distribution of Λ given $Y = y$ is the gamma distribution with parameters $\alpha + y$ and $\beta + 1$.

∎

15.5. Theorem. Bayes' Rule.

Let X and Y be continuous random variables having a joint density function $f_{XY}(x, y)$. Then

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y|x)\, dx}. \tag{15.3}$$

Proof. Reversing the roles of X and Y in the definition we see that

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}.$$

From the definitions,

$$f_{XY}(x,y) = f_X(x) f_{Y|X}(y|x)$$

and so

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx \\ &= \int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y|x)\, dx \end{aligned}$$

The result follows immediately upon substitution into equation (15.3). ∎

15. Conditional Densities: Problems.

1. Let X and Y be independent, identically distributed random variables having common density function f. Find the joint density function for X and $Z = X + Y$.

2. Let X and Y be independent random variables each having the an exponential distribution with parameter λ. Find the conditional density of X given that $Z = X + Y = z$.

3. In example 15.5 show that Y has a negative binomial distribution with parameters α and $p = \beta/(1 + \beta)$.

4. Let U and V be independent random variables having the normal distribution with $\mu = 0$ and $\sigma^2 = 1$. With $-1 < \rho < 1$ set

$$Z = \rho U + \sqrt{1 - \rho^2} V.$$

(a) Find the density of Z.
(b) Find the joint density of U and Z.
(c) Find the joint density of $X = \mu_1 + \sigma_1 U$ and $Y = \mu_2 + \sigma_2 V$ where $\sigma_1 > 0$ and $\sigma_2 > 0$.
(d) Find the conditional density of Y given that $X = x$.

If X is discrete there is a reasonably direct argument that leads to the formula

$$E(X) = \sum_x x f_X(x).$$

For a continuous random variable it is less obvious how one would deduce what the definition of the 'expected value' of X 'should' be. By analogy, if X is continuous and has a density function $f_X(x)$ then it seems reasonable that the expectation of X 'should' be

$$E(X) \int_{\mathbb{R}} x f_X(x)\, dx$$

provided that the integral converges absolutely. However, instead of simply reasoning by analogy we can actually define the expectation in a slightly more general way that connects back to the more intuitive discrete case. It will turn out that the more general definition is same as the one we get by analogy, i.e., that $E(X)$ really is $\int_{\mathbb{R}} x f_X(x)\, dx$ when X has a density function.

Our more general definition is based on step functions. For clarity, we formally define step functions next.

16.1. Definition.

Let $\{I_1, I_2, \cdots, I_n\}$ be disjoint intervals in $(-\infty, \infty)$ and let $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ be real numbers. A step functionis a function of the form

$$\varphi(t) = \begin{cases} \lambda_i & \textit{if } t \in I_i \\ 0 & \textit{otherwise} \end{cases}$$

If X is a continuous random variable and if φ is a step function, then the random variable $Y = \varphi(X)$ is a discrete random variable with state space $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$. Since

$$Y = \varphi(X) = \begin{cases} \lambda_j & \text{if } X \in I_j \\ 0 & \text{otherwise} \end{cases}$$

it follows that Y has density function

$$\begin{aligned} f_Y(y) &= \mathfrak{Pr}\,(Y = y) \\ &= \begin{cases} \mathfrak{Pr}\,(X \in I_j) & \text{if } y = \lambda_j \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \int_{I_j} f(t)\,dt & \text{if } y = \lambda_i = j \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Thus the definition of expectation discrete random variaibles applies to Y and

$$\begin{aligned} E(Y) &= \sum_{j=1}^{n} \lambda_j f_Y(\lambda_j) \\ &= \sum_{j=1}^{n} \lambda_j\, \mathfrak{Pr}\,(X \in I_j) \\ &= \sum_{i=j}^{n} \lambda_i \int_{I_j} f(t)\,dt \\ &= \sum_{i=j}^{n} \int_{I_j} \varphi(t) f(t)\,dt \\ &= \int_{-\infty}^{\infty} \varphi(t) f(t)\,dt \end{aligned}$$

In particular we have established the following simple proposition:

16.2. Proposition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ be a probability space and let X be a continuous random variable defined on $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ having density function $f_X(x)$. Let $\varphi(t)$ be a step function and set $Y = \varphi(X)$. Then Y is a discrete random variable and

$$E(Y) = \int_{-\infty}^{\infty} \varphi(t) f_X(t)\,dt.$$

As we have seen, in general a continuous random variable X need not have a density function. Further if φ is continuous function then the random variable $\varphi(X)$ might not have a density function even if X has a density function. Thus a definition of $E(X)$ that relies on X having a density function is deficient since it would not necessarily extend to a random variable of the form $\varphi(X)$. Because of this, while it would be tempting to define the expectation of a continuous random variable in terms of the density function

$$\int_{-\infty}^{\infty} x f_X(x)\, dx$$

there is a slightly better approach which turns out to be equivalent if X has a density (and, of course, if the above integral converges absolutely).

16.3. Definition.

Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and let X be a continuous random variable defined on $(\Omega, \mathcal{E}, \mathfrak{Pr})$. If φ is a non-negative continous function then we define the expectation of $\varphi(X)$ to be

$$E(X) = \sup\{E(g(X)) : \text{where } g \text{ is a step function and } g(t) \le \varphi(t) \text{ for all } t\}$$

provided that the above supremum is finite.

We will begin by only considering continuous, non-negative transformations φ. However, if one decomposes a continuous function into its positive and negative parts

$$\varphi^+(t) = \max\{\varphi(t), 0\}$$

and

$$\varphi^-(t) = -\min\{\varphi(t), 0\}$$

then it is trivial that $\varphi^+ \ge 0$, $\varphi^- \ge 0$, that both φ^+ and φ^- are continuous, and that

$$\varphi(t) = \varphi^+(t) - \varphi^-(t).$$

We can then define

$$E(\varphi(X)) = E(\varphi^+(X)) - E(\varphi^-(X))$$

provided that both expectations are finite. Proceding in this manner permits results for the specialized case $\varphi \ge 0$ to be generalized to any continuous φ.

We will need list two critical facts before continuing. The first is an assertion about continuous, non-negative functions and the second about monotone sequences of non-negative functions.

16.4. Lemma.

Let $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ be a continuous, non-negative function. Then there is a sequence of non-negative step functions $\{\varphi_n(t)\}$ with the following properties:
(i) for each $t \in \mathbb{R}$, $\lim_n \varphi_n(t) = \varphi(t)$
(ii) $\varphi_n(t) \leq \varphi_{n+1}(t)$

Proof. Fix n and consider φ on the interval $[-n, n]$. Since φ is continuous it follows that φ is *uniformly* continuous on $[-n, n]$. Thus there is a $\delta > 0$ so that if $-n < s, t < n$ and $|s - t| < \delta$ then

$$|\varphi(s) - \varphi(t)| < \frac{1}{n}.$$

Pick m so small that $1/m < \delta$ and divide $[-n, n]$ into $2mn$ subintervals $\{I_k : k = 1, \cdots, 2mn\}$ each of length $1/m$. In addition, select the intervals so that

$$I_k \cap I_j = \phi \qquad \text{if } i \neq j \text{ and} \qquad \bigcup_k I_k = [-n, n].$$

On each interval I_k set

$$\lambda_k = \min\{f(s) : s \in I_k\}$$

and set

$$\varphi_n = \sum_{k=1}^{2mn} \lambda_k \chi_{I_k}$$

where χ_{I_k} is the indicator function of I_k:

$$\chi_{I_k}(t) = \begin{cases} 1 & \text{if } t \in I_k \\ 0 & \text{otherwise} \end{cases}$$

By the way that φ_n is constructed

$$\varphi_n(t) \leq \varphi(t)$$

for all t. Further, if $s \in I_k$ then

$$|\varphi_n(s) - \varphi(s)| \leq |\lambda_k - \varphi(s)| \leq \frac{1}{n}.$$

Since every $s \in [-n, n]$ must be in some I_k this proves the result.

∎

Suppose that $\varphi \geq 0$ is continous except at a finite number of points $\{p_1, \cdots, p_n\}$. A straightforward modification of the above argument will extend the result to this class. We note this fact here for future reference.

16.5. Corollary.

Suppose that φ is continous except at a finite number of points $\{p_1, \cdots, p_n\}$. Then there is a sequence of non-negative step functions $\{\varphi_n(t)\}$ with the following properties:
(i) for each $t \in \mathbb{R}$, $\lim_n \varphi_n(t) = \varphi(t)$
(ii) $\varphi_n(t) \leq \varphi_{n+1}(t)$

Note that the step functions in question may include "atoms," i.e., "steps" where the interval is a single point $[p_i, p_i]$.

The second fact needed is a theorem from advanced analysis.

16.6. Theorem. Monotone Convergence Theorem.

Let $\{\varphi_n\}$ be a monotone sequence of non-negative integrable functions and suppose that $\lim_n \varphi_n(t) = \varphi(t)$ exists for each $t \in \mathbb{R}$. Then $\int_{\mathbb{R}} \varphi(t)\ dt$ exists if and only if $\lim_n \int_{\mathbb{R}} \varphi_n(t)\ dt$ exists and is finite, in which case

$$\lim_n \int_{\mathbb{R}} \varphi_n(t)\ dt = \int_{\mathbb{R}} \varphi(t)\ dt$$

We can now deduce our first result about expectations of continuous random variables.

16.7. Theorem.

Let X be an absolutely continuous, random variable having density function $f_X(x)$ and let φ be a non-negative continuous function. Set $Y = \varphi(X)$ and suppose that $E(Y) < \infty$. Then

$$E(Y) = \int_{\mathbb{R}} \varphi(x) f(x)\, dx.$$

Further there is a sequence of discrete random variables Y_n for which
(i) $\lim_{n\to\infty} E(Y_n) = E(Y)$; and
(ii) For each $\epsilon > 0$, $\lim_{n\to\infty} \mathfrak{Pr}\left(|Y_n - Y| > \epsilon\right) = 0$.

A sequence a random variables $\{Y_n\}$ satisfying *(ii)* with respect to some random variable Y is said to *converge in probability* to Y. (Mathematicians would say $\{Y_n\}$ *converges in measure* to Y.)

converge!in probability
converge!in measure

Proof. Fix $\epsilon > 0$. Since $E(Y) < \infty$ we can choose a step function $g_0(t)$ such that $g_0(t) \leq \varphi(t)$ and

$$\begin{aligned} E(Y) &= E(\varphi(X)) \\ &= \sup\{E(g(X)) : g \quad \text{is a step function and} \quad g \leq \varphi\} \\ &\leq E(g_0(X)) + \epsilon \\ &= \int_{-\infty}^{\infty} g_0(x) f_X(x)\, dx + \epsilon \\ &\quad \text{(since } g \text{ is a step function)} \\ &\leq \int_{-\infty}^{\infty} \varphi(x) f_X(x)\, dx + \epsilon \\ &\quad \text{(since } g \leq \varphi \text{)} \end{aligned}$$

Thus, since $\epsilon > 0$ was arbitrary,

$$E(Y) \leq \int_{-\infty}^{\infty} \varphi(x) f_X(x)\, dx.$$

For the reverse inequality, we select a monotone non-decreasing sequence $\{\varphi_n\}$ of non-negative step functions such that

$$\lim_n \varphi_n(t) = \varphi(t)$$

for each t. Set $Y_n = \varphi_n(X)$. By the definition of expectation, $E(Y_n) \leq E(Y)$ for each n. By the proposition, each Y_n has finite expectation and

$$E(Y_n) = \int_{-\infty}^{\infty} \varphi_n(t) f(t)\, dt.$$

By the Monotone Convergence Theorem,

$$\lim_n \int_{-\infty}^{\infty} \varphi_n(t) f(t)\, dt = \int_{-\infty}^{\infty} \varphi(t) f(t)\, dt$$

and so

$$\begin{aligned}E(Y) &\geq \lim_n E(Y_n)\\ &= \lim_n \int_{-\infty}^{\infty} \varphi_n(t) f(t)\, dt\\ &= \int_{-\infty}^{\infty} \varphi(t) f(t)\, dt\end{aligned}$$

showing the reverse inequality. Thus

$$E(Y) = \int_{-\infty}^{\infty} \varphi(t) f(t)\, dt.$$

For the second conclusion, fix $\epsilon > 0$ and, with Y_n as above, set

$$E_n = \{\omega \in \Omega : Y(\omega) - Y_n(\omega) > \epsilon\}.$$

Note that if $\omega \in E_{n+1}$ then

$$Y(\omega) - Y_n(\omega) \geq Y(\omega) - Y_{n+1}(\omega) \geq \epsilon$$

so that $E_{n+1} \subset E_n$. Then

$$\begin{aligned}\lim_{n\to\infty} \mathfrak{Pr}\left(|Y_n - Y| > \epsilon\right) &= \lim_{n\to\infty} \mathfrak{Pr}\left(Y - Y_n > \epsilon\right)\\ &= \lim_{n\to\infty} \mathfrak{Pr}\left(E_n\right)\\ &= \mathfrak{Pr}\left(\bigcap_n E_n\right)\end{aligned}$$

But for any $\omega \in \Omega$, $Y_n(\omega) \to Y(\omega)$ and hence for n sufficiently large $\omega \notin E_n$. Thus $\cap_n E_n = \phi$. From this

$$\lim_{n\to\infty} \mathfrak{Pr}\left(|Y_n - Y| > \epsilon\right) = 0.$$

∎

16.8. Definition.

If X is an arbitrary continuous random variable, then

$$X^+ = \max\{X, 0\} \quad \textit{and} \quad X^- = -\min\{X, 0\}$$

are both non-negative random variables. If both X^+ and X^- have finite expectation, then we say that X has finite expectation and define

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

We can now show not only that our definition of expectation coincides with the more conventional one involving densities, but we can also deduce a formula for the expectation of $\varphi(X)$.

16.9. Theorem.

Let X be a continuous random variable having density function $f_X(x)$. Then X has finite expectation if and only if

$$\int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty$$

in which case

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

Proof. Take

$$\psi^+(t) = \begin{cases} t & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\psi^-(t) = \begin{cases} -t & \text{if } t \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that $X^+ = \psi^+(X)$ and that $X^- = \psi^-(X)$. Note that X has finite expectation if and only if both $E(\psi^+(X))$ and $E(\psi^-(X))$ are finite. Further

$$\begin{aligned} E(\psi^+(X)) &= \int_{-\infty}^{\infty} \psi^+(x) f_X(x)\, dx \\ &= \int_0^{\infty} x f_X(x)\, dx \end{aligned}$$

and

$$E(\psi^-(X)) = -\int_{-\infty}^{0} x f_X(x)\, dx.$$

Thus

$$\int_{-\infty}^{\infty} |x| f_X(x)\, dx = E(\psi^+(X)) + E(\psi^-(X))$$

and

$$\int_{-\infty}^{\infty} x f_X(x)\, dx = E(\psi^+(X)) - E(\psi^-(X)) = E(X)$$

as desired.

■

16.10. Theorem.

Let X and Y be continuous random variables having density functions $f_Y(x)$ and $f_Y(y)$ respectively. Then
(i) if $\lambda \in \mathbb{R}$ then $E(\lambda X) = \lambda E(X)$
(ii) if $\mathfrak{Pr}\,(X = \lambda) = 1$ then $E(X) = \lambda$
(iii) If X and Y have finite expectation, then $X + Y$ has finite expectation and

$$E(X + Y) = E(X) + E(Y)$$

(iv) if $\mathfrak{Pr}\,(X \geq Y) = 1)$ then $E(X) \geq E(Y)$
(v) $|E(X)| \leq E(|X|)$

Proof. The proof of *(i)* and *(ii)* is essentially the same as in the discrete case, with integrals replacing sums.

Conclusion *(iii)* follows in the same way as the discrete case, using Theorem 12.4 to obtain a formula for the density of $Z = X + Y$ (see the exercises). However it is also easy to deduce *(iii)* directly from the definition. First it suffices to consider the case where X and Y are non-negative. For any step function g_{X+Y} with

$$0 \leq g_{X+Y}(t) \leq t$$

it follows from the observation that $g_{X+Y}(X+Y)$ is discrete and the discrete case that

$$E(g_{X+Y}(X+Y)) = E(g_{X+Y}(X)) + E(g_{X+Y}(Y)) \leq E(X) + E(Y).$$

Since $E(X+Y)$ is the supremum over all such functions g_{X+Y} it follows that $E(X+Y) < \infty$ and

$$E(X+Y) \leq E(X) + E(Y).$$

For the reverse inequality choose step functions g_X and g_Y so that

$$0 \leq g_X(t) \leq t \quad \text{and} \quad 0 \leq g_Y(t) \leq t$$

Then if

$$(g_X \vee g_Y)(t) = \max\{g_X(t), g_Y(t)\}$$

it follows that $(g_X \vee g_Y)(t)$ is a step function. Thus

$$\begin{aligned} E(X+Y) &\geq E(g_X \vee g_Y)(X+Y)) \\ &= E(g_X \vee g_Y)(X)) + E(g_X \vee g_Y)(Y)) \\ &\geq E(g_X(X)) + E(g_Y(Y)) \end{aligned}$$

The last inequality follows from the fact that

$$(g_X \vee g_Y)(X) \geq g_X(X)$$

with probability one and

$$(g_X \vee g_Y)(Y) \geq g_Y(Y)$$

and the discrete case of *(iv)*. Now taking the suprema over all such functions g_X and g_Y we can conclude

$$E(X) + E(Y) \leq E(X+Y)$$

showing *(iii)*.

Conclusions *(iv)* and *(v)* follow exactly as in the discrete case.

∎

Of course, the whole point of the results of this section is the following theorem.

16.11. Theorem.

Let X be a continuous random variable having a density function f_X. Then X has finite expectation if and only if

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty$$

in which case

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

Some examples will help to make this more concrete.

16.12. Example.

Let X have the gamma density with parameters α and λ. Then

$$E(X) = \frac{\alpha}{\lambda}.$$

Proof. By the theorem,

$$\begin{aligned} E(X) &= \int_0^{\infty} x \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\,dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{\infty} x^\alpha e^{-\lambda x}\,dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} \\ &= \frac{\alpha}{\lambda} \end{aligned}$$

∎

16.13. Example.

Let X have the uniform density on the interval $[a, b]$. Then

$$E(X) = \frac{a+b}{2}.$$

Proof. Again from the theorem

$$\begin{aligned} E(X) &= \int_a^b x \left(\frac{1}{b-a} \right) dx \\ &= \left(\frac{1}{b-a} \right) \frac{x^2}{2} \bigg|_{x=a}^{x=b} \\ &= \frac{a+b}{2} \end{aligned} .$$

∎

Finally, not every continuous random variable has a finite expectation.

16.14. Example.

Let X have the Cauchy density

$$f_X(x) = \frac{1}{\pi(1+x^2)}.$$

Then X does not have finite expectation.

Proof. This again follows from the theorem:

$$\begin{aligned}\int_{-\infty}^{\infty} |x| \frac{1}{\pi(1+x^2)}\, dx &= \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2}\, dx \\ &= \frac{2}{\pi} \lim_{u\to\infty} \int_0^{u} \frac{x}{1+x^2}\, dx \\ &= \frac{1}{\pi} \lim_{u\to\infty} \ln(1+x^2)\Big|_{x=0}^{x=u} \\ &= \infty.\end{aligned}$$

∎

16. Expectations: Continuous Random Variables: Problems.

1. Deduce formula 15.5.

2. Deduce 15.9(iii) from 12.4.

3. Let X be a normally distributed random variable with $\mu = 0$ and $\sigma = 1$. Find $E(X)$.

4. Let X be normally distributed with $\mu = 0$. Find the mean and variance of each of the following:
(a) $|X|$
(b) X^2
(c) e^{tX}

5. Let X be a non-negative continuous random variable having density function f_X and distribution function F_X. Show that X has finite expectation if and only if

$$\int_0^\infty (1 - F_X(x))\, dx < \infty$$

in which case

$$E(X) = \int_0^\infty (1 - F_X(x))\, dx.$$

6. Let X have the gamma distribution with parameters α and λ. For what values of t does

$$Y = e^{tx}$$

have finite expectation? For those values of t find

$$E\left(e^{tX}\right)$$

.

17. Moment Generating Functions

Laplace transform

The probability generating function is useful in studying discrete random variables because it's behavior is "regular" with respect differentiation. In a similar way, moment generating functions are useful in studying continuous random variables.

17.1. Definition.

Let X be a continuous random variable having probability density function $f_X(t)$ and suppose that

$$\int_{\mathbb{R}} e^t f_X(t)\ dt < \infty$$

Then the moment generating function for X is the random variable

$$M_X(t) = E(e^{tX}) = \int_{\mathbb{R}} e^{tx} f_X(x)\ dx$$

Of course $M_X(-t)$ is the Laplace transform of the probability density function for X.

17.2. Example.

Suppose that X is an exponential random variable having parameter λ. Then the moment generating function of X is

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

Proof. By definition,

$$\begin{aligned}
M_X(t) &= E({}^{tX}) \\
&= \lambda \int_0^\infty e^{tx} e^{-\lambda x}\, dx \\
&= \lambda \int_0^\infty e^{-(\lambda - t)x}\, dx \\
&= \qquad \text{making the change of variables } u = (\lambda - t)x \\
&= \frac{\lambda}{\lambda - t} \int_0^\infty e^{-u}\, du \\
&= \frac{\lambda}{\lambda - t}
\end{aligned}$$

■

17.3. Example.

Suppose that $a < b$ and that X is a random variable which has a uniform density on $[a, b]$, i.e., that X has probability density function

$$f_X(x) = \begin{cases} \frac{1}{(b-a)} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Then the moment generating function of X is

$$e^t \frac{e^b - e^a}{t(b-a)}$$

Proof. Applying the definition

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) \\ &= \int_a^b e^{tx} f_X(x)\, dx \\ &= \frac{1}{b-a} \int_a^b e^{tx}\, dx \\ &= \frac{1}{b-a}\frac{1}{t}\, e^{tx}\Big|_{x=a}^{x=b} \\ &= \frac{1}{b-a}\frac{1}{t}\left(e^{bt} - e^{at}\right) \\ &= e^t \frac{e^b - e^a}{t(b-a)} \end{aligned}$$

■

Since one equivalent definition of the moment generating is

$$M_X(t) = E\left(e^{tX}\right)$$

it is possible to find the moment generating function of discrete random variables. Where they both exist, there is a simple formula that connect moment generating and probability generating functions.

17.4. Theorem.

Let X be a discrete random variable and suppose that X has a probability generating function $\Phi_X(t)$. Then X has a moment generating function given by the formula

$$M_X(t) = \Phi_X(e^t)$$

The proof is immediate from the definitions.

Moment generating functions are so named precisely because of their regular behavior relative to the moments of the random variable X.

17.5. Theorem.

Let X be a random variable having finite moment generating function $M_X(t)$. Then

$$M_X^{(n)}(0) = E(X^n) \qquad *$$

Further if either side of $(*)$ is finite, then the other side is finite and equality holds.

Proof. Suppose for the moment that $M_X'(t) < \infty$ and that M_X' is continuous. Then

$$\begin{aligned} M_X'(t) &= \frac{d}{dt} E(e^{tX}) \\ &= \frac{d}{dt} \int_{\mathbb{R}} e^{tx} f_X(x)\, dx \\ &= \int_{\mathbb{R}} \frac{d}{dt} e^{tx} f_X(x)\, dx \\ &= \int_{\mathbb{R}} x e^{tx} f_X(x)\, dx \end{aligned}$$

The interchange of the integral and differential operator is justified by a theorem in higher analysis. Evaluating the derivative at $t = 0$ gives the result. A simple induction extends the result for arbitrary values of n.

∎

17.6. Corollary.

Let X be a random variable having finite mean μ and variance σ^2. Then

$$M_X'(0) = \mu \quad \text{and} \quad M_X''(0) = \sigma^2 + \mu^2.$$

Proof. It follows from the previous theorem that $\mu = E(X) = M'_X(0)$. Since

$$\begin{aligned} \sigma^2 &= E\left((X - E(X)\right) \\ &= E\left(X^2 - XE(X) + (E(X))^2\right) \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2) \\ &= E(X^2) - \mu^2 \end{aligned}$$

Rearranging gives the conclusion.

■

17.7. Theorem.

Let X be a random variable having finite moment generating function $M_X(t)$ and let a and b be real numbers. If $Y = aX + b$ then the moment generating function $M_Y(t)$ for Y is

$$M_Y(t) = e^{bt} M_X(at)$$

Proof. This follows readily from the definition:

$$\begin{aligned} M_Y(t) &= E\left(e^{tY}\right) \\ &= E\left(e^{atX+bt}\right) \\ &= e^{bt} E\left(e^{atX}\right) \\ &= e^{bt} M_X(at) \end{aligned}$$

■

17.8. Example.

If X is a random variable having an exponential distribution with parameter λ then

$$\mu = \frac{1}{\lambda} \quad \text{and} \quad \sigma^2 = \frac{1}{\lambda}$$

Proof. Since

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

it follows that

$$M_X'(t) = \frac{\lambda}{(\lambda - t)^2}$$

and

$$M_X''(t) = \frac{2\lambda}{(\lambda - t)^3}$$

Thus

$$M_X'(0) = \frac{1}{\lambda}$$

and

$$M_X''(0) = \frac{2}{\lambda^2}.$$

The conclusion is immediate from the above and the preceding corollary. ∎

17.9. Example.

Let Z be a normally distributed random variable having mean zero and standard deviation one. Then

$$M_Z(t) = e^{\frac{t^2}{2}}.$$

Proof. The density function for Z is

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

so

$$\begin{aligned}
M_Z(t) &= E\left(e^{tZ}\right) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz} e^{-\frac{z^2}{2}}\, dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}+tz}\, dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(\frac{z^2}{2}-tz+\frac{t^2}{2})} e^{\frac{t^2}{2}}\, dz \\
&= e^{\frac{t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{(z-t)^2}{2}}\, dz \\
&= e^{\frac{t^2}{2}}
\end{aligned}$$

17.10. Corollary.

Let X be a normally distributed random variable having mean μ and standard deviation σ. Then

$$M_X(t) = e^{\mu t} e^{\frac{\sigma^2 t}{2}}$$

and $E(X) = \mu$ and the variance of X is σ^2.

17.11. Theorem.

Let X and Y be independent random variables having moment generating functions $M_X(t)$ and $M_Y(t)$ respectively. Then

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

Proof. Let $Z = X + Y$, so that the density function for Z is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx.$$

Then

$$
\begin{aligned}
M_Z(t) &= E\left(e^{tZ}\right)\\
&= \int_{-\infty}^{\infty} e^{tz} f_Z(z)\, dz\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{tz} f_X(x) f_Y(z-x)\, dx\, dz\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t(z-x)} e^{tx} f_X(x) f_Y(z-x)\, dx\, dz\\
&= \int_{-\infty}^{\infty} e^{tx} f_X(x) \int_{-\infty}^{\infty} e^{t(z-x)} f_Y(z-x)\, dz\, dx\\
&= \int_{-\infty}^{\infty} e^{tx} f_X(x) \int_{-\infty}^{\infty} e^{tu} f_Y(u)\, du\, dx\\
&= \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx \int_{-\infty}^{\infty} e^{tu} f_Y(u)\, du\\
&= M_X(t) M_Y(t)
\end{aligned}
$$

■

17.12. Example.

Let X_1 and X_2 be independent, normally distributed random variables having parameters μ_1, σ_1^2 and μ_2, σ_2^2 respectively. Then $X + Y$ is a normally distributed random variable having parameters $\mu_1 + \mu_2$ and $\sigma_1^2 + \sigma_2^2$.

Proof. We know that

$$M_{X_1}(t) = e^{\mu_1 t} e^{\frac{\sigma_1^2 t}{2}}$$

and

$$M_{X_2}(t) = e^{\mu_2 t} e^{\frac{\sigma_2^2 t}{2}}.$$

Thus

$$
\begin{aligned}
M_{X_1+X_2}(t) &= M_{X_1}(t) M_{X_2}(t)\\
&= e^{\mu_1 t} e^{\frac{\sigma_1^2 t}{2}} e^{\mu_2 t} e^{\frac{\sigma_2^2 t}{2}}\\
&= e^{(\mu_1+\mu_2)t} e^{\frac{(\sigma_1^2+\sigma_2^2)t}{2}}.
\end{aligned}
$$

Since the latter is the moment generating function of a normally distributed random variable with parameters $\mu_1 + \mu_2$ and $\sigma_1^2 + \sigma_2^2$ this proves the result.

∎

moment generating function!gamma random variables

17. Moment Generating Functions: Problems.

1. Let X have the Gamma density with parameters α and λ. Show that

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha}.$$

2. Show that if X and Y are independent exponentially distributed random variables with common parameter λ and if $Z = X + Y$ then Z has a gamma distribution with parameters λ and $\alpha = 2$.

3. Let X have the Gamma distribution with parameters α_X and λ and let Y have the Gamma distribution with parameters α_Y and λ. If X and Y are independent and if $Z = X + Y$ show that Z has the Gamma distribution with parameters $\alpha_X + \alpha_Y$ and λ.

4. Let X be a random variable having moment generating function $M_X(t)$ that is finite for all t. Show that

$$\mathfrak{Pr}\left(X \geq x\right) \leq e^{-tx} M_X(t)$$

for all $t \geq 0$.

Hint: Fix x and let $t > 0$ be any real number. Define a new random variable

$$Y = \begin{cases} 1 & \text{if } e^{-tx}e^{tX} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Show that $E(Y) \leq E\left(e^{-tx}e^{tX}\right)$.

5. Let X have the Gamma distribution with parameters α and λ. Show that

$$\mathfrak{Pr}\left(X \geq \frac{2\alpha}{\lambda}\right) \leq \left(\frac{2}{e}\right)^{\alpha}.$$

Recall our earlier definition of expected value.

18.1. Definition.

The expectation of a random variable is the number $E(X)$. The expectation gives a measure of the 'central tendency' of X. The mean of a random variable X is the number

$$\mu = E(X)$$

provided that X has finite expectation.

A closely related measure is the *variance* of X, σ^2.

18.2. Definition.

Let X be a random variable. If

$$\sigma^2 = E\left((X - E(X))^2\right) < \infty$$

then the variance of X is the number

$$\sigma^2 = E\left((X - E(X))^2\right)$$

We recall that an easy computation relates σ^2 and $E(X^2)$:

$$\begin{aligned} \sigma^2 = E\left((X - E(X))^2\right) &= E\left(X^2 - 2XE(X) + \mu^2\right) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

or equivalently,

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

provided that $E(X^2) < \infty$. Random variables that have a finite mean and a finite variance are of especial importance in applications. This section is devoted to some important inequalities pertaining to random variables having finite mean or finite mean and variance. These inequalities, in turn, often turn out to have surprisingly powerful consequences in applications.

The first theorem in this section, the Cauchy-Schwarz Inequality, is similar to the inequality of the same name relating the 'inner product' of two vectors with their norm. Recall that if X and Y are vectors, then

$$|\langle X, Y\rangle| \leq \|X\|^2\|Y\|^2$$

where $\langle X, Y\rangle$ is the inner product of the vectors X and Y. This inequality has a long history, with early versions in the works of Augustin Cauchy (1789-1857) , Herman Schwarz (1843-1921) and Viktor Bunyakovsky (1804-1889). The first statement and proof in its modern form appears to have been due to Hermann Weyl (1885-1955). There are many ways that the geometry of n-dimensional vector spaces are similar to the geometry of random variables having finite second moment – although as one might expect the latter is somewhat more complex!

18.3. Theorem. Cauchy-Schwarz Inequality.

Suppose that X and Y are random variables and that both $E(X^2) < \infty$ and $E(Y^2) < \infty$. Then

$$E(XY)^2 \leq E(X^2)E(Y^2).$$

Equality holds if and only if there is a real number λ for which $X = \lambda Y$ with probability one.

Proof. First observe that if either $E(X^2) = 0$ or $E(Y^2) = 0$ then the conclusion is trivial. Thus we may assume without loss of generality that both $E(X^2) > 0$ and $E(Y^2) > 0$.

Let $\lambda > 0$ be any real number and observe that

$$0 \leq E\left((X + \lambda Y)^2\right) = E(X^2) + 2\lambda E(XY) + \lambda^2 E(Y^2)$$

The above quadratic in λ is minimized when

$$\lambda = -\frac{E(XY)}{E(Y^2)}$$

Thus, with this value for λ

$$\begin{aligned} 0 &\le E(X^2) - 2\frac{E(XY)}{E(Y^2)}E(XY) + \left(\frac{E(XY)}{E(Y^2)}\right)^2 E(Y^2) \\ &= E(X^2) - \frac{E(XY)^2}{E(Y^2)} \end{aligned}$$

From this

$$E(XY)^2 \le E(X^2)E(Y^2).$$

■

The second major inequality in this section is due to the Russian mathematician Pafnuty Lvovich Cebysev (1821-1894). While Čebysev is probably best known for the following inequality, his interests were wide-ranging and included prime number theory, mechanics, quadratic forms and integration. The transliteration of Čebysev's name from the Russian Чебыщвё is particularly challenging; in addition to the spelling we have used, you will variously find his name spelled as Tchebycheff, Chebyshev or even Tchebyscheff.

18.4. Theorem. Čebysev's Inequality.

Suppose that X is a random variable having finite mean μ and finite variance σ^2. Then for any $\epsilon > 0$

$$\mathfrak{Pr}\left(|X - \mu| > \epsilon\right) \le \frac{\sigma^2}{\epsilon^2}$$

The proof of Čebysev's inequality is an easy consequence of an inequality due to Markov. Andrei Markov (1856-1922) was a student of Čebysev. In addition to extending Čebysev's work in number theory and continued fractions, he is probably best known for the random processes called Markov Chains that we will study later in this text. The proof below of Markov's inequality is deceptively simple: the hard work has been put into the theorems in the previous sections on expectations in the discrete and continuous cases. Markov's inequality can provide some surprisingly sharp estimates, as we shall see in the section on Chernoff Bounds.

18.5. Theorem. Markov's Inequality.

Suppose X is a non-negative random variable having finite mean μ and suppose $\epsilon > 0$ is any positive real number. Then

$$\mathfrak{Pr}\,(X \geq \epsilon) \leq \frac{\mu}{\epsilon}$$

Proof. Let U be the random variable

$$U = \begin{cases} \epsilon & \text{if } X \geq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

Then $U \leq X$ so

$$\begin{aligned} \epsilon\,\mathfrak{Pr}\,(X \geq \epsilon) &= E(U) \\ &\leq E(X) \\ &= \mu \end{aligned}$$

from which the result is immediate. ∎

Proof of Čebysev's Inequality. Let Y be the random variable defined by

$$Y = (X - \mu)^2$$

Then from Markov's inequality,

$$\mathfrak{Pr}\,(Y \geq \epsilon^2) \leq \frac{E(y)}{\epsilon^2}$$

which reduces to the result. ∎

18. Estimation: Problems.

1. Let X be a random variable having finite mean μ and finite variance σ^2.

(a) Verify that at least 75% of the observations must fall within two standard deviations of the mean.

(b) For $n \geq 2$ verify that

$$\mathfrak{Pr}\left(|X-\mu| \leq \sigma n\right) \geq 1 - \frac{1}{n^2}$$

for any natural number n.

2. The voltage in a certain circuit is a random variable with mean 120 and standard deviation 5. Expensive equipment will be damaged if the voltage is not between 112 and 128. Use Čebyshev's inequality to estimate the liklihood of damage occuring.

3. A binary transmission channel will erroneously transmit a bit of data with probability of 10%. Use Čebyshev's inequality to estimate the probability that there are between 4 and 16 errors in 100 bits of data.

4. About 2% of a certain type of RAM chips are defective. A manufacturer needs 50 chips for a certain circuit board. How many should she buy in order for there to be at least a 99% chance of having at least 49 working chips?

5. A pharmaceutical company manufactures viagra pills which contain an average of 50 mg of viagra with a standard deviation of 0.75 mg. Find the proportion of the pills that have between 47 and 53 mg.

6. Let X and Y be random variables having finite second moment. Verify the parallelogram law:

$$E((X+Y)^2) + E((X-Y)^2) = 2E(X^2) + 2E(Y^2)$$

19. Sampling Distributions.

In many situations statisticians will be interested in studying a population which they have reason to believe is normally distributed but for which the parameters μ and σ^2 are unknown. An early example arose in the Prussian army in which the quartermasters wished to predict how many uniforms of each size needed to be kept in inventory. Since height and weight – and hence uniform size – were thought to be distributed normally among the soldierly population, this amounted estimating probabilities like

$$\mathfrak{Pr}\,(a - 1/2 < X \leq a + 1/2)$$

where X measures “uniform size" and “a" represents a uniform size. The above probability then tells the quartermaster what proportion of the inventory should be size “a." If sizes are normally distributed with parameters μ and σ^2 then normalizing the above

$$\mathfrak{Pr}\left(\frac{a - 1/2 - \mu}{\sigma} < Z \leq \frac{a + 1/2 - \mu}{\sigma}\right) \tag{19.1}$$

where

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

is a normal random variable with $\mu = 0$ and $\sigma^2 = 1$ reduces the problem to calculating areas under the “standard" normal curve

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

In order to the normalization to Z it is necessary to know the values of μ and σ^2. Learning the exact values would, of course, involve a census of the entire soldierly population, present, past and future. Since this is impractical, the problem reduces to *estimating* the values of μ and σ with sampling data. In this section we will gather together some important sampling distributions that help to address this problem. Among the random variables we will need to consider are

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \qquad \text{sample mean}$$

$$S^2 = \frac{1}{n-1}\sum_{i=i}^{n}\left((X_i - \bar{X})^2\right) \qquad \text{sample variance}$$

$$t = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \qquad \text{Student's } t$$

While our ultimate interest in this section is the case when the underlying distribution is normal, we collect some useful facts about the random variables $\bar{X}$ and S^2.

19.1. Theorem.

Let $\{X_1, X_2, \cdots, X_n\}$ be independent and identically distributed random variables having common mean μ and common variance σ^2. Then the random variable

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

has mean μ and variance σ^2/n. If S^2 is the random variable

$$S^2 = \frac{1}{n-1}\sum_{i=i}^{n}((X_i - \bar{X})^2)$$

then $E(S^2) = \sigma^2$.

19.2. Definition.

We say that $\bar{X}$ is an **unbiased estimator for** *μ since $E(\bar{X}) = \mu$ and we say that S^2 is an* **unbiased estimator for** *σ^2 since $E(S^2) = \sigma^2$.*

Note that in contrast with how σ^2 is calculated, we must divide S^2 by $n-1$ in order for S^2 to be an unbiased estimator for σ^2.

Proof. First note that

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=i}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=i}^{n} E(X_i) \\ &= \mu. \end{aligned}$$

For the rest of the conclusions we first compute $E(\bar{X}^2)$:

$$\begin{aligned}
E\left(\bar{X}^2\right) &= E\left(\bar{X}\bar{X}\right) \\
&= \frac{1}{n^2} E\left(\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} X_j\right)\right) \\
&= \frac{1}{n^2} E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} E(X_i X_j) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} (n-1)\mu^2 + \mu^2 + \sigma^2 \\
&= \mu^2 + \frac{\sigma^2}{n}.
\end{aligned}$$

Using the fact that

$$\text{var}(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2$$

we may immediately conclude that the variance of $\bar{X}$ is σ^2/n.

Next calculate $E(S^2)$:

$$\begin{aligned}
E(S^2) &= \frac{1}{n-1} E\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right) \\
&= \frac{1}{n-1} E\left(\sum_{i=1}^{n} X_i^2 - 2X_i\bar{X} + \bar{X}^2\right) \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \left(E(X_i^2) - 2E(X_i\bar{X}) + E(\bar{X}^2)\right) \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \left(\mu^2 + \sigma^2 - 2E(X_i\bar{X}) + \mu^2 + \frac{\sigma^2}{n}\right) \qquad (19.2.)
\end{aligned}$$

Now consider the middle term in the sum:

$$\begin{aligned}
E(X_i\bar{X}) &= \frac{1}{n}E\left(\sum_{j=1}^{n} X_iX_j\right)\\
&= \frac{1}{n}\sum_{j=1}^{n} E(X_iX_j)\\
&= \frac{1}{n}\left((n-1)\mu^2+\mu^2+\sigma^2\right)\\
&= \mu^2+\frac{1}{n}\sigma^2.
\end{aligned}$$

Substituting into (19.2):

$$\begin{aligned}
E(S^2) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\mu^2+\sigma^2-2E(X_i\bar{X})+\mu^2+\frac{\sigma^2}{n}\right)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sigma^2-\frac{1}{n}\sigma^2\right)\\
&= \frac{n}{n-1}\left(1-\frac{1}{n}\right)\sigma^2\\
&= \sigma^2
\end{aligned}$$

as desired.

∎

Since $E(\bar{X}) = \mu$ and $E(S^2) = \sigma^2$, it seems reasonable that $\bar{X} \approx \mu$ and $S^2 \approx \sigma^2$ for large values of n. This intuitive notion is what is behind “laws of large numbers." Laws of large numbers refer to a class of theorems that determine when a statement like

$$\bar{X} \to E(\bar{X}) \quad \text{as} \quad n \to \infty$$

is true. Čebysev’s inequality provides a weak form of a law of large numbers for the mean that asserts that the above statement is true “in probability."

19.3. Corollary.

$\{X_1, X_2, \cdots, X_n\}$ be independent and identically distributed random variables having common mean μ and common variance σ^2. Then for any $\epsilon > 0$

$$\mathfrak{Pr}\left(|\bar{X} - \mu| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

Proof. This is immediate from Čebysev's inequality and the previous theorem. ∎

The above not only says that $\mathfrak{Pr}\left(|\bar{X} - \mu| > \epsilon\right) \to 0$ as $n \to \infty$, it even gives an estimate as to how large n must be in order to be "sure" (up to some small chance of error $\sigma^2/2\epsilon^2$) that $|\bar{X} - \mu| < \epsilon$. Returning to equation (19.1) and the Prussian quartermaster's problem, applying weak laws of large numbers would permit us to estimate the desired probabilities by replacing the unknown parameters μ and σ with $\bar{X}$ and $\sqrt{S^2}$:

$$\begin{aligned}
&\mathfrak{Pr}\left(\frac{a - 1/2 - \mu}{\sigma} < \frac{X - \mu}{\sigma} \leq \frac{a + 1/2 - \mu}{\sigma}\right) \\
\approx\ &\mathfrak{Pr}\left(\frac{a - 1/2 - \bar{X}}{\sqrt{S^2}} < \frac{X - \bar{X}}{\sqrt{S^2}} \leq \frac{a + 1/2 - \bar{X}}{\sqrt{S^2}}\right)
\end{aligned}$$

However, the Prussian quartermaster was actually able to infer that the underlying distributions were *normally distributed*. In this case we can say more about the distributions of $\bar{X}$ and S^2 and can actually find the density function for

$$\frac{X - \bar{X}}{\sqrt{S^2}}$$

19.4. Corollary.

Let $\{X_1, X_2, \cdots, X_n\}$ be independent and identically distributed normal random variables having common mean μ and common variance σ^2. Then the random variable

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

is normally distributed with mean μ and variance σ^2/n. If S^2 is the random variable

$$S^2 = \frac{1}{n-1}\sum_{i=i}^{n} \left((X_i - \bar{X})^2\right)$$

then S^2 has a gamma distribution with parameters $\alpha = n/2$ and $\lambda = n/2\sigma^2$.

Proof. Since each $X_i - \bar{X}$ is normally distributed with mean $\mu = 0$ and, by the proof of Theorem 1, variance $\sigma^2(n-1)/n$, it follows from 10.8 and the definitions that $(X_i - \bar{X})^2$ is a gamma random variable with parameters $\alpha = 1/2$ and

$$\lambda = \frac{n}{2(n-1)\sigma^2}.$$

Then the random variable

$$\sum_{i=i}^{n} \left((X_i - \bar{X})^2\right)$$

is necessarily a gamma random variable with parameters $\alpha = n/2$ and

$$\lambda = \frac{n}{2(n-1)\sigma^2}$$

from which S^2 is a gamma random variable with parameters $\alpha = n/2$ and $\lambda = n/2\sigma^2$. ∎

Finally we turn to the question of the distribution of $\bar{X}/\sqrt{S^2}$. For simplicity, we normalize both the numerator and denominator.

19.5. Theorem.

Let X be a normally distributed random variable having mean $\mu = 0$ and variance $\sigma^2 = 1$ and let Y be a Chi-squared random variable with n degrees of freedom, i.e., Y has a gamma distribution with parameters $\alpha = n/2$ and $\lambda = 1/2$. Then the random variable

$$T = \frac{X}{\sqrt{Y/n}}$$

has the density function

$$f_T(t) = \frac{\Gamma\left(\frac{1}{2}(n+1)\right)}{\sqrt{n\pi}\Gamma(n/2)\left(1 + (t^2/n)\right)^{(n+1)/2}}$$

Proof. We will first consider

$$W^2 = \frac{X^2}{Y}.$$

Since X is normally distributed, X^2 is a gamma random variable with parameters $\alpha = 1/2$ and $\lambda = 1/2$. Then W^2 is the ratio of two gamma distributions and so, via 14.16,

$$f_{W^2}(z) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{z^{\alpha_2 - 1}}{(z+1)^{\alpha_1 + \alpha_2}} \qquad 0 < z < \infty$$

where $\alpha_1 = n/2$ and $\alpha_2 = 1/2$. Using $\Gamma(1/2) = \sqrt{\pi}$ this then simplifies to

$$f_{W^2}(z) = \frac{\Gamma((1+n)/2)}{\sqrt{\pi}\Gamma(n/2)} \frac{z^{-1/2}}{(z+1)^{(n+1)/2}} \qquad 0 < z < \infty$$

Thus

$$\mathfrak{Pr}\,(W^2 < z) = \int_0^z f_{W^2}(s)\,ds.$$

On the other hand,

$$\mathfrak{Pr}\,(W^2 < z) = \mathfrak{Pr}\,(-\sqrt{z} < W < \sqrt{z})$$

and so

$$\mathfrak{Pr}\,(-\sqrt{z} < W < \sqrt{z}) = \int_0^z f_{W^2}(s)\,ds.$$

Student's t distribution
Gosset, William Sealey
Neyman, Jerzy
Fisher, Ronald

Differentiating both sides gives

$$\frac{1}{2\sqrt{z}}\left(f_W(-\sqrt{z})+f_W(\sqrt{z})\right)=f_{W^2}(z).$$

Setting $z=w^2$ gives

$$\frac{1}{2}\left(f_W(-w)+f_W(w)\right)=|w|f_{W^2}(w^2).$$

One can readily show that f_W is symmetric about $w=0$ and so this becomes

$$\begin{aligned}f_W(w)&=|w|f_{W^2}(w^2)\\&=|w|\frac{\Gamma((n+1)/2)}{\sqrt{\pi}\Gamma(n/2)}\frac{|w|^{-1}}{(w^2+1)^{(n+1)/2}}\\&=\frac{\Gamma((n+1)/2)}{\sqrt{\pi}\Gamma(n/2)}\frac{1}{(w^2+1)^{(n+1)/2}}\end{aligned}$$

To complete the proof, note that $T=\sqrt{n}W$ and hence

$$\begin{aligned}f_T(t)&=\frac{1}{\sqrt{n}}f_W(t/\sqrt{n})\\&=\frac{\Gamma\left(\frac{1}{2}(n+1)\right)}{\sqrt{n\pi}\Gamma(n/2)\left(1+(t^2/n)\right)^{(n+1)/2}}\end{aligned}$$

∎

The above distribution is most often called “Student's t” distribution. It was devised by William Sealey Gosset, a chemist and statistician employed by the Guinness Brewery in Dublin. He devised important statistical methods for analyzing the small sample sizes available for monitoring quality control at the brewery. While he had degrees from Oxford in both Chemistry and Mathematics, in 1906-07 he traveled to University College in London to study further under Karl Pearson. He continued his correspondences with Pearson and with Jerzy Neyman and Ronald Fisher.

Gosset published his statistical work under the pseudonym "A. Student," hence the name for the distribution. Most statistics books today will contain "Student's t-tables" for working with small samples, so that today he is much better known by his pseudonym than by his real name. Tradition has it that he published under a pseudonym because his employer, while encouraging and funding his research, discouraged scholarly publication; however there is no documentary evidence of this.

modulus

20. Characteristic Functions

Moment generating functions

$$M_X(t) = E(e^{tX})$$

are useful in that the existence of the n^{th} moment $E(X^n)$ exactly corresponds to the existence of the n^{th} derivative of M_X evaluated at 0, i.e.,

$$E(X^n) = M_X^{(n)}(0)$$

However, moment generating functions have the limitation that the integral or sum that is implicit in the definition may not converge absolutely. The characteristic function – analogous to the Fourier transform – provides a more regularized approach that avoids this difficulty. While slightly more complex computationally, characteristic functions provide a powerful tool in that the characteristic function will always exist, provided that X has a density function.

Since characteristic functions involve complex variables, we begin with a quick summary of the basic facts that we will need.

A *complex number* z is a number of the form $z = x + iy$ where x and y are real numbers and $i = \sqrt{-1}$. The *modulus* of a complex number is

$$|z| = \sqrt{x^2 + y^2}.$$

The *modulus* of a complex number is analogous to the absolute value of a real number. There is also a one-to-one correspondence between complex numbers $z = x + iy$ and points on the plane where the modulus of $z = x + iy$ is similar to the norm $\|v\|$ of the vector originating at the origin and terminating at (x, y). However the interaction between the algebra of complex numbers and the metric geometry of the modulus is somewhat different than conventional Euclidian geometry on the plane.

Starting with a complex number z we can define e^z using the Taylor's series

$$z = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

With this definition it is possible to show that the usual properties of the exponential function apply, it viz.,

$$e^{u+v} = e^u e^v \qquad \frac{d}{dz} e^{\lambda z} = \lambda e^{\lambda z}$$

For example, for the former formula:

$$\begin{aligned}
e^{u+v} &= \sum_{n=0}^{\infty} \frac{(u+v)^n}{n!} \\
&= \sum_{n=0}^{\infty} \sum_{m=0}^{n} \frac{1}{n!} \binom{n}{m} u^m v^{n-m} \\
&= \sum_{m=0}^{\infty} \sum_{n=m}^{\infty} \frac{1}{m!(n-m)!} u^m v^{n-m} \\
&= \sum_{m=0}^{\infty} \frac{u^m}{m!} \sum_{n=m}^{\infty} \frac{v^{n-m}}{(n-m)!} \\
&= \sum_{m=0}^{\infty} \frac{u^m}{m!} \sum_{j=0}^{\infty} \frac{v^j}{(j)!} \\
&= e^u e^v
\end{aligned}$$

Examining the Taylor's series for e^{it} where $t \in \mathbb{R}$ reveals an important formula due to De Moivre.

$$\begin{aligned}
e^{it} &= \sum_{n=0}^{\infty} \frac{(it)^n}{n!} \\
&= 1 + \frac{it}{1} - \frac{t^2}{2!} - \frac{it^3}{3!} + \frac{t^4}{4!} + \cdots \\
&= \left(1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \cdots\right) + i\left(\frac{t}{1!} - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots\right) \\
&= \cos(t) + i\sin(t).
\end{aligned}$$

From this we can immediately deduce that $\left|e^{it}\right| = 1$. The above also shows that

$$\begin{aligned}
e^{s+it} &= e^s e^{it} \\
&= e^s\left(\cos(t) + i\sin(t)\right)
\end{aligned}$$

Finally, given a complex number $z = x + iy$ it is an easy calculation to see that $e^{u+iv} = z$ whenever $u = \ln(|z|)$ and $v = \arctan(y/x)$ (this is a simple change to polar coordinates). There are, of course, multiple values of v for which $v = \arctan(y/x)$. If we assume that $-\pi \leq v < \pi$ then we can define $\ln(z) = u + iv$ where $u = \ln(|z|$ and

$$v = \arctan\left(\frac{Im(z)}{Re(z)}\right)$$

and $-\pi \leq v < \pi$. With this definition $\ln(z)$ is a well-defined single-valued function and $e^{\ln(z)} = z$. Differentiating on both sides gives

$$\begin{aligned}
1 &= \frac{d}{dz}e^{\ln(z)} \\
&\quad \text{apply the chain rule} \\
&= \frac{d}{dz}\ln(z)e^{\ln(z)} \\
&\quad \text{apply } e^{\ln(z)} = z \\
&= \frac{d}{dz}\ln(z)z
\end{aligned}$$

from which

$$\frac{d}{dz}\ln(z) = \frac{1}{z}.$$

The point of these more-or-less heuristic arguments is that both e^z and $\ln(z)$ are well-defined and have the usual relationships and derivatives for complex arguments.

20.1. Definition.

If X is a random variable a density function $f_X(x)$ then the characteristic function of X is the function $\varphi_X(t)$ given by

$$\varphi_X(t) = E\left(e^{itX}\right)$$

If X is discrete, then

$$\varphi_X(t) = \sum_{x=-\infty}^{\infty} e^{ixt} f_X(x)$$

while if X is continuous then

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} f_X(x)\, dx.$$

Characteristic function!uniform distribution

In either case, since $\left|e^{itx} f_X(x)\right| = f_X(x)$, it follows that $E\left(e^{itX}\right)$ is finite and well-defined for all $t \in \mathbb{R}$. The reason that the characteristic function is always finite and well-defined is because e^{it} has modulus one for all t. In contrast, e^t is unbounded and hence the moment generating function can fail to exist because the associated integral or sum might diverge.

If X in addition does have a moment generating function $M_X(t)$, however, then

$$\varphi_X(t) = M_X(it).$$

This means that we can immediately deduce the characteristic functions for several classes of random variables. For example,if X is normally distributed with mean μ and standard deviation σ^2 then

$$\varphi_X(t) = \exp\left(it\mu - \frac{\sigma^2 t^2}{2}\right)$$

while if X is exponentially distributed with parameter λ then

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}$$

and if X is uniformly distributed on $(-1, 1)$ then

$$\varphi_X(t) = \frac{\sin(t)}{t}.$$

Of course, the characteristic function can also be related to the probability generating function

$$\Phi_X(t) = E(t^X)$$

that we considered for discrete random variables. Indeed,

$$\Phi_X(e^t) = M_X(t) \quad \text{and so} \quad \varphi_X(t) = \Phi_X(e^{it})$$

Thus if X is a binomial random variable with parameters n and p then the characteristic function for X is

$$\varphi_X(t) = (pe^{it} + 1 - p)^n$$

and if X is has a Poisson distribution with parameter λ then

$$\varphi_X(t) = \exp(\lambda(e^{it} - 1)$$

If X and Y are independent random variables, then so are the random variables e^{itX} and e^{itY} and so

$$\begin{aligned}\varphi_{X+Y}(t) &= E\left(e^{it(X+Y)}\right)\\ &= E\left(e^{itX}e^{itY}\right)\\ &= E\left(e^{itX}\right)E\left(e^{itY}\right)\\ &\varphi_X(t)\varphi_Y(t)\end{aligned}$$

or

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$$

Further, the moments of X can be inferred from the derivatives of φ_X provided that the moments exist (or equivalently provided that the derivatives exist). For example,

$$\begin{aligned}\frac{d}{dt}\varphi_X(t)|_{t=0} &= \frac{d}{dt}E\left(e^{itX}\right)\\ &= E\left(iXe^{itX}\right)|_{t=0}\\ &= iE(X)\end{aligned}$$

provided that the derivatives exist (or that the first moment is finite). Similarly,

$$\varphi_X^{(n)}(0) = i^n E(X^n)$$

If we can expand $M_X(t)$ in a power series on some open interval about $t \in \mathbb{R}$

$$M_X(t) = \sum_{n=0}^{\infty} \frac{E(X^n)}{n!}t^n$$

then we can also expand the characteristic function

$$\varphi_X(t) = \sum_{n=0}^{\infty} \frac{i^n E(X^n)}{n!}t^n$$

on the same open interval about t.

Finally, note that since

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

and

$$\varphi_X''(0) = i^2 E(X^2)$$

it follows that

$$\varphi_X''(0) = -(\sigma^2 + \mu^2)$$

20. Characteristic Functions: Problems.

1. Suppose that X is a random variable and that $M_X(t) < \infty$ for all t. Show that for all $t \geq 0$

$$\mathfrak{Pr}\left(X \geq x\right) \leq e^{-tx} M_X(t).$$

(Hint: try using an argument similar to the one used for Čebysev's inequality.)

2. Let X be a gamma random variable having parameters α and λ. Show that

$$\mathfrak{Pr}\left(X \geq \frac{2\alpha}{\lambda}\right) \leq \left(\frac{2}{e}\right)^{\alpha}$$

3. Find the characteristic function for a geometric random variable X having parameter p.

4. Let $\{X_1, X_2, \cdots, X_n\}$ be independent random variables each having a geometric distribution with parameter p. Find the characteristic function of the sum $X_1 + X_2 + \cdots + X_n$.

5. Let X be any random variable.

(a) Show that

$$\varphi_X(t) = E\left(\cos(tX)\right) + iE\left(\sin(tX)\right)$$

(b) Show that

$$\varphi_{-X}(t) = E\left(\cos(tX)\right) - iE\left(\sin(tX)\right)$$

(c) Show that

$$\varphi_{-X}(t) = \varphi_X(-t)$$

21. Central Limit Theorem

We earlier used Čebysev's inequality in order to deduce, for example, estimates like

$$\mathfrak{Pr}\left(|X-\mu|<\frac{\sigma}{2}\right)\geq 0.5$$

These estimates are useful precisely because they work even in the absence of knowledge about the underlying distribution of X.

The Central Limit Theorem is a much more powerful result of the above general type. Applications of the theorem often arise when a random sample of size n is drawn from a population and individual measurements are taken. An opinion poll is an example of such a situation where the sample mean $\bar{X}$ can be thought of estimating the mean of the underlying population μ. In this particular situation, if we know that X has a mean and a finite variance then we can conclude that $\bar{X}$ is approximately normally distributed for n sufficiently large. The astounding power of this result derives from the fact that it applies regardless of the distribution of X: all we need to know is that X has finite mean and variance.

21.1. Theorem.

Suppose that $\{X_1, X_2, \cdots, X_n\}$ are independent, identically distributed random variables having common characteristic function $\varphi_X(t)$. Set

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then the characteristic function of S_n is

$$\varphi_{S_n}(t) = \left(\varphi_X(t)\right)^n.$$

Proof. Since the random variables $\{X_n\}$ are independent, the characteristic function of the sum is the product of the characteristic functions:

$$\begin{aligned}\varphi_{S_n}(t) &= E\left(\exp\left(it(X_1+X_2+\cdots+X_n)\right)\right)\\ &= E\left(\exp(itX_1)\right)E\left(\exp(itX_2)\right)\cdots E\left(\exp(itX_n)\right)\\ &= \left(\varphi_X(t)\right)^n\end{aligned}$$

as desired.

∎

21.2. Corollary.

Suppose that $\{X_1, X_2, \cdots, X_n\}$ *are independent, identically distributed random variables having common characteristic function* $\varphi_X(t)$*. Set*

$$\bar{X}_n = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right).$$

Then the characteristic function of $\bar{X}_n$ *is*

$$\varphi_{\bar{X}_n}(t) = \left(\varphi_X\left(\frac{t}{n}\right)\right)^n.$$

Proof. This follows immediately from the theorem above and the definition:

$$\begin{aligned}
\varphi_{\bar{X}_n}(t) &= E\left(exp\left(i\frac{t}{n}(X_1 + X_2 + \cdots + X_n)\right)\right) \\
&= \varphi_{S_n}\left(\frac{t}{n}\right) \\
&= \left(\varphi_X\left(\frac{t}{n}\right)\right)^n.
\end{aligned}$$

∎

The next theorem will not be used directly in the proof of the Central Limit Theorem but is the fundamental fact needed to establish the Continuity Theorem, which we will use.

21.3. Theorem. Inversion Theorem.

Let X be a random variable having probability density function $f_X(x)$ and characteristic function $\varphi_X(t)$. If X is discrete having state space $\mathbb{Z}$ then

$$f_X(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-int}\varphi_X(t)\ dt \tag{1}$$

while if X is continuous then

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)\ dt. \tag{2}$$

provided that

$$\int_{-\infty}^{\infty} |\varphi_X(t)|\, dt < \infty.$$

This is called an *inversion theorem* since it describes how to recover the density function for X from the characteristic function or how to 'invert' the characteristic function. It is actually fairly easy to argue that (1) should be true. For example, if we expand $\varphi_X(t)$ inside the integral in (1) we get:

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-int}\varphi_X(t)\ dt &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-int}\left[\sum_{j=-\infty}^{\infty} e^{ijt} f_X(j)\right]\ dt \\ &= \sum_{j=-\infty}^{\infty} f_X(j)\ \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(j-n)t}\ dt\end{aligned}$$

The interchange of the sum and integral in last step above is justified by Fubini's theorem, which is proved in advanced analysis classes. To complete the argument, it is enough to show that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(j-n)t} = \begin{cases} 1 & \text{if } j = n \\ 0 & \text{otherwise}\end{cases}$$

which is left as an easy exercise for the reader.

The continuous case of the inversion theorem is much more difficult to verify and involves fairly deep results in advanced analysis. While difficult to verify in general, specific cases are easy – see the exercises.

The next theorem is the central fact needed in order to establish the Central Limit Theorem.

21.4. Theorem. Continuity Theorem.

Let X be a random variable having characteristic function $\varphi_X(t)$. Suppose that $\{X_1, X_2, \cdots, X_n. \cdots\}$ is a sequence of random variables having characteristic functions $\{\varphi_{X_n}(t)\}$. If

$$\lim_{n\to\infty} \varphi_{X_n}(t) = \varphi(t)$$

for all t then

$$\lim_{n\to\infty} P(X_n \le x) = P(X \le x).$$

for all points x at which $F_X(x)$ is continuous.

Thus if the characteristic functions converge then the distributions also converge. It is important to note that none of the random variables in question need be continuous! This is a very deep and complex result that relies heavily on the inversion theorem. It is called a 'continuity' theorem since it says, roughly, that the distribution function F_X depends continuously on the characteristic function φ_X since convergence of the latter implies convergence of the former.

21.5. Theorem. Central Limit Theorem.

Suppose that $\{X_1, X_2, \cdots, X_n\}$ are independent, identically distributed random variables having common finite mean μ and common finite variance σ^2. Set

$$\bar{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right).$$

Then

Thus for large values of n

$$\frac{\bar{X} - \mu}{(\sigma/\sqrt{n})}$$

is approximately normally distributed with mean zero and variance one. For practical purposes, n should be at least 30 in order to apply this result. Another way to state this is that the 'sampling distribution' of $\bar{X}$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. Notice that this is true *even if the common distribution shared by the random variables*

$\{X_n\}$ *is discrete!*

In order to establish the Central Limit Theorem we first establish a simple lemma about the behavior of characteristic functions near zero.

21.6. Lemma.

Let X be a random variable having characteristic function $\varphi_X(t)$. If X has finite mean μ then

$$\lim_{t\to 0}\frac{\ln(\varphi_X(t)) - i\mu t}{t} = 0. \tag{3}$$

If in addition X has finite variance σ^2 then

$$\lim_{t\to 0}\frac{\ln(\varphi_X(t)) - i\mu t}{t^2} = -\frac{\sigma^2}{2}. \tag{4}$$

Proof. Clearly $\varphi_X(0) = 1$ so $\ln(\varphi_X(t)$ is well-defined for t near zero and $\ln(\varphi_X(0)) = 0$. If X has finite expectation μ then $\varphi_X(t)$ is differentiable and $\varphi'_X(0) = i\mu$. Thus

$$\begin{aligned}\lim_{t\to 0}\frac{\ln(\varphi_X(t))}{t} &= \lim_{t\to 0}\frac{\ln(\varphi_X(t)) - \ln(\varphi_X(0))}{t-0}\\ &= \frac{d}{dt}\ln(\varphi_X(t))|_{t=0}\\ &= \frac{\varphi'_X(0)}{\varphi_X(0)}\\ &= i\mu\end{aligned}$$

or

$$\lim_{t\to 0}\frac{\ln(\varphi_X(t))}{t} = i\mu$$

From this equation (3) follows upon subtracting

$$i\mu = \lim_{t\to 0}\frac{i\mu t}{t}$$

from each side of the above equation.

Next suppose that X has finite variance σ^2. Applying l'Hôspital's rule

$$\begin{aligned}\lim_{t\to 0}\frac{\ln(\varphi_X(t)) - i\mu t}{t^2} &= \lim_{t\to 0}\frac{1}{2t}\left(\frac{\varphi_X'(t)}{\varphi_X(t)} - i\mu\right)\\ &= \lim_{t\to 0}\frac{\varphi_X'(t) - i\mu\varphi_X(t)}{2t\varphi_X(t)}\end{aligned}$$

Now apply l'Hôspital's rule a second time to get

$$\begin{aligned}\lim_{t\to 0}\frac{\ln(\varphi_X(t)) - i\mu t}{t^2} &= \lim_{t\to 0}\frac{\varphi_X''(t) - i\mu\varphi_X'(t)}{2\varphi_X(t) + 2t\varphi_X'(t)}\\ &= \frac{\varphi_X''(0) - (i\mu)^2}{2}\\ &= \frac{-(\sigma^2+\mu^2)+\mu^2}{2}\\ &= -\frac{\sigma^2}{2}.\end{aligned}$$

establishing equation (4).

■

Proof of the Central Limit Theorem. Let Y_n be the random variable

$$Y_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Let Z be a random variable having a normal distribution with mean zero and variance one. Then the characteristic function for Z is

$$\varphi_Z(t) = e^{-\frac{t^2}{2}}$$

and the distribution function for Z is

$$F_Z(z) = \int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\,dt.$$

Thus by the continuity theorem, it suffices to show that

$$\lim_{n\to\infty}\varphi_{Y_n}(t) = e^{-\frac{t^2}{2}}. \tag{5}$$

Let $\varphi_X(t)$ denote the common characteristic function shared by $\{X_n\}$. The characteristic function of Y_n is

$$\begin{aligned}
\varphi_{Y_n}(t) &= E\left(e^{itY_n}\right) \\
&= E\left(\exp\left(it\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\right)\right) \\
&= E\left(\exp\left(it\frac{(X_1+X_2+\cdots+X_n)-n\mu}{\sigma\sqrt{n}}\right)\right) \\
&= \exp\left(-it\frac{n\mu}{\sigma\sqrt{n}}\right)E\left((X_1+X_2+\cdots+X_n)\frac{it}{\sigma\sqrt{n}}\right) \\
&= \exp\left(-it\frac{n\mu}{\sigma\sqrt{n}}\right)\left(\varphi_X\left(\frac{it}{\sigma\sqrt{n}}\right)\right)^n \\
&= \exp\left[n\ln\left(\varphi_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right)-i\mu n\left(\frac{t}{\sigma\sqrt{n}}\right)\right]
\end{aligned}$$

Now

$$\begin{aligned}
n\ln\varphi_X\left(\frac{t}{\sigma\sqrt{n}}\right)-i\mu n\left(\frac{t}{\sigma\sqrt{n}}\right) &= n\left(\ln\left(\varphi_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right)-i\mu\frac{t}{\sigma\sqrt{n}}\right) \\
&= \frac{t^2}{\sigma^2}\left(\frac{\ln\left(\varphi_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right)-i\mu\frac{t}{\sigma\sqrt{n}}}{\frac{t^2}{\sigma^2 n}}\right) \\
&\longrightarrow -\frac{t^2}{\sigma^2}\frac{\sigma^2}{2} \text{ as } n\to\infty \\
&= -\frac{t^2}{2}
\end{aligned}$$

by (4) of the lemma. From this it follows that

$$\begin{aligned}
\lim_{n\to\infty}\varphi_{Y_n}(t) &= \lim_{n\to\infty}\exp\left(n\ln\varphi_X\left(\frac{t}{\sigma\sqrt{n}}\right)-i\mu n\left(\frac{t}{\sigma\sqrt{n}}\right)\right) \\
&= \exp\left(-\frac{t^2}{2}\right)
\end{aligned}$$

proving (5) and hence the result.

■

Under the weaker assumption that X has finite mean μ but omitting the assumption that X have finite variance, we can still conclude that $\bar{X}$ converges 'in probability' (or 'in measure') to μ. This result is known as the *Weak Law of Large Numbers*.

21.7. Theorem. The Weak Law of Large Numbers.

Suppose that $\{X_1, X_2, \cdots, X_n\}$ *are independent, identically distributed random variables having common finite mean* μ*. Set*

$$\bar{X}_n = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right).$$

Then for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathfrak{Pr}\left(|\bar{X}_n - \mu| > \epsilon\right) = 0.$$

Proof. Let $\varphi_X(t)$ be the common characteristic function for the random variables $\{X_n\}$ and let Y_n be the random variable

$$Y_n = \bar{X} - \mu$$

so that the characteristic function of Y_n is

$$\varphi_{Y_n}(t) = e^{-i\mu t}\left(\varphi_X\left(\frac{t}{n}\right)\right)^n.$$

For fixed $t \in \mathbb{R}$ we may choose n sufficiently large that $\ln(\varphi_x(t/n))$ is well-defined and

$$e^{-i\mu t}\left(\varphi_X\left(\frac{t}{n}\right)\right)^n = \exp\left(n\left[\ln\left(\varphi_X\left(\frac{t}{n}\right)\right) - i\mu\left(\frac{t}{n}\right)\right]\right)$$

Now observe that, if $t \neq 0$,

$$\begin{aligned}\lim_{n\to\infty} n\ln\left(\varphi_X\left(\frac{t}{n}\right)\right) - i\mu\left(\frac{t}{n}\right) &= t\lim_{n\to\infty}\frac{\ln\left(\varphi_X\left(\frac{t}{n}\right)\right) - i\mu\left(\frac{t}{n}\right)}{t/n}\\ &= 0\end{aligned}$$

by (3) in the Lemma. As a consequence, the characteristic function of

$$\bar{X} - \mu$$

approaches 1 as $n \to \infty$.

Now if X is the discrete random variable having density function

$$f_X(x) = \begin{cases} 1 & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

then X has characteristic function 1. The distribution function for X is

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ 1 & 0 < x \end{cases}$$

so $F_X(x)$ is continuous everywhere except at $x = 0$.

Fix $\epsilon > 0$. By the Continuity Theorem $\bar{X} - \mu$ approaches X as $n \to \infty$ and hence

$$\lim_{n \to \infty} \mathfrak{Pr}\left(\bar{X} - \mu \leq -\epsilon\right) = F_X(-\epsilon) = 0 \tag{7}$$

and

$$\lim_{n \to \infty} \mathfrak{Pr}\left(\bar{X} - \mu \leq \epsilon\right) = F_X(\epsilon) = 1.$$

The second inequality implies that

$$\lim_{n \to \infty} \mathfrak{Pr}\left(\bar{X} - \mu > \epsilon\right) = 0$$

which, when combined with (7) gives

$$\lim_{n \to \infty} \mathfrak{Pr}\left(|\bar{X}_n - \mu| >\right) = 0$$

as desired.

∎

21. Central Limit Theorem: Problems.

1. Let X be a continuous random variable having density function

$$f_X(t) = \frac{1}{2}e^{-|t|} \quad t \in \mathbb{R}.$$

(a) Show that

$$\varphi_X(t) = \frac{1}{1+t^2}$$

(b) Use the inversion formula to show that

$$e^{-|x|} = \int_{\mathbb{R}} e^{-ixt} \frac{1}{\pi(1+t^2)}\, dt$$

(b) Using (b) show that

$$e^{-|x|} = \int_{\mathbb{R}} e^{ixt} \frac{1}{\pi(1+t^2)}\, dt$$

2. Suppose that X has the Cauchy density

$$f_X(x) = \frac{1}{\pi(1+t^2)} \quad x \in \mathbb{R}$$

Show that

$$\varphi_X(t) = e^{-|t|} \quad t \in \mathbb{R}$$

3. Vessels in the US nuclear submarine fleet cruise for six months at a time without surfacing. As a consequence the vessels must have sufficient on board inventory of essential items to be reasonably confident that the inventory will last for the 180 day duration of the cruise. Suppose that an essential electrical component on the submarine has a lifespan, from initial power-on, that is exponentially distributed with a mean of 8 days. Once the component burns out, a new one is immediately installed and power is applied. How many such components should the supply officer have in inventory in order to have a 99% certainty that the inventory will last the duration of the cruise?

4. Show that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(j-n)t} = \begin{cases} 1 & \text{if } j = n \\ 0 & \text{otherwise} \end{cases}$$

completing the argument for equation (2).

polling

22. Applications to Polling

The Central Limit Theorem is critical to the modern practice of opinion polling. In this section we will present two examples out of many possibilities.

22.1. Example.

After a major event, the news networks usually do 'instant' polling, asking 600 randomly selected respondents a set of questions about the event. The networks then report the responses, usually including a 'margin of error' to the poll. This process can be understood – and evaluated – using the Central Limit Theorem.

Solution. Generally the questions can be answered 'yes' or 'no.' Thus any one question is a Bournoulli trial with p being the proportion of the entire population that would answer 'yes' given the opportunity. Of course, p is unknown, which is the point of the poll – to try to estimate p. For simplicity we will suppose that our poll has only one such question.

Since the question is asked of 600 randomly selected respondents, we can think of each answer as a Bernoulli random variable

$$X_i = \begin{cases} 1 & \text{if the answer is 'agree'} \\ 0 & \text{if the answer is 'disagree'} \end{cases}$$

The the proportion of the sample responding 'agree' is

$$\bar{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots, X_{600}\right).$$

Now if the network reports that the 'margin of error' for the poll is, for example, 5%, then they are reporting that

$$-0.05 \le \bar{X} - p \le +0.05$$

where p is the probability that a randomly selected respondent would respond 'agree.'

Since $\bar{X}$ is deduced from incomplete data (a sample as opposed to a census), there is necessarily uncertainty built into the estimate $\bar{X}$ – after all, $\bar{X}$ is a random variable! Thus it makes sense to ask how often $\bar{X}$ would meet the publicized standard of being accurate within 5%, i.e., to ask what is the value of

$$\mathfrak{Pr}\left(-0.05 \le \bar{X} - p \le +0.05\right). \qquad 1$$

The Central Limit Theorem gives a way to approximate this probability.

Since each X_i is a Bernoulli trial, the random variables $\{X_i\}$ are independent and identically distributed with common mean p and variance $p(1-p)$. Thus we could re-write (1) as

$$\begin{aligned}
&\mathfrak{Pr}\left(-0.05 \le \bar{X} - p \le +0.05\right) = \\
&\quad = \mathfrak{Pr}\left(-\frac{0.05}{\sqrt{p(1-p)}/\sqrt{600}} \le \frac{\bar{X}-p}{\sqrt{p(1-p)}/\sqrt{600}} \le +\frac{0.05}{\sqrt{p(1-p)}/\sqrt{600}}\right) \\
&\quad \approx \mathfrak{Pr}\left(-\frac{0.05\sqrt{600}}{\sqrt{p(1-p)}} \le Z \le +\frac{0.05\sqrt{600}}{\sqrt{p(1-p)}}\right)
\end{aligned}$$

where Z is a standard normal random variable.

Now $p(1-p)$ is largest when $p = 0.5$. Thus

$$\frac{0.05\sqrt{600}}{\sqrt{p(1-p)}} \ge \frac{0.05\sqrt{600}}{0.5}$$

and

$$-\frac{0.05\sqrt{600}}{\sqrt{p(1-p)}} \le -\frac{0.05\sqrt{600}}{0.5}$$

so

$$\left(-\frac{0.05\sqrt{600}}{\sqrt{p(1-p)}}, \frac{0.05\sqrt{600}}{\sqrt{p(1-p)}}\right) \supseteq \left(-\frac{0.05\sqrt{600}}{0.5}, \frac{0.05\sqrt{600}}{0.5}\right).$$

From this,

$$\begin{aligned}
\mathfrak{Pr}\left(-\frac{0.05\sqrt{600}}{\sqrt{p(1-p)}} \le Z \le +\frac{0.05\sqrt{600}}{\sqrt{p(1-p)}}\right) &\ge \mathfrak{Pr}\left(-\frac{0.05\sqrt{600}}{0.5} \le Z \le \frac{0.05\sqrt{600}}{0.5}\right) \\
&= \mathfrak{Pr}\left(-2.45 < Z < 2.45\right)] \\
&= 0.9602.
\end{aligned}$$

The last calculation above deduced from either standard normal tables or from built-in spreadsheet functions. For example, in Microsoft Excel the `NORMDIST` function returns the areas under the normal curve.

In particular we can conclude from this that the network's claim that their poll is accurate to within a 5% tolerance is true for at least 96.02% of all polls constructed with this methodology. Alternatively, we can conclude that 3.98% of the time a poll constructed with this methodology will have an error greater than 5%. Overall, the network's claim seems to be reasonably credible.

22.2. Example.

A similar but slightly different problem arises in determining the sample size necessary to achieve a desired level of accuracy. Typically prior to national elections the major polling organizations will all report their results with an 'error' of $\pm\mathbf{0.025}$. *In addition, the polls are generally designed to have this level of 'error' with 95% confidence.*

More precisely, if a yes/no polling question is asked of $\boldsymbol{n}$ *randomly selected respondents, then the random variables* $\{\boldsymbol{X_i}\}$ *given by*

$$\boldsymbol{X_i} = \begin{cases} \mathbf{1} & \textit{if the } \boldsymbol{i^{th}} \textit{ respondent answers 'yes'} \\ \mathbf{0} & \textit{otherwise} \end{cases}$$

are independent and each has a Bournoulli distribution with parameter $\boldsymbol{p}$ *where* $\boldsymbol{p}$ *is the probability that any randomly selected member of the population would answer 'yes.'*

Exactly as before, then, the polling organizations claim that

$$\boldsymbol{-0.025 \leq \bar{X} - p \leq +0.025.}$$

In order for this claim to be credible, we need to assess

$$\mathfrak{Pr}\,\boldsymbol{(-0.025 \leq \bar{X} - p \leq +0.025).}$$

If the above probability is at least 95% then we would say that the poll results have '95% confidence.' Thus our goal in this example is to choose $\boldsymbol{n}$ *in such a way that*

$$\mathfrak{Pr}\,\boldsymbol{(-0.025 \leq \bar{X} - p \leq +0.025) \geq 0.95.} \tag*{2}$$

Solution. With exactly the same reasoning as in the previous example, we can approximate (2) with a standard normal random variable

$$\begin{aligned} \mathfrak{Pr}\,(-0.025 \leq \bar{X} - p \leq +0.025) &\approx \mathfrak{Pr}\left(-\frac{0.025\sqrt{n}}{0.2} \leq Z \leq \frac{0.025\sqrt{n}}{0.5}\right) \\ &\geq 0.95 \end{aligned}$$

However, once again using either standard normal tables or a spreadsheet,

$$\mathfrak{Pr}\,\boldsymbol{(-1.96 \leq Z \leq +1.96) = 0.95}$$

confidence level

and hence can conclude that it suffices to choose n so that

$$\frac{0.025\sqrt{n}}{0.5} = 1.96$$

or $\sqrt{n} = 39.2$, from which $n = 1536.64$. Since sample sizes must be integers, we conclude that we should select $n = 1,537$.

∎

A close reading of the results from major polling organizations reveals that they all use exactly this sample size. Thus 95% of the time the poll results are accurate to within $\pm 2.5\%$.

If $\{X_1, X_2, \cdots, X_n\}$ are independent and identically distributed random variables having a common mean μ and a common variance σ^2, then the central limit theorem asserts that

$$\bar{X} = \frac{1}{n}\sum_{1}^{n} X_i$$

is approximately a normal random variable having mean μ and variance σ^2/n provided that n is sufficiently large. Thus for fixed $\gamma > 0$ one could calculate

$$\mathfrak{Pr}\left(\bar{X} - \frac{\gamma\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{\gamma\sigma}{\sqrt{n}}\right) \tag{22.1}$$

by approximating $\bar{X}$ with an appropriate normal distribution.

Usually the goal is to choose $\gamma > 0$ so that the above probability is high, say 95%. The value of the probability is said to be the “confidence level” of the interval

$$\left(\bar{X} - \frac{\gamma\sigma}{\sqrt{n}}, \bar{X} + \frac{\gamma\sigma}{\sqrt{n}}\right). \tag{22.2}$$

The table below gives values of γ from the normal distribution for various levels of confidence.

γ	Confidence Levels
$\pm$ 1.150	75%
$\pm$ 1.281	80%
$\pm$ 1.440	85%
$\pm$ 1.644	90%
$\pm$ 1.695	91%
$\pm$ 1.750	92%
$\pm$ 1.811	93%
$\pm$ 1.881	94%
$\pm$ 1.960	95%
$\pm$ 2.053	96%
$\pm$ 2.170	97%
$\pm$ 2.241	97.5%
$\pm$ 2.326	98%
$\pm$ 2.575	99%
$\pm$ 2.807	99.5%

22.3. Example.

Currently all lab samples from a physician's office are sent to $Tests~R~Us$, a commercial lab specializing in analyzing and producing pathology reports. The physician suspects that $Tests~R~Us$ may be cutting corners, and decides to double check their results against the state laboratory which has essentially a 100% accuracy rate. Of 512 samples, a $Tests~R~Us$ incorrectly identifies 32. Find a 98% confidence interval for the proportion of incorrectly identified samples.

Solution. To do this problem, the random variables in question are Bernoulli trials with an unknown value for p. However, we can estimate p using the data in the problem; in this kind of problem $\bar{X}$ is usually written as $\hat{p}$ since it is estimating the probability of "success." Since the variance is then $p(1-p)$ we can approximate the interval with

$$\left(\hat{p}-\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

with γ chosen so that

$$\mathfrak{Pr}\left(\hat{p}-\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}<p<\hat{p}-\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)=0.98.$$

The Central Limit Theorem implies for large values of n that $\hat{p}$ has a distribution which is approximately normal with $\mu = p$ and $\sigma^2 = p(1-p)$. As a consequence we approximate γ with the value of $\gamma = 2.326$ calculated from a normal distribution.

Using the data in the problem, $\hat{p} = 0.063$ and so this works out to $(0.0381, 0.0879)$. ∎

Note that 22.1 is a probability as long as we are dealing with random variables. In practice, one actually gathers data and calculates, based on the data, numerical values for $\bar{X}$. Substituting these values into 22.2 for $\bar{X}$ results in a specific interval. This interval either contains μ or it doesn't, so once the calculations are done the probability is either zero or one. Thus in the above example it is correct to say that the probability that p falls in the interval

$$\left(\hat{p} - \gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} - \gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

is approximately 98% since $\hat{p}$ represents a random variable. The same statement about $(0.0381, 0.0879)$ would be incorrect.

22. Applications to Polling: Problems.

1. Obstructive sleep apnoea effects approximately 4% of women and 9% of men between the ages of 30 and 60. Researchers were interested in estimating the nightly arousals of sleeping partners of subjects with obstructive sleep apnoea. in a sample of 126 sleeping partners of persons with sleep apnoea, the researcher found there were an average of 21.05 sleep interruptions in an eight hour period with a standard deviation of 4.08 interruptions. Find a 95% confidence interval for the average sleep interruptions of sleeping partners of persons with obstructive sleep apnoea.

2. A researcher was interested in learning whether people believe bad behavior might cause disease. To test this, the researcher showed 260 persons a short movie in which a narrator described a person who “lied, cheated and stole” and then subsequently caught an infectious disease. The narrator then stated ”I believe that serious illnesses happen at least slightly more often to people who deserve them.” The subjects were then asked if they agreed that illness can be “payback” for bad behavior. In the sample 50 agreed. Find a 95% confidence interval for the proportion of all persons who believe that disease can be ”payback” for bad behavior.

3. In 1985 the average US teen-ager drank 0.16 liters of soda per day. By 2005, this average had increased to 0.63 liters per day. Since soda consumption is associated with low intake of vitamins and minerals and high intake of sugar and fat, a school counselor starts program to encourage consumption of fruit juices and milk rather than soft drinks. In a sample of 80 students participating in the program, the soft drink consumption was 0.53 liters per day with a standard deviation of 0.06 liters. Find a 95% confidence interval for the average soda consumption of students participating in the peer-counseling program.

4. At birth infants show a preference for viewing pictures of human faces as opposed to other objects; however it is not until later that the infants distinguish between faces that are right-side-up and those that are upside-down. A sample of 53 randomly selected infants aged 3 months were shown a series of two side-by-side pictures of human faces: one that was right-side-up and another that was up-side-down. The researchers measured which photo the infants preferred by observing which they were more likely to look at. In this group, 23 showed a preference for the right-side-up faces. 95% confidence interval for the proportion of all 3 month-old infants who prefer right-side-up photos of human faces.

23. Laws of Large Numbers

Laws of Large Numbers
Chernoff, Herman

Laws of Large Numbers state that the observed frequencies of events tend to approach the actual probabilities as the number of observations increases. For example, if

$$\{X_1, X_2, \cdots, X_n\}$$

is a collection of independent and identically distributed random variables having common mean μ and if we set $\bar{X}_n$ to be the sample mean

$$\bar{X}_n = \frac{1}{n}(X_1, X_2, \cdots, X_n)$$

then the Weak Law of Large numbers tells us for any $\epsilon > 0$ that

$$\mathfrak{Pr}\left(\left|\bar{X}_n - \mu\right|\right) > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

Thinking of the random variables $\{X_1, X_2, \cdots, X_n\}$ as being observations from a population with unknown mean, the weak law says that we can "learn" the value of μ by calculating the sample mean. This result is certainly of fundamental theoretical importance. However, from a practical standpoint it provides no information about how much data is needed (how large n should be) in order to have our "learned" estimate for μ satisfy a predetermined bound.

Both of Markov's and Čebysev's inequalities are attempts to resolve this estimation problem. Using a technique now known as the Chernoff Method, Herman Chernoff proved the following improvement on Čebysev's inequality in 1952.

23.1. Theorem. Chernoff Bound.

Suppose that $\{X_1, \cdots, X_n\}$ are Bernoulli random variables having parameter p and set

$$S_n = X_1 + X_2 + \cdots + X_n.$$

then for any $\epsilon \geq 1$

$$\begin{aligned} \mathfrak{Pr}(S \geq \epsilon np) &\leq \exp((-\epsilon \ln(\epsilon) + \epsilon - 1)np) \\ &= \left(\frac{e^{\epsilon - 1}}{\epsilon^{\epsilon}}\right)^{np} \end{aligned}$$

Proof. As with Čebysev's Inequality, this follows from Markov's inequality, although the proof is somewhat less straightforward. Markov's inequality implies that

$$\mathfrak{Pr}\,(S \geq \epsilon np) = \mathfrak{Pr}\,\left(\epsilon^S \geq \epsilon^{\epsilon np}\right) \leq \frac{E(\epsilon^S)}{\epsilon^{\epsilon np}}.$$

Now consider for any $i = 1, \cdots n$

$$\begin{aligned}
E\left(\epsilon^{X_i}\right) &= \epsilon^0\,\mathfrak{Pr}\,(X_i = 0) + \epsilon^1\,\mathfrak{Pr}\,(X_i = 1) \\
&= (1 - \mathfrak{Pr}\,(X_i = 1)) + \epsilon\,\mathfrak{Pr}\,(X_i = 1) \\
&= 1 + (\epsilon - 1)\,\mathfrak{Pr}\,(X_i = 1) \\
&\qquad \text{(using } 1 + x \leq e^x \text{ for } x \geq 0\text{)} \\
&\leq \exp\left((\epsilon - 1)\,\mathfrak{Pr}\,(X_i = 1)\right) \\
&= e^{(\epsilon-1)p}
\end{aligned}$$

From this

$$\begin{aligned}
E(\epsilon^S) &= E\left(\epsilon^{X_1+\cdots+X_n}\right) \\
&= E\left(\epsilon^{X_1}\cdots\epsilon^{X_n}\right) \\
&= E\left(\epsilon^{X_1}\right)\cdots E\left(\epsilon^{X_n}\right) \\
&\leq e^{(\epsilon-1)E(X_1)}\cdots e^{(\epsilon-1)E(X_n)} \\
&= e^{(\epsilon-1)(E(X_1)+\cdots E(X_n))} \\
&= \exp\left((\epsilon - 1)np\right)
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathfrak{Pr}\,(S \geq \epsilon np) &\leq \frac{E(\epsilon^S)}{\epsilon^{\epsilon np}} \\
&\leq \frac{\exp\left((\epsilon-1)np\right)}{\exp\left(\epsilon np\ln(\epsilon)\right)} \\
&\leq \exp\left((-\epsilon\ln(\epsilon) + \epsilon - 1)np\right)
\end{aligned}$$

∎

The Chernoff Bound has applications to both networks (see the problems) and to – of course – games of chance. Consider the following two examples.

23.2. Example.

Estimate the chances of getting 75 or more "heads" in 100 rolls of a fair coin?

Solution. In this case $n = 100$ and, since the coin is presumed to be fair, $p = 0.5$. Since we want to compute

$$\mathfrak{Pr}\left(S \geq 75\right)$$

we will take $\epsilon = 1.5$. Then by Chernoff's inequality

$$\begin{aligned}
\mathfrak{Pr}\left(S \geq 75\right) &= \mathfrak{Pr}\left(S \geq \epsilon np\right)\\
&\leq \exp\left((-\epsilon\ln(\epsilon) + \epsilon - 1))np\right)\\
&= \exp\left((-1.5\ln(1.5) + .5)50\right)\\
&= 0.0044
\end{aligned}$$

∎

23.3. Example.

Some states offer a game called 'pick four' in which players purchase tickets selecting four numbers in order. Then after a fixed period of time the game ends and one winning combination is drawn from the possible 10 ,000 numbers (0000 to 9999). Each person who selected the winning numbers gets a fixed payout.

Of course if N tickets are sold, then the expected number of winners is

$$\frac{N}{10,000.}$$

Thus if 1,000,000 tickets are sold each week, the state could expect 100 winners, on average, each week. It is a simple matter to design the payout so that, in the long run, the state makes money on the game.

Regardless of what happens in the long run, however, the state must be prepared to pay off all of the winners even if, in any given week, more than the expected number win. By pure chance, for example, it is certainly possible that 1,000 or even 10,000 people could win the game. How large should the state's cash reserves to be to have a reasonable certainty of being able to pay off all the winners in any given week?

Solution. Chernoff's Bound helps to answer this question. Each ticket may be thought of as a random variable X_i that assumes the value 1 if it is a winner and the value 0 otherwise. Since the winning numbers are randomly selected, any given ticket has one chance in 10,000 of winning. Thus X_i can be thought of as Bernoulli random variable with parameter $p = 0.0001$. For the purposes of this example we will assume that the random variables $\{X_i\}$ are independent.

Under these conditions, then

$$X_1 + \cdots X_N$$

describes exactly the number of winning tickets. If N tickets are sold, then Chernoff's Bound computes

$$\mathfrak{Pr}\left(X_1 + \cdots + X_N \geq \epsilon N \cdot 0.0001\right)$$

If $\epsilon = 1.25$ then Chernoff's Bound tells us how likely it is that any given week's payout exceeds the expected payout by twenty-five per cent.

By Chernoff's Bound,

$$\begin{aligned}
&\mathfrak{Pr}\left(X_1 + \cdots + X_N \geq 1.25N \cdot 0.0001\right) \\
&\qquad \leq \exp\left((-1.25 \ln(1.25) + 0.25)100\right) \\
&= 0.055
\end{aligned}$$

In other words, there is only a 5.5% chance that there will be 125 or more winners in any given week. Thus the state only needs to keep a 25% cushion to be reasonably certain of always having enough cash on hand to pay all of the winners. A 36% cushion would be sufficient in all but 0.3% of the drawings.

∎

The assumption that the tickets are independent in fact is flawed in the above description. There are certain numbers that are more or less likely to be selected, and hence not all numbers are equally likely to be purchased on the tickets. Indeed, by studying the patterns of number selection and by playing games that divided winnings among the winning tickets, Chernoff was able to actually show a long-term profit of nearly 7%!

The above example is essentially one of determining the necessary 'carrying capacity' for the state to operate the lottery. Similar reasoning applies in any setting that can be modeled by Bernoulli trials interacting with a system of limited capacity. An obvious and important example is a telephony network. In this case, some of the N calls on the network will attempt to use a particular switch (say the one here at OU-Tulsa) and the remainder will not. This provides a value for the probability that any one of the N calls will attempt to pass through the OU-Tulsa switch. Chernoff's then gives a way to calculate the necessary capacity of the switch to be assured with high probability that all calls will be completed.

The above form of Chernoff's bound is quite general and widely used. There are numerous other forms, however, many of which can be deduced using the techniques in the next two results.

23.4. Theorem. Generalized Chernoff Bound.

Suppose that $\{X_1, \cdots X_n\}$ *are non-negative, independent, identically distributed random variables having a common moment generating function* $M_X(t)$*. Then for all* $t \in \mathbb{R}$

$$\mathfrak{Pr}\left(\frac{X_1 + \cdots X_n}{n} \geq \xi\right) \leq e^{n(\ln(M_X(t)) - t\xi)}$$

Proof. The proof is much the same as for the previous theorem. Then,

$$\begin{aligned}
\mathfrak{Pr}\left(\frac{X_1 + \cdots X_n}{n} \geq \xi\right) &= \mathfrak{Pr}\left(e^{t(X_1 + \cdots X_n)} \geq e^{nt\xi}\right) \\
&\text{apply Markov's inequality...} \\
&\leq \frac{E\left(e^{t(X_1 + \cdots X_n)}\right)}{e^{nt\xi}} \\
&= \frac{E\left(e^{tX_1} \cdots e^{tX_n}\right)}{e^{nt\xi}} \\
&= \frac{M_X^n(t)}{e^{nt\xi}} \\
&= e^{n(\ln(M_X(t)) - t\xi)}
\end{aligned}$$

∎

It is easy to re-write Chernoff's inequality in the following form:

$$\mathfrak{Pr}\left(\bar{X}_n - p \geq (\epsilon - 1)p\right) \leq \left(\frac{e^{\epsilon - 1}}{\epsilon^{\epsilon}}\right)^{np}.$$

This is sometimes called the *multiplicative Chernoff bound* since the right-hand-side of the inequality inside the probability function involves the product $(\epsilon - 1)p$. Sometimes it is more useful to consider inequalities of the form

$$\mathfrak{Pr}\left(\bar{X}_n - p \geq \epsilon\right).$$

Bounds on the above inequality are called *additive Chernoff bounds* or *Hoeffding's inequality* since no products are involved. In addition the expression

$$\left(\frac{e^{\epsilon-1}}{\epsilon^{\epsilon}}\right)^{np}$$

while quite sharp is difficult to deal with. Thus Hoeffding's bound deduces a somewhat weaker but more tractable bound for the probability. The version of the additive bound we will deduce states that

$$\mathfrak{Pr}\left(\left|\bar{X}_n - p\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

While the proof uses the basic technique pioneered by Chernoff, it is much more technical. We first prove three preliminary results.

23.5. Lemma.

Suppose that f is a real-valued function defined on the interval $[a, b]$ and that $f^{''}$ is continuous and non-negative on $[a, b]$. Then if $0 < t < 1$

$$(1-t)f(a) + tf(b) \leq f\left((1-t)a + tb\right).$$

In particular the graph of f lies beneath the chord joining $(a, f(a)$ and $(b, f(b))$. Such a function is said to be **convex.**

Proof. Fix t and set $x_0 = (1-t)a + tb$. Expanding f in it's Taylor's series about x_0 we obtain

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^{x} (x - s) f''(s)\, ds.$$

The above can also be readily verified by integrating by parts. In view of our assumptions, the last term must always be non-negative, and so we obtain

$$f(x) \leq f(x_0) + f'(x_0)(x - x_0).$$

Applying this formula in the case $x = a$ gives

$$f(a) \leq f(x_0) + f'(x_0)t(a - b)$$

and in the case $x = b$ gives

$$f(b) \leq f(x_0) + f'(x_0)(1-t)(b - a).$$

Multiply the first equation by $(1-t)$ and the second by t and add the equations:

$$\begin{aligned}(1-t)f(a)+tf(b) &\leq (1-t)f(x_0)+t(1-t)(a-b)f'(x_0)+\cdots\\ &\cdots+tf(x_0)+t(1-t)(b-a)f'(x_0)\\ &= f(x_0)\\ &= f\left((1-t)a+tb\right)\end{aligned}$$

proving the result.

■

23.6. Lemma.

Let $\xi(t)$ be defined by

$$\xi(t)=-pt+\ln\left(1-p+pe^t\right).$$

Then

$$\xi(t)\leq\frac{t^2}{8}$$

for all values of t.

Proof. To prove this we will expand ξ in a Taylor's series. To this end, note that

$$\begin{aligned}\xi'(t) &= -p+\frac{pe^t}{1-p+pe^t}\\ &= -p+\frac{p}{((1-p)e^{-t}+p)}\end{aligned}$$

and that

$$\xi''(t)=\frac{p(1-p)e^{-t}}{((1-p)e^{-t}+p)^2}$$

Expanding $\xi(t)$ in a Taylor's series about $t=0$ we obtain for some $t_0\in[0,t]$ that

$$\xi(t)=\xi(0)+\xi'(0)t+\xi''(t_0)\frac{t^2}{2}.$$

Since $\xi(0)=0=\xi'(0)$ it follows that

$$\xi(t)=\xi''(t_0)\frac{t^2}{2}$$

for some $t_0 \in [0, t]$. But now observe that

$$\begin{aligned}
\xi''(t) &= \frac{p(1-p)e^{-t}}{((1-p)e^{-t}+p)^2} \\
&= \frac{p}{((1-p)e^{-t}+p)} \frac{(1-p)e^{-t}}{((1-p)e^{-t}+p)} \\
&= \frac{p}{((1-p)e^{-t}+p)} \left(1 - \frac{p}{((1-p)e^{-t}+p)}\right)
\end{aligned}$$

which is in the form $u(1-u)$ with

$$u = \frac{p}{((1-p)e^{-t}+p)}.$$

Since $u(1-u)$ is has a maximum value of $1/4$ it follows that

$$\begin{aligned}
\xi(t) &= \xi''(t_0)\frac{t^2}{2} \\
&\leq \frac{1}{4}\frac{t^2}{2}
\end{aligned}$$

showing $\xi(t) \leq t^2/8$ as desired.

∎

23.7. Lemma.

Suppose that $E(X) = 0$ and $a < X < b$. Then for any $s > 0$

$$E\left(e^{sX}\right) \leq e^{s^2(b-a)^2/8}.$$

Proof. For $0 \leq \lambda \leq 1$ and fixed x with $a < x < b$ define

$$\lambda = \frac{b-x}{b-a}.$$

Then for any $s > 0$ it follows that

$$sx = \lambda sa + (1-\lambda)sb.$$

Since e^u is a convex function in u, this implies that

$$e^{sx} \le \frac{b-x}{b-a} e^{sa} + \frac{x-a}{b-a} e^{sb}.$$

In particular then since $E(X) = 0$

$$E\left(e^{sX}\right) \le \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb}.$$

Taking

$$p = -\frac{a}{b-a}$$

and using the fact that

$$-ps(b-a) = as$$

we can rewrite this as

$$\begin{aligned} E\left(e^{sX}\right) &\le (1-p)e^{sa} + pe^{sb} \\ &= \left(1 - p + pe^{s(b-a)}\right) e^{-ps(b-a)} \\ &= \exp\left(-ps(b-a) + \ln\left(1 - p + pe^{s(b-a)}\right)\right) \end{aligned} \qquad \textbf{(23.1.)}$$

Now set $\xi(t)$ to be the value of the exponential function with $t = s(b-a)$:

$$\xi(t) = -pt + \ln\left(1 - p + pe^t\right).$$

By the Lemma

$$\xi(t) \le \frac{t^2}{8}$$

for all values of t.

To complete the proof, set $t = s(b-a)$ and substituting into (23.1):

$$\begin{aligned} E\left(e^{sX}\right) &\le \exp\left(-ps(b-a) + \ln\left(1 - p + pe^{s(b-a)}\right)\right) \\ &= \exp\left(\xi(s(b-a))\right) \\ &\le \exp\left(\frac{s^2(b-a)^2}{8}\right) \end{aligned}$$

∎

In view of the above lemma, we can now state and prove Hoeffding's inequality.

23.8. Theorem. Hoeffding's Inequality (I).

Let $\{X_1, X_2, \cdots, X_n\}$ be independent and identically distributed random variables having common expectation μ. Suppose for some number $R > 0$ that $0 \leq X_i \leq R$. If

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

then for every $\epsilon > 0$

$$\mathfrak{Pr}\left(\bar{X}_n - \mu \geq \epsilon\right) \leq \exp\left(\frac{-2n\epsilon^2}{R^2}\right)$$

and

$$\mathfrak{Pr}\left(\mu - \bar{X}_n \geq \epsilon\right) \leq \exp\left(\frac{-2n\epsilon^2}{R^2}\right)$$

Combining the two inequalities gives

$$\mathfrak{Pr}\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-2n\epsilon^2}{R^2}\right)$$

In the special case that each X_i is a Bernoulli trial (and hence that $R = 1$) we obtain the following special case.

23.9. Theorem. Hoeffding's Inequality (II).

Let $\{X_1, X_2, \cdots, X_n\}$ *be independent Bernoulli trials having common probability of success* p*. If*

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

then for every $\epsilon > 0$

$$\mathfrak{Pr}\left(\bar{X}_n - p \geq \epsilon\right) \leq e^{-2n\epsilon^2}$$

and

$$\mathfrak{Pr}\left(p - \bar{X}_n \geq \epsilon\right) \leq e^{-2n\epsilon^2}.$$

Combining the two inequalities gives

$$\mathfrak{Pr}\left(\left|\bar{X}_n - p\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Proof. We prove only the first inequality in the first version of the inequality. The second inequality is deduced in exactly the same way, and the last is the sum of the first two.

For each i take $U_i = X_i - \mu$. Then

- $E(U_i) = 0$ for each i;
- $-\mu \leq U_i \leq R - \mu$ for each i.

Then via the lemma for each i and any $s > 0$

$$E\left(e^{sU_i}\right) \leq e^{s^2 R^2/8}.$$

Legendre transform

Now for any $\epsilon > 0$ and any $s > 0$ we can apply Markov's inequality

$$\begin{aligned}
\mathfrak{Pr}\left(\bar{X}_n - p > \epsilon\right) &= \mathfrak{Pr}\left(\sum_{i=1}^{n} U_i > n\epsilon\right) \\
&= \mathfrak{Pr}\left(\exp\left(s\sum_{i=1}^{n} U_i\right) > e^{n\epsilon s}\right) \\
&\leq \frac{E\left(\exp\left(s\sum_{i=1}^{n} U_i\right)\right)}{e^{ns\epsilon}} \\
&= \frac{\prod_{i=1}^{n} E\left(\exp\left(sU_i\right)\right)}{e^{ns\epsilon}} \\
&\leq \frac{e^{ns^2R^2/8}}{e^{ns\epsilon}} \\
&= \exp\left(\frac{ns^2R^2}{8} - ns\epsilon\right)
\end{aligned}$$

Now

$$\frac{ns^2R^2}{8} - ns\epsilon$$

assumes a minimum value of $\frac{-2n\epsilon^2}{R^2}$ when $s = 4\epsilon/R^2$ which completes the proof.

■

We remark that in proving Hoeffding's inequality the essential technique involves examining

$$I(\xi) = \min_{t\in\mathbb{R}^+} \{(\ln(M_X(t)) - t\xi)\}$$

in the generalized Chernoff bound. The function $I(\xi)$ above is the Legendre transform of $\ln\left(M_X(t)\right)$.

23. Laws of Large Numbers: Problems.

1. Suppose that a dishonest gambler wishes to produce a die on which the number "6" never appears when the die is rolled. However the gambler is uncertain how to do this, so he weights the die and then rolls it 25 times to see if a "6" ever appears. If it never appears in these 25 tries, then he assumes that he has successfully rigged the die, otherwise he repeats the process. Comment on the gambler's strategy for "learning"

$$\mathfrak{Pr}\,(\text{rolling a six}).$$

In particular what are the chances that the gambler stops after 25 trials even if the die is still fair?

2. Suppose that a coin is flipped 100 times and we observe 75 heads and 25 tails. Estimate the chances of getting 75 or more heads in 100 rolls of a fair die. Is this evidence that the coin is biased in favor of heads?

3. Repeat the pick four example but with the following assumptions:
(a) Suppose that the number of players is $N = 500,000$.
(b) Suppose that the number of players is $N = 250,000$.
(c) Suppose that the number of players is $N = 100,000$.

Explain what, if any, difference you see in the above answers.

4. Let $\{X_1, X_2, \cdots, X_n\}$ be independent Bernoulli trials having common probability of success p. Let $S = \sum_{i=1}^{n} X_i$ and let δ be a real number with $0 < \delta < 1$.
(a) For arbitrary $\alpha > 0$ use the Chernoff technique (applied to $e^{\alpha S}$) to show that

$$\mathfrak{Pr}\,(S \geq (1+\delta)np) \leq \left(\frac{1 + p(e^{\alpha} - 1)}{e^{(1+\delta)\alpha p}}\right)^{n}.$$

(b) Use the fact that $1 + x \leq e^{x}$ for $x \geq 0$ to conclude that

$$\mathfrak{Pr}\,(S \geq (1+\delta)np) \leq \exp\left(np(e^{\alpha} - 1 - (1+\delta)\alpha)\right).$$

(c) Show that the function $\xi(\alpha) = e^{\alpha} - 1 - (1+\delta)\alpha)$ has minimum value $\delta - (1-\delta)\ln(1-\delta)$ and hence that

$$\mathfrak{Pr}\,(S \geq (1+\delta)np) \leq \exp\left(\delta - (1-\delta)\ln(1-\delta)\right).$$

(d) Expand $\delta - (1-\delta)\ln(1-\delta)$ in a Taylor's series to conclude that

$$\delta - (1-\delta)\ln(1-\delta) \leq -\frac{\delta^2}{3}$$

and hence that

$$\mathfrak{Pr}\left(S \geq (1+\delta)np\right) \leq \exp\left(-\frac{\delta^2}{3}\right).$$

While this provides a sharper estimate than the corresponding one-sided inequality in Hoeffding's inequality, it lacks the symmetry needed to deduce the version with absolute values (which is needed for confidence intervals).

5. Let $\{X_n\}_{n=1}^{1000}$ be a collection of independent Bernoulli random variables having parameter 0.84.

(a) Use Markov's inequality to estimate

$$\mathfrak{Pr}\left(X_1 + X_2 + \cdots X_{1000} \geq 900\right)$$

(b) Use Čebysev's inequality to estimate

$$\mathfrak{Pr}\left(X_1 + X_2 + \cdots X_{1000} \geq 900\right)$$

(c) Use the Chernoff Bound to estimate

$$\mathfrak{Pr}\left(X_1 + X_2 + \cdots X_{1000} \geq 900\right)$$

(d) Which gives the better estimate?

(For example, $X_1 + X_2 + \cdots X_{1000}$ might count the number of college applicants with ACT scores higher than 25. This would assist a University in estimating their scholarship budget.)

6.

(a) Suppose that a telephony network handles one billion calls on a typical day. We are designing a switch on the network that will handle, on average, one million of those calls. If we design the switch with 1% over-capacity, i.e., with the ability to handle 1,001,000 calls, then estimate the chances that we will ever exceed the capacity of the switch.

(b) What assumptions did you make about the distribution and independence of the calls? Are these reasonable assumptions? Do you think these assumptions would be more or less reasonable for a data network as opposed to a telephony network?

classification rule

24. Application to Learning Theory

Many prediction problems involve using a set of observations X in order to predict an outcome Y. For example, one might use atmospheric conditions such as barometric pressure, humidity, temperature, wind direction and speed (compiled into a composite measure X) to predict whether or not it is going to rain (so Y is "rain" or "no rain."Other examples might involve using characteristics of an email message to decide whether or not it is spam, or patient symptoms to decide whether or not a disease is present, or network characteristics to decide whether or not a signal is successfully transmitted.

In all of these cases, X and Y are subject to random fluctuations and so can be thought of as random variables. The prediction involves observing the outcome X and then applying a decision rule – based on the outcome – to predict the value of Y. In all the examples we have given Y is a binary classification and so we may assume that Y assumes the values 0 or 1. A decision rule can then be thought of as a function γ

$$\gamma : X \mapsto \{0, 1\}.$$

The function γ is sometimes called a **classification rule** since it classifies each X by using X to predict one of the two binary outcomes. Thus the assignment of X to $\gamma(X)$ places X in the "0" class or in the "1" class.

Of course there are many different ways in which one might combine the characteristics making up X in order to predict the outcome Y. Further the relationship between the predictor (X) and the outcome (Y) cannot generally expected to be perfect, so that there will always be some observations X for which $\gamma(X) \neq Y$, i.e., $\gamma(X)$ will occasionally err. Thus any prediction scheme $\gamma(X)$ necessarily must assess the risk associated with random fluctuations in both the predictor X and the outcome Y as well as the inherent inaccuracies of the predictor.

Suppose that there are two possible decision rules, γ_1 and γ_2. The problem we will discuss in this section deals with deciding which of the two rules is preferable, or less "risky." Thus a more rigorous understanding of "risk" is needed.

Generally the framework for assessing the risk associated with a rule γ involves a collection of test data. The test data can be thought of as a set of observations $\{(X_1, Y_1), (X_1, Y_2), \cdots (X_n, Y_n)\}$ – for example composite atmospheric conditions (X_i) for day i and whether or not it rained (Y_i) on day i. Each test observation X_i results in a prediction $\gamma(X_i)$. If $\gamma(X_i) = Y_i$ then the rule γ accurately predicted the outcome, while if $\gamma(X_i) \neq Y_i$ the rule failed. The **empirical loss** is the number of times that rule γ is not accurate in these n trials, while the **empirical risk** is the average loss. Given two competing prediction rules γ_1 and γ_2 we would then prefer the less risky, i.e., the one with smaller empirical risk.

weak law of large numbers

While this framework makes intuitive sense for choosing between γ_1 and γ_2, before verifying that it is also reasonable mathematically it is necessary to provide a somewhat more structured context.

24.1. Definition.

The **loss function** *is the mapping* $\ell : \mathbb{R} \times \mathbb{R} \longrightarrow \{0, 1\}$ *given by*

$$\ell(x, y) = \begin{cases} 0 & \textit{if } x = y \\ 1 & \textit{if } x \neq y \end{cases}$$

In the sequel we will suppose that we have two probability spaces $(\Omega_1, \mathcal{E}_1, \mathfrak{Pr}_1)$ and $(\Omega_2, \mathcal{E}_2, \mathfrak{Pr}_2)$ and two random variables $X : \Omega_1 \longrightarrow \mathbb{R}$ and $Y : \Omega_2 \longrightarrow \{0, 1\}$ defined on Ω_1 and Ω_2 respectively. The random variables (X, Y) then have a joint distribution. Our test data $\{(X_1, Y_1), (X_1, Y_2), \cdots (X_n, Y_n)\}$ will be assumed to be an independent and identically distributed collection from this joint distribution $\varphi_{XY}(x, y)$. In most practical applications very little is known about the joint distribution. Indeed, the learning problem can be conceived at least in part as the problem of deducing properties of the joint distribution from the test data.

24.2. Definition.

Let $\gamma : \mathbb{R} \longrightarrow \{0, 1\}$. *The* **loss** *associated with* γ *and* (X, Y) *is* $\ell(\gamma(X), Y)$ *and the* **risk** *associated with* γ *is*

$$R(\gamma) = E(\ell(\gamma(X), Y))$$

Thus given two rules γ_1 and γ_2 we would prefer the one with smaller risk. Since the underlying distribution is unknown, this criteria does not lead to a directly computable conclusion. However, if we gather sufficient test data the decision may be based on empirical evidence. Since test data is necessarily incomplete – not being census data – the empirical evidence necessarily includes random error. The weak law of large numbers[and learning theory] tells us that this error vanishes as $n \longrightarrow \infty$ and Hoeffding's inequality quantifies the rate at which the error vanishes.

agnostic learning

24.3. Definition.

Suppose that

$$\{(X_1, Y_1), (X_1, Y_2), \cdots (X_n, Y_n)\}$$

are independent and identically distributed random variables as described above. We will call $\{(X_1, Y_1), (X_1, Y_2), \cdots (X_n, Y_n)\}$ *our* **test set**. *The* **empirical loss** *associated with a test set is the random variable*

$$\sum_{i=1}^{n} \ell(\gamma(X_i), Y_i)$$

and the **empirical risk** *is the random variable*

$$\hat{R}_n(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \ell(\gamma(X_i), Y_i)$$

Assuming that $R(\gamma) < \infty$, then it follows (see Section (19.1), p. 166) that

$$E\left(R_n(\gamma)\right) = R(\gamma)$$

i.e., that $\hat{R}_n(\gamma)$ is an unbiased estimator for $R(\gamma)$.

Further the weak law of large numbers tells us for any $\epsilon > 0$ that

$$\lim_{n \to \infty} \mathfrak{Pr}\left(\left|R(\gamma) - \hat{R}_n(\gamma)\right| > \epsilon\right) = 0.$$

Thus it is possible to learn the risk associated with γ from the empirical risk. The basic question is "how fast does $\hat{R}_n(\gamma)$ converge to $R(\gamma)$?" This is where Hoeffding's inequality comes into play. This approach is sometimes called **agnostic learning** since it relies on no prior assumptions about the underlying distribution of (X, Y) (other than that the risk is finite).

24.4. Theorem.

Suppose that $\{(X_1, Y_1), (X_1, Y_2), \cdots, (X_n, Y_n)\}$ *are independent and identically distributed random variables as described above and suppose that* $R(\gamma) < \infty$*. Let* $\epsilon > 0$ *be arbitrary. Then*

$$\mathfrak{Pr}\left(\hat{R}(\gamma) - R(\gamma) \geq \epsilon\right) \leq e^{-2n\epsilon^2}$$

and

$$\mathfrak{Pr}\left(R(\gamma) - \hat{R}(\gamma) \leq \epsilon\right) \leq e^{-2n\epsilon^2}.$$

Thus

$$\mathfrak{Pr}\left(\mid \hat{R}(\gamma) - R(\gamma) \mid \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Proof. We consider the random variables

$$\{\ell(\gamma(X_1), Y_1), \ell(\gamma(X_1), Y_1), \cdots, \ell(\gamma(X_n), Y_n)\}$$

so

$$\begin{aligned} E(\hat{R}(\gamma)) &= E\left(\frac{1}{n}\sum_{i=1}^{n} \ell(\gamma(X_i), Y_i)\right) \\ &= E(\ell(\gamma(X), Y)) \\ &= R(\gamma). \end{aligned}$$

Further for each i

$$0 \leq \ell(\gamma(X_i), Y_i) \leq 1.$$

Thus the assumptions of Hoeffding's inequality (23.8) apply with $R = 1$ and so (24.4) is just a restatement of (23.8).

∎

The value of (24.4)is that it provides a way of estimating the risk for any decision rule γ up to any desired degree of accuracy ϵ. The estimate depends only on the size of the test sample and not on any other *a priori* assumptions or knowledge about the underlying distribution. Thus (24.4)gives a practical way to evaluate the risk of any classification scheme γ and to differentiate between different classification schemes γ_1 and γ_2.

Given that one can empirically estimate risks, a related question becomes whether or not there is a classification scheme that minimizes risk. It turns out that there is such a scheme, although the proof is not constructive since it relies on knowledge of the underlying distribution.

24.5. Definition.

The **Bayes' risk** *is the infimum of the risk for all classifiers under consideration:*

$$R^* = \inf_{\gamma} R(\gamma).$$

24.6. Definition.

The **Bayes' classifier** *is the following mapping:*

$$\gamma^*(x) = \begin{cases} 1 & \textit{if } \eta(x) \geq 0.5 \\ 0 & \textit{otherwise} \end{cases}$$

where

$$\eta(x) = \mathfrak{Pr}\left(Y = 1 \middle| X = x\right).$$

Using these concepts we can prove that the Bayes' classifier is the optimal classifier in the sense that it attains the value of Bayes' risk.

24.7. Theorem.

The Bayes' classifier $\gamma*$ attains the Bayes' risk, i.e.,

$$R(\gamma*) = R^* = \inf_{\gamma} R(\gamma).$$

Proof. Let g be any classifier. We will first show that

$$\mathfrak{Pr}\left(g(x) \neq Y \middle| X = x\right) \geq \mathfrak{Pr}\left(\gamma^*(x) \neq Y \middle| X = x\right).$$

For any g and any x

$$\begin{aligned}
&\mathfrak{Pr}\left(g(x) \neq Y \middle| X = x\right) = \\
&\qquad = 1 - \mathfrak{Pr}\left(g(x) = Y \middle| X = x\right) \\
&\qquad = 1 - \Big(\mathfrak{Pr}\left(Y = 1, g(x) = 1 \middle| X = x\right) + \mathfrak{Pr}\left(Y = 0, g(x) = 0 \middle| X = x\right)\Big) \\
&\qquad = 1 - \Big(1_{\{g(x)=1\}}\,\mathfrak{Pr}\left(Y = 1 \middle| X = x\right) + 1_{\{g(x)=0\}}\,\mathfrak{Pr}\left(Y = 0 \middle| X = x\right)\Big) \\
&\qquad = 1 - \left(1_{\{g(x)=1\}}\eta(x) + 1_{\{g(x)=0\}}(1 - \eta(x))\right)
\end{aligned}$$

In particular,

$$\mathfrak{Pr}\left(\gamma^*(x) \neq Y \middle| X = x\right) = 1 - \left(1_{\{\gamma^*(x)=1\}}\eta(x) + 1_{\{\gamma^*(x)=0\}}(1 - \eta(x))\right).$$

Upon taking the difference it follows that

$$\begin{aligned}
&\mathfrak{Pr}\ \left(g(x) \neq Y \middle| X = x\right) - \mathfrak{Pr}\ \left(\gamma^*(x) \neq Y \middle| X = x)\right)) \\
&\quad = \eta(x)\left[1_{\{\gamma*(x)=1\}} - 1_{\{g(x)=1\}}\right] + (1 - \eta(x))\left[1_{\{\gamma*(x)=0\}} - 1_{\{g(x)=0\}}\right] \\
&\quad = \eta(x)\left[1_{\{\gamma*(x)=1\}} - 1_{\{g(x)=1\}}\right] + (1 - \eta(x))\left[1 - 1_{\{\gamma*(x)=1\}} - 1 + 1_{\{g(x)=1\}}\right] \\
&\quad = \eta(x)\left[1_{\{\gamma*(x)=1\}} - 1_{\{g(x)=1\}}\right] - (1 - \eta(x))\left[1_{\{\gamma*(x)=1\}} - 1_{\{g(x)=1\}}\right] \\
&\quad = (2\eta(x) - 1)\left(1_{\{\gamma*(x)=1\}} - 1_{\{g(x)=1\}}\right)
\end{aligned}$$

For future reference we summarize the above as

$$\begin{aligned}
&\mathfrak{Pr}\ \left(g(x) \neq Y \middle| X = x\right) - \mathfrak{Pr}\ \left(\gamma^*(x) \neq Y \middle| X = x)\right)) \\
&\qquad\qquad = (2\eta(x) - 1)\left(1_{\{\gamma*(x)=1\}} - 1_{\{g(x)=1\}}\right) \qquad\qquad \mathbf{(24.1.)}
\end{aligned}$$

Recalling that

$$\gamma^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

we can distinguish two cases.

Case I. $\eta(x) \geq 0.5$

In this case

$$\underbrace{(2\eta(x) - 1)}_{\geq 0}\underbrace{\left(\underbrace{1_{\{\gamma*(x)=1\}}}_{=1} - \underbrace{1_{\{g(x)=1\}}}_{=0 \text{ or } 1}\right)}_{\geq 0}$$

Case II. $\eta(x) < 0.5$

$$\underbrace{(2\eta(x)-1)}_{<0}\underbrace{\left(\underbrace{\mathbf{1}_{\{\gamma*(x)=1\}}}_{=0}-\underbrace{\mathbf{1}_{\{g(x)=1\}}}_{=0 \text{ or } 1}\right)}_{\leq 0}$$

Thus in either case

$$\mathfrak{Pr}\left(g(x) \neq Y \middle| X = x\right) \geq \mathfrak{Pr}\left(\gamma^*(x) \neq Y \middle| X = x)\right))$$

To see that this suffices to prove the result, let φ_{XY} be the joint density of X and Y. Then by definition

$$\varphi_{XY}(x,y) = \varphi_X(x)\varphi_{Y|X}(y|x).$$

Then for any classifier g

$$\begin{aligned}
\ell(g(x),y)\varphi_{XY}(x,y) &= \ell(g(x),y)\varphi_X(x)\varphi_{Y|X}(y|x) \\
&= \ell(g(x),y)\varphi_X(x)\,\mathfrak{Pr}\left(Y = y \middle| X = x\right) \\
&= \varphi_X(x)\,\mathfrak{Pr}\left(Y \neq g(x) \middle| X = x\right).
\end{aligned}$$

In the above we use the observation that the random variable $Y|_{X=x}$ is discrete and hence that the density function $\varphi_{Y|X}(y|x)$ evaluated at y is just the probability that $Y|_{X=x}$ assumes the value y.

Now if, for example, X is continuous the conclusion follows from the fact that for all classifiers g

$$\begin{aligned}
R(g) &= E(\ell(g(X),Y)) \\
&= \sum_y \int_{\mathbb{R}} \ell(g(x),y)\varphi_{XY}(x,y)\,dx \\
&= \sum_y \int_{\mathbb{R}} \varphi_X(x)\,\mathfrak{Pr}\left(Y \neq g(x) \middle| X = x\right)dx \\
&\geq \sum_y \int_{\mathbb{R}} \varphi_X(x)\,\mathfrak{Pr}\left(Y \neq \gamma^*(x) \middle| X = x\right)dx \\
&= \sum_y \int_{\mathbb{R}} \ell(\gamma^*(x),y)\varphi_{XY}(x,y)\,dx \\
&= E(\ell(\gamma^*(X),Y)) \\
&= R(\gamma^*)
\end{aligned}$$

The proof for the case that X is discrete is similar.

∎

It is not possible generally to calculate the Bayes' classifier γ^* since φ_{XY} is not known. However the definition of γ^*

$$\gamma^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

where

$$\eta(x) = \mathfrak{Pr}\left(Y = 1 \middle| X = x\right)$$

does suggest that one can approximate γ^* by approximating η. That is the content of the next theorem.

24.8. Theorem.

Let $\tilde{\eta} : \mathbb{R} \longrightarrow [0, 1]$ be an approximation of η. If

$$g(x) = \begin{cases} 1 & \textit{if } \tilde{\eta}(x) \geq 0.5 \\ 0 & \textit{otherwise} \end{cases}$$

then

$$R(g) - R^* \leq E\left(|\tilde{\eta}(X) - \eta(X)|\right)$$

Proof. We apply (24.1):

$$\begin{aligned} \mathfrak{Pr}\left(g(X) \neq Y \middle| X = x\right) &- \mathfrak{Pr}\left(\gamma^*(X) \neq Y \middle| X = X\right) \\ &= (2\eta(x) - 1)\left(1_{\{\gamma*(x)=1\}} - 1_{\{g(x)=1\}}\right) \\ &= |2\eta(x) - 1| 1_{\{\gamma^*(x) \neq g(x)\}} \\ &= \frac{1}{2}\left|\eta(x) - \frac{1}{2}\right| 1_{\{\gamma^*(x) \neq g(x)\}} \end{aligned}$$

Notice next for any classifier g we can apply the definition of expectation to the discrete random variable $\ell(g(X), Y)$ to obtain

$$R(g) = E(\ell(g(X), Y)) = \mathfrak{Pr}\left(g(X) \neq Y\right)$$

Then using the definition of conditional densities as in the previous theorem

$$\begin{aligned} R(g) - R^* &= \mathfrak{Pr}\,(g(X) \ne Y) - \mathfrak{Pr}\,(\gamma^*(X) \ne Y) \\ &= \int_{\mathbb{R}} \left(\mathfrak{Pr}\,\left(g(x) \ne Y \middle| X = x\right) - \mathfrak{Pr}\,\left(\gamma^*(x) \ne Y \middle| X = X\right)\right) \varphi_X(x)\,dx \\ &= \int_{\mathbb{R}} \frac{1}{2}\left|\eta(x) - \frac{1}{2}\right| 1_{\{\gamma^*(x) \ne g(x)\}} \varphi_X(x)\,dx \end{aligned}$$

Now if $\gamma^*(x) \ne g(x)$ then either

$$\left(g(x) = 1 \quad \text{and} \quad \gamma^*(x) = 0\right) \implies \left(\tilde{\eta}(x) \ge 0.5 \quad \text{and} \quad \eta(x) < 0.5\right)$$

or else

$$\left(g(x) = 0 \quad \text{and} \quad \gamma^*(x) = 1\right) \implies \left(\tilde{\eta}(x) < 0.5 \quad \text{and} \quad \eta(x) \ge 0.5\right)$$

In either case

$$\left|\eta(x) - \frac{1}{2}\right| 1_{\{\gamma^*(x) \ne g(x)\}} \le |\eta(x) - \tilde{\eta}(x)| 1_{\{\gamma^*(x) \ne g(x)\}}.$$

On the other hand, if $\gamma^*(x) = g(x)$ then

$$\mathfrak{Pr}\,\left(g(x) \ne Y \middle| X = x\right) - \mathfrak{Pr}\,\left(\gamma^*(x) \ne Y \middle| X = X\right) = 0.$$

Thus

$$\begin{aligned} R(g) - R^* &\le \int_{\mathbb{R}} 2|\eta(x) - \tilde{\eta}(x)| 1_{\{\gamma^*(x) \ne g(X)\}} \varphi_X(x)\,dx \\ &\le \int_{\mathbb{R}} 2|\eta(x) - \tilde{\eta}(x)| \varphi_X(x)\,dx \\ &= 2E(|\tilde{\eta}(x) - \eta(X)|) \end{aligned}$$

∎

training set
Einstein

Learning theory has evolved into a rich literature and this section only touches on the most basic concepts. Our discussion has assumed that we already have candidate classifiers γ to consider. Deducing a candidate classifier from data is much more difficult, especially where real-time computational constraints are imposed. Typically a **training set** is first used to deduce a candidate classifier using some algorithm, then a test set is used to assess the risk of that classifier. These topics belong to a more advanced course specializing in machine learning.

We comment further that we have restricted our approach to binary outcomes Y. There are other possible learning problems. A classical example is estimate the signal f under noisy conditions:

$$Y = f(X) + W$$

where X is continuous and W is random noise independent of X. The loss function in this case is often

$$\ell(X, Y) = \|X - Y\|^2$$

and the signal estimation becomes a least-squares regression problem. This case is further complicated since the signal is often governed by an evolution equation such as

$$\int_0^t X(s)\,ds + X(t) = W(t)$$

The forcing function $W(t)$ might represent, for example, “white noise”, introduced by Einstein in his studies of Brownian motion. Again, these are topics for more advanced courses.

25. Two State Markov Chains

stochastic process
phase space
state space
time series
Markov property!for chains

Broadly speaking, a *stochastic process* is a random function. These functions generally fall into two classes:

(a) **Chains**, in which the domain of the function is the integers; and
(a) **Processes**, in which the domain of the function is the real numbers.

In the former case we usually write $X(n)$ or X_n and in the latter case we usually write $X(t)$. In each case, X represents a random variable. The domain (either $\mathbb{Z}$ or $\mathbb{R}$) is sometimes called the **phase space**, while the range is called the **state space**. Stochastic processes are also sometimes called **time series**.

There are many examples of stochastic processes. The closing prices in a stock exchange, the spot price of oil, and the number of telephone calls passing through a switch at time t are just a few of the many examples that arise in applications.

We will begin by studying **chains** $X(n)$ or X_n because of their greater simplicity. However, chains are also useful in approximating more complex continuous-time processes.

With no further assumptions, very little could be said about random chains. However many examples of interest share an important property: the future evoltuion of the system (the value of $X(n+1)$) can be predicted knowing only the present state of the system (the value of $X(n)$); knowledge of the history prior to the present is irrelevant. Mathematically, this says that

$$\begin{aligned} \mathfrak{Pr}\left(X(n+1) = y \middle| X(0) = x_0, X(1) = x_1, \cdots \right.\\ \left. \cdots X(n-1) = x_{n-1}, X(n) = x_n\right) = \\ = \mathfrak{Pr}\left(X(n+1) = y \middle| X(n) = x_n\right). \end{aligned}$$

This is called the **Markov property** and we will assume that this property holds henceforward.

In this section we will consider a very simple example of a *Markov Chain.* While the example is simple, it illustrates some of the most basic concepts and questions that arise in the study of Stochastic Processes. The simplicity of the example has the additional advantage of providing an elegant solution. While the solution we present generalizes, it is usually too complex computationally to be useful for more complex processes.

25.1. Example. Two-state Markov chain.

Suppose that a machine has two states, working and not working. Suppose further that

$$\mathfrak{Pr}\left(\textit{Machines works today}\middle|\textit{Machine did not work yesterday}\right) = p$$

and

$$\mathfrak{Pr}\left(\textit{Machines does not work today}\middle|\textit{Machine did work yesterday}\right) = q.$$

Suppose further that whether or not the machine works today depends only on whether or not it worked yesterday. Then if

$$X(n) = \begin{cases} 0 & \textit{machine is } \textbf{not} \textit{ working on day } n \\ 1 & \textit{machine } \textbf{is} \textit{ working on day } n \end{cases}$$

we can conclude

$$\mathfrak{Pr}\left(X(n) = 1\middle|X(n-1) = 0\right) = p$$

and

$$\mathfrak{Pr}\left(X(n) = 0\middle|X(n-1) = 1\right) = q.$$

Since the machine either works or not

$$\mathfrak{Pr}\left(X(n) = 0\middle|X(n-1) = 0\right) = 1 - p$$

and

$$\mathfrak{Pr}\left(X(n) = 1\middle|X(n-1) = 1\right) = 1 - q.$$

Finally, we suppose that these probabilities do not depend on the particular day when we we look:

$$\mathfrak{Pr}\left(X(n+m) = 1\middle|X(n+m-1) = 0\right) = p$$

and

$$\mathfrak{Pr}\left(X(n+m) = 0\middle|X(n+m-1) = 1\right) = q.$$

for all choices of $n, m > 0$.

Are there any possible conclusions about the long term behavior of this machine?

For example, can we conclude in the long run what the chances are that the machine will be out of service? This is analogous to asking if the limit

$$\lim_{n\to\infty} \mathfrak{Pr}\left(X(n) = 0\right)$$

exists. If the limit exists, we would certainly like to know what it is. Another question might be how many days, on average, is the machine in service, i.e., is there a value to the limit

$$\lim_{n\to\infty} E\left(\frac{1}{n+1}\sum_{k=0}^{n} X(k)\right).$$

Intuitively one might expect these two limits to be related in some manner, and it turns out that they are.

When the machine is first installed, there is some initial probability that it will work (or not) on installation. In terms $X(n)$ this can be described by

$$\pi_0(k) = \mathfrak{Pr}\left(X(0) = k\right)$$

where k can assume values from the set $\{0, 1\}$. For example, if the machine is manufactured with a high degree of reliability, then we might have

$$\pi_0(0) = 0.02 \quad \text{and} \quad \pi_0(1) = 0.98.$$

More broadly speaking, π_n is the distribution of X_n, i.e.,

$$\pi_n(k) = \mathfrak{Pr}\left(X(n) = k\right)$$

where k can assume values from the set $\{0, 1\}$.

Given an initial probability distribution π_0 we can easily calculate the probabilities that the machine is working or not on day one:

$$\begin{aligned}
\mathfrak{Pr}\left(X(1) = 1\right) &= \mathfrak{Pr}\left(X(1) = 1 \text{ and } X(0) = 0\right) + \mathfrak{Pr}\left(X(1) = 1 \text{ and } X(0) = 1\right) \\
&= \mathfrak{Pr}\left(X(1) = 1 \middle| X(0) = 0\right) \mathfrak{Pr}\left(X(0) = 0\right) + \cdots \\
&\quad \cdots + \mathfrak{Pr}\left(X(1) = 1 \middle| X(0) = 1\right) \mathfrak{Pr}\left(X(0) = 1\right) \\
&= p\pi_0(0) + (1-q)\pi_0(1) \\
&= p(1-\pi_0(1)) + (1-q)\pi_0(1) \\
&= p + (1-p-q)\pi_0(1).
\end{aligned}$$

In fact a more general result follows readily.

25.2. Proposition.

For $n > 1$

$$\mathfrak{Pr}\left(X(n)=1\right)=\frac{p}{p+q}+(1-p-q)^n\left(\pi_0(1)-\frac{p}{p+q}\right).$$

Proof. We can proceed by induction, with the previous calculation providing a basis for the case $n=1$:

$$\begin{aligned}\mathfrak{Pr}\left(X(1)=1\right)&=p+(1-p-q)\pi_0(1)\\&=\frac{p(p+q)}{p+q}+(1-p-q)^1\left(\pi_0(1)-\frac{p}{p+q}\right)+(1-p-q)\frac{p}{p+q}\\&=\frac{p}{p+q}+(1-p-q)^1\left(\pi_0(1)-\frac{p}{p+q}\right).\end{aligned}$$

Since this bases the induction in the case $n=1$ we may now assume the case n and deduce the case $n+1$:

$$\begin{aligned}\mathfrak{Pr}\left(X(n+1)=1\right)&=\mathfrak{Pr}\left(X(n+1)=1\text{ and }X(n)=0\right)+\cdots\\&\qquad\cdots+\mathfrak{Pr}\left(X(n+1)=1\text{ and }X(n)=1\right)\\&=\mathfrak{Pr}\left(X(n+1)=1\middle|X(n)=0\right)\mathfrak{Pr}\left(X(n)=0\right)+\cdots\\&\qquad\cdots+\mathfrak{Pr}\left(X(n+1)=1\middle|X(n)=1\right)\mathfrak{Pr}\left(X(n)=1\right)\\&=p\,\mathfrak{Pr}\left(X(n)=0\right)+(1-q)\,\mathfrak{Pr}\left(X(n)=1\right)\\&=p+(1-p-q)\,\mathfrak{Pr}\left(X(n)=1\right)\\&=p+(1-p-q)\left(\frac{p}{p+q}+(1-p-q)^n\left(\pi_0(1)-\frac{p}{p+q}\right)\right)\\&=\frac{p}{p+q}+(1-p-q)^{n+1}\left(\pi_0(1)-\frac{p}{p+q}\right).\end{aligned}$$

∎

25.3. Corollary.

For $n > 1$

$$\mathfrak{Pr}\left(X(n) = 0\right) = \frac{q}{p+q} + (1-p-q)^n \left(\pi_0(0) - \frac{q}{p+q}\right).$$

25.4. Corollary.

If $2 > p + q > 0$ *then*

$$\lim_{n\to\infty} \mathfrak{Pr}\left(X(n) = 1\right) = \frac{p}{p+q}$$

and

$$\lim_{n\to\infty} \mathfrak{Pr}\left(X(n) = 0\right) = \frac{q}{p+q}.$$

Notice that if $p + q = 2$, then a machine that is working today is always broken tomorrow and a machine that is broken today is always working tomorrow. Thus no matter what the initial state, the value of $X(n)$ oscillates between zero and one and hence there can be no limiting behavior.

Note that for any $x, y \in \{0, 1\}$ we can define

$$p(x, y) = \mathfrak{Pr}\left(X(1) = y \middle| X(0) = x\right).$$

The function p gives the *transition probabilities* for the Markov Chain. The transition probabilities in turn give rise to a *transition matrix* P. In our example,

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

Note that if we multiply the row vector

$$\begin{pmatrix} \pi_0(0) & \pi_0(1) \end{pmatrix}$$

by the transition matrix, we obtain the state of the system at time one:

$$\begin{pmatrix} \pi_1(0) & \pi_1(1) \end{pmatrix}$$

i.e.,

$$\begin{pmatrix} \pi_0(0) & \pi_0(1) \end{pmatrix} \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} = \begin{pmatrix} (1-p)\pi_0(0) + q\pi_0(1) \\ p\pi_0(0) + (1-q)\pi_0(1) \end{pmatrix}.$$

More generally,

$$\begin{pmatrix} \mathfrak{Pr}\,(X(n)=0) & \mathfrak{Pr}\,(X(n)=1) \end{pmatrix} = \begin{pmatrix} \pi_0(0) & \pi_0(1) \end{pmatrix} \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}^n$$

25.5. Definition.

A **stationary distribution** *for the two-state Markov Chain is a distribution on* $\{0,1\}$ *satisfying*

$$\begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix} \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} = \begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix}.$$

25.6. Proposition.

Suppose that

$$\begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix}$$

satisfies

$$\lim \begin{pmatrix} \pi_n(0) & \pi_n(1) \end{pmatrix} = \begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix}.$$

Then $\begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix}$ *is a stationary distribution.*

Proof. Suppose that the initial distribution corresponds to $\begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix}$. Then

$$\begin{aligned} \lim_{n\to\infty} \begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix} P^n &= \lim_{n\to\infty} \begin{pmatrix} \pi_n(0) & \pi_n(1) \end{pmatrix} \\ &= \begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix} \end{aligned}$$

Then it follows that

$$\begin{aligned} \begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix} P &= \lim_{n\to\infty} \begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix} P^n P \\ &= \lim_{n\to\infty} \begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix} P^{n+1} \\ &= \begin{pmatrix} \pi(0) & \pi(1) \end{pmatrix}. \end{aligned}$$

∎

In other words, if a limiting distribution exists, then it must also be a stationary distribution. Under mild assumptions, the converse is true, i.e., that if a stationary distribution exists then it is a limiting distribution. We will illustrate this by using some results from linear algebra.

There is a technique from linear algebra that reduces the problem of finding the stationary distribution to one of finding the eigenvalues and corresponding eigenvectors of the transition matrix. To do this we need to recall the following theorem.

25.7. Theorem.

Let M be a two-by-two matrix and suppose that M has two distinct eigenvalues λ_1 and λ_2 with corresponding eigenvectors v_1 and v_2. Let E be the matrix whose columns consist of v_1 and v_2 respectively and let D be the matrix

$$D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

Then

$$M = E \cdot D \cdot E^{-1}$$

and, in particular,

$$M^n = E \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix} E^{-1}.$$

25.8. Example.

In our example,

$$M = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

We will find the eigenvalues, eigenvectors and the matrix E and E^{-1} to illustrate the above theorem.

Solution. For this choice of M the characteristic equation is

$$\lambda^2 - (2-p-q)\lambda + 1 - p - q = 0$$

which leads to eigenvalues of

$$\lambda_1 = 1 \quad \text{and} \quad \lambda_2 = 1 - p - q$$

with corresponding eigenvectors of

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -p \\ q \end{pmatrix}.$$

Then

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 1 - p - q \end{pmatrix}$$

and

$$E = \begin{pmatrix} -p & 1 \\ q & 1 \end{pmatrix}.$$

Assmuing that $p + q > 0$, it then follows that

$$\begin{aligned} M^n &= E \cdot D^n \cdot E^{-1} \\ &= \begin{pmatrix} 1 & -p \\ 1 & q \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & (1 - p - q)^n \end{pmatrix} \cdot \frac{1}{p+q} \begin{pmatrix} q & p \\ -1 & 1 \end{pmatrix}. \end{aligned}$$

Assuming that $2 > p + q > 0$ it follows that

$$\begin{aligned} \lim_{n\to\infty} M^n &= \lim_{n\to\infty} E \cdot \begin{pmatrix} 1 & 0 \\ 0 & (1 - p - q)^n \end{pmatrix} E^{-1} \\ &= E \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot E^{-1} \\ &= \frac{1}{p+q} \begin{pmatrix} q & p \\ q & p \end{pmatrix} \end{aligned}$$

Now for *any* initial distribution

$$\begin{aligned} \lim_{n\to\infty} \left(\pi_n(0) \quad \pi_n(1) \right) &= \lim_{n\to\infty} \left(\pi_0(0) \quad \pi_0(1) \right) M^n \\ &= \left(\pi_0(0) \quad \pi_0(1) \right) \frac{1}{p+q} \begin{pmatrix} q & p \\ q & p \end{pmatrix} \\ &= \left(\tfrac{q}{p+q} \quad \tfrac{p}{p+q} \right) \end{aligned}$$

since

$$\pi_0(0) + \pi_0(1) = 1.$$

This establishes both the existence of the stationary distribution and the fact that

$$\lim_{n\to\infty} \left(\pi_n(0) \quad \pi_n(1) \right) = \left(\tfrac{q}{p+q} \quad \tfrac{p}{p+q} \right)$$

∎

Many of the results of interest about Markov Chains can be deduced using linear algebra techniques similar to the above. While this is an elegant and self-contained approach, we will restrict this approach to the problem sets. There are two main reasons reasons for not using this approach in the main arguments in this text. First, this approach presumes the reader is familiar with fundamental concepts from linear algebra such as canonical forms and diagonalization which may not be true for all or even most readers. The second reason has to do with the values some of the most important processes can assume. In our present example, the process $\{X(n)\}$ could assume only two values: 0 or 1. As we shall see in the next section, there are many important examples of random process which can assume a finite number of values ($\{0, 1, \cdots, n\}$) or an infinite number of values ($\{0, 1, \cdots, n, \cdots\}$). In the latter case the resulting “matrix”of transition probabilities is infinite rather than finite. While there are advanced algebraic techniques for analyzing the behavior of such matrices, they will certainly be unfamiliar to most readers of this text. For these reasons, we relegate the use of linear algebra in analyzing Markov Chains to the exercies.

25. Two State Markov Chains: Problems.

1. In the two-state chain, find

$$\mathfrak{Pr}\,(X_1 \neq X_2).$$

2. In the two-state chain, find

$$\mathfrak{Pr}\,(X_1 \geq X_2).$$

3. Let T_0 denote the first time $n > 0$ that the the two-state chain is in state 0. Find

$$\mathfrak{Pr}\,(T_0 = n \big| X(0) = 0).$$

If the two-state chain starts in state 0, how long, on average, will it be before it first returns to state zero?

4. Let T_1 denote the first time $n > 0$ that the the two-state chain is in state 1. Find

$$\mathfrak{Pr}\,(T_1 = n\big).$$

phase space

26. Markov Chains – Definitions and Examples

In this section we will introduce the basic definitions of Markov Chains and present several important examples.

26.1. Definition.

A **discrete stochastic process** *is a collection* $X(n)$, $n = 0, 1, 2, \ldots$ *of random variables*

$$X(n) : \Omega \to \mathcal{S}$$

where $(\Omega, \mathcal{E}, \mathfrak{Pr}\,)$ *is a probability space and*

$$\mathcal{S} = \{x_0, x_1, x_2, \ldots\}$$

is a finite or countably infinite collection of symbols or "states." For convenience, we will write X_n *for* $X(n)$. *We assume that the one-step* **transition probabilities**

$$\mathfrak{Pr}\,(X_{n+1} = x_j \big| X_n = x_i)$$

are **stationary**, *i.e., that they are independent of* n*:*

$$\mathfrak{Pr}\,(X_{n+1} = x_j \big| X_n = x_i) = \mathfrak{Pr}\,(X_{m+1} = s_j \big| X_m = x_i).$$

for all choices of m *and* n. *We will write* p_{ij} *for the transition probabilities:*

$$p_{ij} \equiv P(x_i, x_j) \equiv \mathfrak{Pr}\,(X_{n+1} = x_j \big| X_n = x_i).$$

Generally speaking we will not need to explicitly reference Ω or $\mathcal{E}$. In most examples the state space $\mathcal{S}$ can be taken to be either a finite set of integers $\{0, 1, \ldots, N\}$ or the non-negative integers $\{0, 1, \ldots\}$. The set of integers

$$\{n \;\;:\;\; X(n) \text{ is defined}\}$$

is called the *phase space*. For now, both the state space and the phase space are discrete; eventually we will consider processes in which either the phase space or both the phase space and the state space are continuous.

In order to provide additional structure, some additional assumptions are usually required. The most important of these is the Markov Property.

26.2. Definition.

A stochastic process $\{X_n\}$ *has the* **Markov Property** *if*

$$\mathfrak{Pr}\left(X_n = x_n \middle| X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0\right) = \cdots$$
$$\cdots = \mathfrak{Pr}\left(X_n = x_n \middle| X_{n-1} = x_{n-1}\right).$$

In particular, if the Markov Property holds, then the one-step transition probability from time $n-1$ *to time* n *depends only on the state of the system at the immediately preceding time* $n-1$ *and not on the entire prior history. A discrete process having the Markov Property is said to be a* **Markov Chain.**

26.3. Definition.

The **initial distribtution** *for a Markov Chain is*

$$\pi_0(x_k) = \mathfrak{Pr}\left(X_0 = x_k\right).$$

The distribution of X_n *is commonly denoted by*

$$\pi_n(x_k) = \mathfrak{Pr}\left(X_n = x_k\right).$$

The joint distribution of $\{X_0, X_1, \ldots, X_n\}$ can be readily described in terms of the initial distribution and the transition probabilities.

26.4. Proposition.

For any Markov Chain

$$\mathfrak{Pr}\left(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n\right) =$$
$$= \pi(x_0)P((x_0, x_1)P(x_1, x_2)\cdots P(x_{n-1}, x_n).$$

Proof. For example,

$$\begin{aligned}\mathfrak{Pr}\left(X_0 = x_0, X_1 = x_1\right) &= \mathfrak{Pr}\left(X_0 = x_0\right)\mathfrak{Pr}\left(X_1 = x_1 \middle| \mathfrak{Pr}\left(X_0 = x_{x_0}\right)\right.\\ &= \pi_0(x_0)p_{0,1}\\ &(= \pi_0(x_0)P(x_0, x_1))\end{aligned}$$

Proceding by induction, we now assume that

$$\mathfrak{Pr}\left(X - 0 = x_0, X_1 = x_1, \dots, X_n = x_n\right) =$$
$$= \pi(x_0)P(x_0, x_1)P(x_1, x_2)\cdots P(x_{n-1}, x_n)$$

and verify

$$\mathfrak{Pr}\left(X_0 = x_0, X_1 = x_1, \dots, X_{n+1} = x_{n+1}\right) =$$
$$= \pi(x_0)P(x_0, x_1)P(x_1, x_2)\cdots P(x_{n+1}, x_{n+1})$$

Applying the definition of conditional probability,

$$\begin{aligned}&\mathfrak{Pr}\left(X_0 = x_0, X_1 = x_1, \dots, X_{n+1} = x_{n+1}\right) =\\ &\quad= \mathfrak{Pr}\left(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n\right)\cdots\\ &\qquad\cdots\mathfrak{Pr}\left(X_{n+1} = x_{n+1} \middle| X(0) = x_0, X(1) = x_1, \dots, X(n) = x_n\right)\\ &\quad= \pi(x_0)P(x_0, x_1)P(x_1, x_2)\cdots P(x_{n-1}, x_n)\cdots\\ &\qquad\cdots\mathfrak{Pr}\left(X_{n+1} = x_{n+1} \middle| X(n) = x_n\right)\\ &\quad= \pi(x_0)P(x_0, x_1)P(x_1, x_2)\cdots P(x_n, x_{n+1})\end{aligned}$$

applying, in order, the definition of condtional probability, the inductive hypothesis, the Markov Property, and stationary transition probabilities.

Next we informally define the n-step transition probabilities. We will give a more rigorous defnition in the next section.

∎

26.5. Definition.

The n-step transition probabilities are

$$P^n(x_i, x_j) = \mathfrak{Pr}\left(X_n = x_j \middle| X_0 = x_i\right).$$

Intuitively this follows from stationary transition probablities,

$$\begin{aligned} P^n(x_i, x_j) &= \mathfrak{Pr}\left(X_n = x_j \middle| X_0 = x_i\right) \\ &= \mathfrak{Pr}\left(X_{m+n} = x_j \middle| X_m = x_i\right) \end{aligned}$$

for any choice of m, a fact we shall deduce in the next section.

26.6. Proposition. Chapman-Kolmogorov Equation.

Let $\{X_n\}$ be a Markov chain having transition function P. Then

$$\pi_1(x_k) = \sum_i \pi_0(x_i) P(x_i, x_k).$$

Proof. Note that

$$\begin{aligned} \pi_1(x_k) &= \mathfrak{Pr}\left(X_1 = x_k\right) \\ &= \sum_i \mathfrak{Pr}\left(X_1 = x_k, \ \ X_0 = x_i\right) \\ &= \sum_i \mathfrak{Pr}\left(X_0 = x_i\right) \mathfrak{Pr}\left(X_1 = x_k \middle| X_0 = x_i\right) \\ &= \sum_i \pi_0(x_i) P(x_i, x_j). \end{aligned}$$

∎

26.7. Definition.

A distribution π is **stationary** *if*

$$\pi(k) = \sum_i \pi(i) p_{i,k}$$

for all k.

Notice that if X_0 has a stationary distribution π, then the Chapman-Kolmogorov equation implies that X_1 has the same distrubtuion. An easy induction then implies that all random variables X_n must have π as their distribution. It is also an easy exercise to verify that if a stationary distriubtion exists it must be unique.

26.8. Example. Ehrenfest Chain.

Suppose that we have two urns and $2R$ balls, numbered $\{1, 2, \ldots, 2R\}$. Initially some of the balls are in one urn and some are in the other. Select a number at random from $\{1, 2, \ldots, 2R\}$ and move the ball with that number from the urn it is in to the other urn. Let X_n denote the number of balls in the first urn after n repetitions of this process. We will find the transition probabilities $p_{i,j}$ and the stationary distribution for this chain.

Solution. Notice that the state space is

$$\mathcal{S} = \{0, 1, \ldots, 2R\}.$$

Now if $X(n)$ is zero, then all the balls are in the other urn, so there is a probability of one that $X(n+1) = 1$. Similarly if $X(n) = 2R$, then all of the balls are in urn one, and so there is a probability of one that $X(n+1) = 2R - 1$. One theother hand, if $0 < i < 2R$, then

$$p_{i,j} = \begin{cases} (2R - i)/2R & \text{if } j = i + 1 \\ i/2R & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases}$$

Notice that $p_{i,j}$ defines a $(2R+1) \times (2R+1)$ matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1/2R & 0 & 1-1/2R & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 2/2R & 0 & 1-2/2R & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1-2/2R & 0 & 2/2R & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1-1/R & 0 & 1/2R \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \end{pmatrix}$$

A stationary distribution will correspond to a row vector π that satisfies

$$\pi P = \pi$$

and $\sum \pi(i) = 1$. We leave it to the exercises to verify that

$$\pi(k) = \frac{(2R)!(0.5)^{2R}}{k!(2R-k)!}$$

is the stationary distribution.

▮

P. and T. Ehrenfest introduced this chain as an example of a stochastic (probabilistic) approach to the physical notion of equilibrium. Instead of using balls and urns, they used fleas jumping between two dogs. Despite the apparent contrived nature of the example, the Ehrenfest chain can be useful in many applications such as heat transfer. It can also be regarded asa discrete approximation of an important continuous-time, continuous state process known as the Ornstein-Uhlenbeck process.

The assumption that he Ehrenfest chain starts with $2R$ balls is historical rather than implicit in the model and some textbooks will refer to the Ehrenfest chain with d balls instead of $2R$ balls.

26.9. Definition.

A state $x_i \in \mathcal{S}$ is said to be **absorbing** *if $p_{i,i} = 1$ or, equivalently, if $p_{i,j} = 0$ whenever $i \neq j$. In particular, if ever $X_n = s_i$ then the chain never leaves the state s_i, i.e., $X_m = s_i$ for all $m \geq n$.*

26.10. Example. Gambler's Ruin Chain.

Suppose that a gambler starts out with a fixed number of dollars and makes a series of one dollar bets. Assume that

$$p = \mathfrak{Pr}\,(\textit{gambler wins \$1})$$

and so

$$1 - p = \mathfrak{Pr}\,(\textit{gambler loses \$1}).$$

Let X_n be the number of dollars the gambler has after n bets. If the gambler ever runs out of dollars – i.e., if ever $X_n = 0$ – then then no further bets are possible so $X_m = 0$ for $m > n$. In particular, 0 is an absorbing state. Otherwise, for $i > 0$

$$p_{i,j} = \begin{cases} 1 - p & \textit{if } j = i - 1 \\ p & \textit{if } j = i + 1 \\ 0 & \textit{otherwise} \end{cases} \tag{26.1}$$

This is called the **Gambler's ruin chain** *on the state space $\mathcal{S} = \{0, 1, 2, \cdots\}$. The gambler might seek to limit losses by deciding to cease playing if his or her total dollars ever reach $\$d$. In this latter case d is also an absorbing state and (26.1) holds for $0 < i < d$.*

We will eventually show that if $p \leq 0.5$ then the gambler eventually goes bankrupt ($X_n = 0$ for some n) with probability one, hence the name of the chain.

Both the Ehrenfest Chain and the Gambler's Ruin chain are special cases of the more general birth and death chain.

26.11. Example. Birth and Death Chain.

Suppose that $\mathcal{S} = \{0, 1, 2, \cdots\}$ or that $\mathcal{S} = \{0, 1, 2, \cdots, d\}$. Supose that if $X_n = i$ then X_{n+1} can assume only the values $i - 1$, i or $i + 1$ and that the probability of the transitions are given by

$$p_{i,j} = \begin{cases} q_i & \text{if } j = i - 1, \\ r_i & \text{if } j = i, \\ p_i & \text{if } j = i + 1, \text{ and} \\ 0 & \text{elsewhere} \end{cases}$$

where $p_i \geq 0$, $r_i \geq 0$, $q_i \geq 0$ and $p_i + r_i + q_i = 1$. It is usually also assumed that 0 is absorbing, i.e., that $r_0 = 1$ and $p_0 = 0 = q_0$. If $\mathcal{S} = \{0, 1, 2, \cdots, d\}$ then d is assumed to be absorbing as well. This latter case is sometimes called the birth and death chain with absorbing boundaries.

The name of this chain comes from applications where X_n is the population of the n^{th} generation of living organisms. In this case, the transition from i to $(i + 1)$ constitutes a birth and the transition from i to $(i - 1)$ constitutes a death.

In many settings we will study objects which can generate new objects. These objects could be, for example, neutrons, bacteria, or the male lineage in a royal family. In each case the number of objects in the $(n + 1)^{st}$ generation depends on the number of objects in the n^{th} generation as well as the probability that any object (neutron, bacterium, king) survives to the $(n + 1)^{st}$ generation and – possibly – generates new particles (neutrons, bacteria, male heirs) in the $(n + 1)^{st}$ generation. This gives rise to the next example.

26.12. Example. Branching Chain.

Suppose that $\{\xi_n\}_{n=1}^{\infty}$ *are independent, identically distributed, non-negative, integer-valued random variables having a common density function* f*. We set*

$$p_{i,j} = \mathfrak{Pr}\left(\xi_1 + \cdots + \xi_i = j\right)$$

for all j *and for* $i > 0$ *and set* $p_{0,0} = 1$*. If we assume an initial distribution* π_0 *on* $\mathcal{S} = \{0, 1, 2, \cdots\}$ *and define recursively*

$$\pi_n(j) = \mathfrak{Pr}\left(X_n = j\right) = \sum_i p_{i,j}\pi_{n-1}(i)$$

then the reader can readily verify that X_n *is a Markov Chain with stationary transition probabilities. Indeed,*

$$p_{1,j} = f(j)$$

for $j \geq 0$.

It is possible that, after some number of generations, the original particle and all of its descendents will have died out. If this happens, the particle is said to have become *extinct*. An interesting problem is computing the probability of extinction for a particular paricle, i.e., computing the probability that a branching chain that starts in state 1 will eventually be absorbed to state zero.

26.13. Example. Renewal Chain.

Suppose that a process arises in the following manner: a item, such as a light bulb, is checked periodically to see if it is working. If it has failed after check at time n *it is replaced at time* $n+1$*. From the time it is first installed, it has it has probability* p_j *of failing between the* j^{th} *and* $(j+1)^{st}$ *period of its working life. Let* X_n *be the age, in periods, of the item in use at time* n*. Note that* $X_n = 0$ *for an item that has been replaced at time* n*. Then*

$$\mathfrak{Pr}\left(X_n = k \middle| X_{n-1} = j\right) = \begin{cases} p_j & k = 0 \\ 1 - p_j & k = j+1 \\ 0 & \text{otherwise} \end{cases}$$

Using this model, on average what would you expect to be the average lifetime of a lightbulb in use? Are there any implications for replacement scheduling?

In the next example we suppose that customers (airplanes, phone calls) arrive at a service queue (airport, switch). In any particular unit of time, exactly one customer will be served. However, during that time some random number ξ of customers will have arrived. The result is a simple model for many applications involving queues – and provides a discrete approximation for more complex models.

26.14. Example. Queuing Chain.

Suppose that $\{\xi_n\}_{n=1}^{\infty}$ are independent, identically distributed, non-negative, integer-valued random variables having a common density function f. We set

$$X_{n+1} = X_n + \xi_{n+1} - 1$$

. Then X_n is a Markov Chain with state space $\mathcal{S} = \{0, 1, \cdots\}$ and

$$p_{i,j} = \begin{cases} f(j) & \text{if } i = 0 \text{ ,} \\ f(j - i + 1) & \text{if } i \geq 1 \end{cases}$$

26. Markov Chains – Definitions and Examples: Problems.

1. Show that if a stationary distribution exists it must be unique.

2. Verify that

$$\pi(k) = \frac{(2R)!(0.5)^{2R}}{k!(2R-k)!}$$

defines the stationary distribution for the Ehrenfest Chain.

3. Suppose that we have a queue with a fixed service time per customer – for convenience, think of this as a taxi stand in which taxis arrive one per minute and serve exactly one customer at a time. If there is a customer waiting, exactly one customer is served with probability one in that interval; if there are no customers waiting in the queue, the taxi departs and there's another minute lapse before the next taxi arrives. Meantime, customers are arriving at the queue according to the Poison distribution, i.e, the probability that k new customers arrive in any service interval is

$$\mathfrak{Pr}\,(\text{exactly } k \text{ new arrivals}) = \frac{\lambda^k}{k!}e^{-\lambda}$$

for $k = 0, 1, \ldots$. We suppose that the number of new customers arriving in the n^{th} interval is independent of the number arriving in the m^{th} interval, for $m \neq n$. Let $\{X_n\}$ be the number of customers waiting in the queue at the start of the n^{th} interval. Show that $\{X_n\}$ is a Markov Chain and find the transition function.

4. In the renewal chain, suppose that

$$p_i = p(1-p)^i \quad i = 0, 1, 2, \cdots$$

This distribuiton is not unreasonable when modelling electrical components. Suppose that the model is revised so that the bulb is automatically replaced, whether it has failed or not, if its age ever reaches $\frac{2p}{1-p}$ intervals. What are the transition probabilities in this case?

5. Let $\{X_n\}$ be a Markov chain whose state space $\mathcal{S}$ is a subset of the non-negative integers and whose transition function satisfies the idenity

$$\sum_y yP(x,y) = Ax + B$$

for all $x \in \mathcal{S}$ and for some constants A and B.
(a) Show that $E(X_{n+1} = AE(X_n) + B$.
(b) Show that, if $A \neq 1$, then

$$E(X_n) = \frac{B}{1-A} + A^n\left(E(X_0) - \frac{B}{1-A}\right).$$

6. Let X_n be the Ehrenfest Chain on $\{0, 1, \cdots, d\}$ and show that the assumption of the previous exercise holds. Use this to calculate $E_x(X_n) \equiv E(X_n \big| X_0 = x)$.

27. Calculations with Transition Functions

Because of the assumptions regarding transition functions outlined at the start of the previous section, there is a one-to-one correspondence between a Markov chain X_n and its transition function $P(x, y)$. Thus for every chain there is a unique transition function and for every transition function there is a unique Markov chain. For this reason the study of Markov chains involves deducing properties of the transition function.

We begin by collecting some basic formulae.

27.1. Proposition.

Let $\{X_n\}$ be a Markov chain on a state space $\mathcal{S}$ having transition function P. Then

$$\begin{aligned}\mathfrak{Pr}\left(X_{n+1} = x_{n+1}, \cdots, X_{n+m} = x_{n+m} \middle| X_0 = x_0, \cdots, X_n = x_n\right) = \\ = P(x_n, x_{n+1}) \cdots P(x_{n+m-1}, x_{n+m}).\end{aligned}$$

Proof. First write the left-hand-side of the conclusion as

$$\frac{\mathfrak{Pr}\left(X_0 = x_0, \cdots, X_{n+m} = x_{n+m}\right)}{\mathfrak{Pr}\left(X_0 = x_0, \cdots, X_n = x_n\right)}.$$

By Proposition 26.4, this reduces to

$$\frac{\pi_0(x_0)P(x_0, x_1) \cdots P(x_{n+m-1}, x_{n+m})}{\pi_0(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n)}$$

which proves the result upon canceling terms.

∎

The following proposition is useful in working deducing properties of the transition function.

27.2. Proposition.

Let $(\Omega, \mathcal{M}, \mathfrak{Pr})$ be a probability space and let the events in the conclusions listed below all be in $\mathcal{M}$. Then

(a) If $\{D_i\}$ are all disjoint and if, for each i, $\mathfrak{Pr}\,(C\big|D_i) = p$ then $\mathfrak{Pr}\,(C\big|\cup_i D_i) = p$.

(b) If $\{C_i\}$ are all disjoint then

$$\mathfrak{Pr}\,(\cup_i C_i\big|D) = \sum_i \mathfrak{Pr}\,(C_i\big|D).$$

(c) If $\{E_i\}$ are all disjoint and $\cup_i E_i = \Omega$ then

$$\mathfrak{Pr}\,(C\big|D) = \sum_i \mathfrak{Pr}\,(E_i\big|D)\,\mathfrak{Pr}\,(C\big|E_i \cap D).$$

(d) If $\{C_i\}$ are all disjoint and if $\mathfrak{Pr}\,(A\big|C_i) = \mathfrak{Pr}\,(B\big|C_i)$ for all i, then

$$\mathfrak{Pr}\,(A\big|\cup_i C_i) = \mathfrak{Pr}\,(B\big|\cup_i C_i).$$

Proof. We will prove only (a) and leave the remaining similar proofs to the exercises. Observe that

$$\begin{aligned}
\mathfrak{Pr}\,(C\big|\cup_i D_i) &= \frac{\mathfrak{Pr}\,(C \cap (\cup_i D_i))}{\mathfrak{Pr}\,(\cup_i D_i)} \\
&= \frac{\mathfrak{Pr}\,(\cup_i (C \cap D_i))}{\mathfrak{Pr}\,(\cup_i D_i)} \\
&= \frac{\sum_i \mathfrak{Pr}\,(C \cap D_i)}{\sum_i \mathfrak{Pr}\,(D_i)} \\
&= \frac{\sum \mathfrak{Pr}\,(C\big|D_i)\,\mathfrak{Pr}\,(D_i)}{\sum_i \mathfrak{Pr}\,(D_i)} \\
&= p
\end{aligned}$$

∎

27.3. Proposition.

For subsets $\{A_0, A_1, \ldots, A_{n-1}\}$ *of the state space* $\mathcal{S}$*,*

$$\mathfrak{Pr}\left(X_{n+1} = y_1, \ldots, X_{n+m} = y_m \middle| X_0 \in A_0, \ldots X_{n-1} \in A_{n-1}, X_n = x\right)$$
$$= P(x, y_1)P(y_1, y_2) \cdots P(y_{m-1}, y_m)$$

Proof. This follows at once from 27.1 and 27.2(a). ∎

27.4. Proposition.

For subsets $\{A_0, A_1, \ldots, A_{n-1}\}$ *and* $\{B_1, \ldots, B_m\}$ *of the state space* $\mathcal{S}$*,*

$$\mathfrak{Pr}\left(X_{n+1} \in B_1, \ldots, X_{n+m} \in B_m \middle| X_0 \in A_0, \ldots X_{n-1} \in A_{n-1}, X_n = x\right)$$
$$= \sum_{y_1 \in B_1} \cdots \sum_{y_m \in B_m} P(x, y_1)P(y_1, y_2) \cdots P(y_{m-1}, y_m)$$

Proof. This follows from 27.3 and 27.2(b). ∎

Recall in the previous section we defined the m-step transition function to be

$$\mathfrak{Pr}\left(X_{n+m} = y \middle| X_n = x\right) = P^m(x, y)$$

which, because of stationary transition probabilities becomes

$$\mathfrak{Pr}\left(X_m = y \middle| X_0 = x\right) = P^m(x, y).$$

The previous proposition lets us rephrase this in an alternative fashion.

27.5. Proposition.

For subsets $\{A_0, A_1, \ldots, A_{n-1}\}$ *of the state space* $\mathcal{S}$*,*

$$\begin{aligned} &P(X_{n+m} = y \big| X_0 \in A_0, \ldots, X_{n-1} \in A_{n-1}, X_n = x) = \\ &= \sum_{y_1} \cdots \sum_{y_{m-1}} P(x, y_1) P(y_1, y_2) \cdots P(y_{m-2}, y_{m-1}) P(y_{m-1}, y). \end{aligned}$$

Proof. This follows upon taking $B_1, \ldots, B_{m-1}$ to be $\mathcal{S}$ and $B_m = \{y\}$ in 27.4. ∎

In particular, then we can deduce the following

27.6. Proposition.

The m**-step transition function** $P^m(x, y)$ *is given by*

$$P^m(x, y) = \sum_{y_1} \cdots \sum_{y_{m-1}} P(x, y_1) P(y_1, y_2) \cdots P(y_{m-2}, y_{m-1}) P(y_{m-1}, y)$$

for $m \geq 2$*, by* $P^1(x, y) = P(x, y)$ *for* $m = 1$ *and by*

$$P^0(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

Proof. This follows upon setting $A_0, A_1, \ldots, A_{n-1}$ to be $\mathcal{S}$ in the previous proposition. ∎

27.7. Proposition.

Let $\{X_n\}$ be a Markov process. Then

$$P^{n+m}(x,y) = \sum_z P^n(x,z)P^m(z,y).$$

Proof. First note that from the above

$$\mathfrak{Pr}\left(X_{n+m}=y \middle| X_0=x, X_n=z\right) = P^m(z,y).$$

Now, by 27.2(c)

$$\begin{aligned} P^{n+m}(x,y) &= \mathfrak{Pr}\left(X_{n+m}=y \middle| X_0=x\right) \\ &= \sum_z \mathfrak{Pr}\left(X_n=z \middle| X_0=x\right) \mathfrak{Pr}\left(X_{n+m}=y \middle| X_0=x, X_n=z\right) \\ &= \sum_z P^n(x,z)P^m(z,y) \end{aligned}$$

∎

As a consequence of the above observations, if a Markov chain has a finite state space then we can think of P^n as the nth power of the transition matrix.

27.8. Example.

Suppose that $\{X_n\}$ is the gambler's ruin chain on $\mathcal{S} = \{0,1,2,3\}$. Find the transition matrix P for X_n.

Solution. In general the transition matrix on a finite state space $\mathcal{S} = \{0,1,\cdots,d\}$ has the form

$$P = \begin{array}{c} \\ 0 \\ \vdots \\ d \end{array}\begin{array}{c} \begin{array}{ccc} 0 & \cdots & d \end{array} \\ \begin{pmatrix} P(0,0) & \cdots & P(0,d) \\ & \vdots & \\ P(d,0) & \cdots & P(d,d) \end{pmatrix} \end{array}$$

In the case $n = 3$ and the gambler's ruin chain this becomes

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ q & 0 & p & 0 \\ 0 & q & 0 & p \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Because of 27.8,

$$P^2(x, y) = \sum_z P(x, z)Pz, y)$$

which is exactly the definition of the product of the matrix P with itself or

$$P^2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ q & qp & 0 & p^2 \\ q^2 & 0 & qp & p \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

∎

We conclude this section with the notion of **hitting times** which will be important in our analysis of Markov chains. First we prove another preliminary proposition.

27.9. Proposition.

Let A, B and C be events. If $\mathfrak{Pr}\,(C \cap A) \neq 0$, then

$$\mathfrak{Pr}\,(A \cap B \,|\, C) = \mathfrak{Pr}\,(A \,|\, C)\, \mathfrak{Pr}\,(B \,|\, A \cap C).$$

Proof. Note that if $\mathfrak{Pr}\,(C \cap A) \neq 0$ then $\mathfrak{Pr}\,(C) \neq 0$. Thus

$$\begin{aligned} \mathfrak{Pr}\,(A \cap B \,|\, C)\, \mathfrak{Pr}\,(C) &= \mathfrak{Pr}\,(A \cap B \cap C) \\ &= \mathfrak{Pr}\,(B \,|\, C \cap A)\, \mathfrak{Pr}\,(C \cap A) \\ &= \mathfrak{Pr}\,(B \,|\, C \cap A)\, \mathfrak{Pr}\,(A \,|\, C)\, \mathfrak{Pr}\,(C). \end{aligned}$$

Dividing by $\mathfrak{Pr}\,(C) \neq 0$ gives the result.

∎

27.10. Definition.

Let $A \subseteq \mathcal{S}$ be a set of states, and consider the event that the chain eventually assumes one of the values in the set A, i.e.

$$\{X_n \in A \text{ for some value of } n > 0\}$$

We will be interested in calculating the probability of this happening, subject to the condition that the event $\{X_0 = x\}$ has occurred. In this case, we will write

$$\mathfrak{Pr}\,(X_n \in A \text{ for some value of } n > 0 | X_0 = x) \equiv P_x(A) \equiv \mathfrak{Pr}\,(A|X_0 = x).$$

Thus the conclusion of the proposition can be re-written as

$$P_x(A \cap B) = P_x(A)P_x(B\big|A). \tag{27.1}$$

27.11. Definition.

Let A be any subset of the state space $\mathcal{S}$. The **hitting time** T_A *of A is*

$$T_A = \min\{n > 0 \;\;:\;\; X_n \in A\}$$

In particular, T_A is the first time $n \geq 1$ that $X_n \in A$. Note that it is possible that X_n is never in A, in which case $T_A = \infty$. In many cases A will be a singleton set $\{x\}$ in which case we will write

$$T_x \equiv T_{\{x\}}.$$

27.12. Proposition.

For any $n \geq 1$ and any states x and y in the state space $\mathcal{S}$

$$P^n(x,y) = \sum_{m=1}^{n} P_x(T_y = m) P^{n-m}(y,y).$$

Proof. For fixed n and for $1 \leq m \leq n$ the events

$$\{T_y = m \quad \text{and} \quad X_n = y\}$$

are disjoint. Thus

$$\{X_n = y\} = \bigcup_{m=1}^{n} \{T_y = m \quad \text{and} \quad X_n = y\}.$$

From this

$$\begin{aligned}
P^n(x,y) &= P_x(X_n = y) \\
&= \sum_{m=1}^{n} P_x(T_y = m \ \text{ and } \ X_n = y) \\
&= \sum_{m=1}^{n} P_x(T_y = m) \, \mathfrak{Pr}\left(X_n = y \middle| X_0 = x \ \text{ and } \ T_y = m\right) \\
&= \sum_{m=1}^{n} P_x(T_y = m) \, \mathfrak{Pr}\left(X_n = y \middle| X_0 = x, X_1 \neq y, \cdots, X_{m-1} \neq y, X_m = y\right) \\
&= \sum_{m=1}^{n} P_x(T_y = m) P^{n-m}(y,y)
\end{aligned}$$

∎

In words, the proposition is saying that in order to go from x to y in n steps, we must be in state y for the first time at some step m where $1 \leq m \leq n$, and then transition from y back to y in the remaining $n - m$ steps. Summing the probabilities of these disjoint events gives the result.

27.13. Example.

If a is an absorbing state, then

$$P^n(x,a) = P_x(T_a \le n)$$

for all $n \ge 1$.

Proof. If a is absorbing, then $P^{n-m}(a,a) = 1$ for all $1 \le m \le n$ so

$$\begin{aligned} P^n(x,a) &= \sum_{m=1}^{n} P_x(T_a = m) P^{n-m}(a,a) \\ &= \sum_{m=1}^{n} P_x(T_a = m) \\ &= P_x(T_a \le n). \end{aligned}$$

■

27. Calculations with Transition Functions: Problems.

1. The assumption in the previous section that the state space for the Ehrenfest chain consists of $\mathcal{S} = \{0, 1, \ldots, 2R\}$ is historical rather than a requirement of the model. What is the transition matrix for an Ehrenfest chain in which the assumption is that one starts with d balls and the the state space is $\mathcal{S} = \{0, 1, \ldots, d\}$?

2. Suppose that the Ehrenfest Chain is described as in the previous exercise with $d = 3$ and hence for $\mathcal{S} = \{0, 1, 2, 3\}$.
(a) Find $P_x(T_0 = n)$ for $x \in \mathcal{S}$ and $1 \leq n \leq 3$.
(b) Find P, P^2 and P^3.
(b) Suppose that X_0 has the uniform distribution

$$\pi_0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right).$$

Find π_1, π_2 and π_3.

3. Prove 27.2(b).

4. Prove 27.2(c).

5. Prove 27.2(d).

6. Suppose that a Markov Process has state space $\mathcal{S} = \{0, 1, 2\}$ and transition matrix

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1-p & 0 & p \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

(a) Find P^2.
(b) Show that $P^4 = P^2$.
(c) Find P^n for all n.

28.1. Definition.

Let $x, y \in \mathcal{S}$ be states. Recalling that the **hitting time** *for x is*

$$T_y = \min\{n > 0 \ : \ X_n = y\}$$

We set

$$\rho_{xy} = P_x(T_y < \infty).$$

A state y is **recurrent** *if $\rho_{yy} = 1$ and* **transient** *if $\rho_{yy} < 1$.*

Thus ρ_{xy} is the probability that the chain can ever get to state y given that it starts in state x. A state is *recurrent* if, given that the chain starts in state y, there is a 100% chance that it returns to y and *transient* if there is a positive chance that it does not return to y. Clearly an abosorbing state is recurrent, but not conversely.

Clearly the state space $\mathcal{S}$ can be divided into two disjoint sets: the transient states

$$\mathcal{S}_T = \{x \in \mathcal{S} \ : \ \rho_{xx} < 1\}$$

and the recurrent states

$$\mathcal{S}_R = \{x \in \mathcal{S} \ : \ \rho_{xx} = 1\}.$$

In this section we will show that if $x \in \mathcal{S}_T$ then the chain X_n visits x at most a finite number of times with probability one, while if $x \in \mathcal{S}_R$ then then the chain X_n starting in state x returns to x an infinite number of times with probability one. Thus, in studying the long-term behavior of the chain, only the recurrent states matter. In addition we will show that

$$\mathcal{S}_R = \bigcup C_i$$

where the sets C_i are disjoint and have the additional property that, once $X_n \in C_i$ for some i then X_n remains in C_i thereafter with probability one.

Dirac delta function
unit impulse function
unit point mass

28.2. Definition.

Let $y \in \mathcal{S}$ be a state and define the **indicator function** 1_y *for the singleton set* $\{y\}$ *to be*

$$1_y(z) = \begin{cases} 1 & \text{if } z = y \\ 0 & \text{otherwise} \end{cases}$$

This function has various names and notations depending on the context. For example, in physics it is sometimes called the *Dirac delta function*, in engineering the *unit impulse function*, and in mathematics the *characteristic function* of the set $\{y\}$ or the *unit point mass*. For our purposes “indicator function”is as good a name as any.

28.3. Definition.

For a state $y \in \mathcal{S}$ we set

$$N(y) = \sum_{n=1}^{\infty} 1_y(X_n)$$

so that

$$N(y) = \ \# \text{ of times } n \geq 1 \text{ that } X_n = y.$$

The following corollary is obvious from the definitions.

28.4. Corollary.

For any states $x, y \in \mathcal{S}$

$$P_x(N(y) \geq 1) = P_x(T_y < \infty) = \rho_{xy}.$$

28.5. Corollary.

For any states $x, y \in \mathcal{S}$

$$P_x(N(y) \geq 2) = \rho_{xy}\rho_{yy}.$$

Proof. In order for $N(y) \geq 2$ the chain must be able get from x to y, then return from y back to y. Thus if we set

$$A_m = \{X_0 = x, T_y = m\}$$

and

$$B_{m,n} = \{X_m = y \text{ and } T_y = m + n\}$$

then the event

$$X_0 = x \text{ and } N(y) \geq 2$$

is exactly the event

$$\cup_{m=1}^{\infty} \cup_{n=1}^{\infty} (A_m \cap B_{m,n}).$$

By the Markov Property, for each fixed n and m the sets A_m and $B_{m,n}$ satisfy

$$\mathfrak{Pr}\left(A_m \cap B_{m,n}\right) = \mathfrak{Pr}\left(A_m\right) \mathfrak{Pr}\left(B_{m,n}\right).$$

Further for each fixed m and n the sets $A_m \cap B_{m,n}$ are disjoint. Thus

$$P_x(N(y) \geq 2) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \mathfrak{Pr}\left(A_m\right) \mathfrak{Pr}\left(B_{m,n}\right).$$

Finally, stationary transition probabilities imply

$$\mathfrak{Pr}\left(B_{m,n}\right) = \mathfrak{Pr}\left(X_0 = y \text{ and } T_y = n\right).$$

Thus

$$\begin{aligned}
P_x(N(y) \geq 2) &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \mathfrak{Pr}\left(A_m\right) \mathfrak{Pr}\left(B_{m,n}\right) \\
&= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} P_x(T_y = m) P_y(T_y = n) \\
&= \sum_{n=1}^{\infty} P_y(T_y = n) \Big(\sum_{m=1}^{\infty} P_x(T_y = m) \Big) \\
&= \Big(\sum_{m=1}^{\infty} P_x(T_y = m) \Big) \Big(\sum_{n=1}^{\infty} P_y(T_y = n) \Big) \\
&= \rho_{xy} \rho_{yy}.
\end{aligned}$$

■

Taking $x = y$, a simple induction verifies

28.6. Corollary.

For any state $y \in \mathcal{S}$ and any $m \geq 2$

$$P_y(N(y) \geq m) = \rho_{yy}^m.$$

Another simple induction verifies

28.7. Corollary.

For any states $x, y \in \mathcal{S}$ and any $m \geq 2$

$$P_x(N(y) \geq m) = \rho_{xy}\rho_{yy}^{m-1}.$$

28.8. Corollary.

For any states $x, y \in \mathcal{S}$

$$P_x(N(y) = m) = \rho_{xy}\rho_{yy}^{m-1}(1 - \rho_{yy}).$$

Proof. Note that

$$P_x(N(y) = 0) = 1 - P_x(N(y) \geq 1) = 1 - \rho_{xy}$$

and hence

$$P_y(N(y) = 0) = 1 - \rho_{yy}.$$

Using arguments similar to 28.5, the result follows. ∎

28.9. Definition.

For a state $x \in \mathcal{S}$ we define the **conditional expectation** E_x *of an event E to be the expected value of the conditional random variable*

$$E\big|_{X_0=x}.$$

For example,

$$E_x(1_y(X_n)) = P_x(X_n = y) = P^n(x,y).$$

28.10. Corollary.

For any states $x, y \in \mathcal{S}$

$$E_x(N(y)) = \sum_n P^n(x,y).$$

Proof. For any states $x, y \in \mathcal{S}$

$$\begin{aligned} E_x(N(y)) &= E_x\left(\sum_n 1_y(X_n)\right) \\ &= \sum_n E_x\left(1_y(X_n)\right) \\ &= \sum_n P^n(x,y). \end{aligned}$$

∎

28.11. Definition.

For any states $x, y \in \mathcal{S}$ set

$$\mathfrak{G}\,(\mathbf{x},\mathbf{y}) = \mathbf{E_x(N(y))}.$$

Note that in view of the preceding corollary

$$\mathfrak{G}\,(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{n}} \mathbf{P}^{\mathbf{n}}(\mathbf{x},\mathbf{y}).$$

28.12. Theorem.

(a) If $y \in \mathcal{S}$ is a transient state then for any state $x \in \mathcal{S}$

$$P_x(N(y) < \infty) = 1$$

and

$$\mathfrak{G}\,(\mathbf{x},\mathbf{y}) = \frac{\rho_{\mathbf{xy}}}{\mathbf{1} - \rho_{\mathbf{yy}}}.$$

(b) If $y \in \mathcal{S}$ is a recurrent state then

$$P_y(N(y) = \infty) = 1$$

and

$$\mathfrak{G}\,(\mathbf{y},\mathbf{y}) = \infty.$$

Further for any state $x \in \mathcal{S}$

$$P_x(N(y) = \infty) = P_x(T_y < \infty) = \rho_{xy}$$

In particular, if $\rho_{xy} = 0$ then $\mathfrak{G}\,(\mathbf{x},\mathbf{y}) = \mathbf{0}$ while if $\rho_{xy} > 0$ then $\mathfrak{G}\,(\mathbf{x},\mathbf{y}) = \infty$.

Proof. For part (a), if y is transient, then

$$0 \le \rho_{yy} < 1$$

and hence

$$\begin{aligned} P_x(N(y) = \infty) &= \lim_{m\to\infty} P_x(N(y) \ge m) \\ &= \lim_{m\to\infty} \rho_{xy}\rho_{yy}^{m-1} \\ &= 0 \end{aligned}$$

Further

$$\begin{aligned}\mathfrak{G}\,(\mathbf{x},\mathbf{y}) &= E_x(N(y)) \\ &= \sum_{m=1}^{\infty} mP_x\big(N(y)=m)\big) \\ &= \sum_{m=1}^{\infty} m\rho_{xy}\rho_{yy}^{m-1}(1-\rho_{yy}) \\ &= \frac{\rho_{xy}}{1-\rho_{yy}}\end{aligned}$$

using the formula

$$\sum_{m=1}^{\infty} mt^{m-1} = \frac{1}{(1-t)^2}$$

for $-1 < t < 1$. This completes the proof of (a).

For part (b), we assume that y is recurrent so $\rho_{yy} = 1$ and hence

$$\begin{aligned}P_x\big(N(y)=\infty\big) &= \lim_{m\to\infty} P_x(N(y) \geq m) \\ &= \lim_{m\to\infty} \rho_{xy}\rho_{yy}^{m-1} \\ &= \rho_{xy}.\end{aligned}$$

In particular, taking $x = y$,

$$P_y(N(y)=\infty) = 1.$$

Since the condtional random variable

$$N(y)\big|_{X_0=y}$$

assumes the value infinity with positive probability it follows that

$$\mathfrak{G}\,(\mathbf{y},\mathbf{y}) = \mathbf{E}_{\mathrm{y}}(\mathbf{N}(\mathbf{y})) = \infty.$$

Now if $\rho_{xy} = 0$ then for each m

$$P_x(T_y = m) = 0$$

which in turn implies that

$$P^n(x,y) = 0$$

and hence that $\mathfrak{G}\,(\mathbf{x},\mathbf{y}) = \mathbf{0}$.

On the other hand, if $\rho_{xy} > \mathbf{0}$, then

$$P_x(N(y) = \infty) = \rho_{xy} > \mathbf{0}$$

which again implies that

$$\mathfrak{G}\,(\mathbf{x},\mathbf{y}) = \mathbf{E}_\mathrm{x}(\mathbf{x},\mathbf{y}) = \infty$$

completing the proof of (b).

■

28.13. Definition.

A Markov chain is said to be **recurrent** *if every state is recurrent, and is said to be* **transient** *if every state is transient.*

28.14. Corollary.

If the state space $\mathcal{S}$ is finite then there must be at least one recurrent state.

Proof. Suppose for contradiction that every state is transient. Then $\mathfrak{G}\,(\mathbf{x},\mathbf{y}) < \infty$ and

$$\mathfrak{G}\,(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{n}} \mathbf{P}^\mathrm{n}(\mathbf{x},\mathbf{y})$$

implies that

$$\lim_n P^n(x,y) = \mathbf{0}.$$

From this,

$$\begin{aligned} 0 &= \sum_{y\in\mathcal{S}} \lim_n P^n(x,y) \\ &= \lim_n \sum_{y\in\mathcal{S}} P^n(x,y) \\ &= \lim_n P_x(y \in \mathcal{S}) \\ &= 1 \end{aligned}$$

a contradiction.

■

28.15. Definition.

Let $x, y \in \mathcal{S}$ and suppose $\rho_{xy} > 0$. Then we say that x **leads to** *y and we write*

$$x \longrightarrow y.$$

The following corollary is immediate from the definitions and the foregoing.

28.16. Corollary.

Let $x, y, z \in \mathcal{S}$. Then $x \longrightarrow y$ if and only if

$$P^n(x, y) > 0$$

for some $n > 0$. Further if $x \longrightarrow y$ and $y \longrightarrow z$, then $x \longrightarrow z$.

28.17. Theorem.

Let $x \in \mathcal{S}$ be a recurrent state and suppose that $x \longrightarrow y$. Then y is recurrent and $\rho_{xy} = \rho_{yx} = 1$.

Proof. Without loss of generality we may assume $x \neq y$. Since

$$0 < \rho_{xy} = P_x(T_y < \infty)$$

it follows that there is an n for which

$$P_x(T_y = n) > 0.$$

Let n_0 be the least such n, i.e.,

$$n_0 = \min\{n \ : \ P_x(T_y = n) > 0\}.$$

Thus $P^{n_0}(x, y) > 0$ and $P^m(x, y) = 0$ if $1 \leq m < n_0$.

This means that we can find states

$$y_1, y_2, \cdots, y_{n_0 - 1}$$

each of which are different from both x and y for which

$$\begin{aligned} P_x(X_1 = y_1, &X_2 = y_2, \cdots, X_{n_0-1} = y_{n_0-1}, X_n = y) \\ &= P(x, y_1)P(y_1, y_2)\cdots P(y_{n_0-1}, y) \\ &> 0 \end{aligned}$$

Now suppose for contradiction that $\rho_{yx} < 1$. Then there is a positive probabilty

$$1 - \rho_{yx}$$

that a chain starting at y never assumes the value x. Thus a chain starting at x has positive probability

$$P(x, y_1)P(y_1, y_2)\cdots P(y_{n_0-1}, y)(1 - \rho_{yx})$$

of never returning to x. However, this contradicts x being recurrent, hence $\rho_{yx} = 1$.

Since $\rho_{yx} = 1$, we can select n_1 so that

$$P^{n_1}(y, x) > 0.$$

Then for any n

$$\begin{aligned} P^{n_0+n+n_1}(y, y) &= P_y(X_{n_0+n+n_1} = y) \\ &\geq P_y(X_{n_1} = x, X_{n_1+n} = x, X_{n_1+n+n_0} = y) \\ &= P^{n_1}(y, x)P^n(x, x)P^{n_0}(x, y). \end{aligned}$$

This implies

$$\begin{aligned} \mathfrak{G}\,(\mathbf{y}, \mathbf{y}) &\geq \sum_{n=n_1+1+n_0}^{\infty} P^n(y, y) \\ &= \sum_{n=1}^{\infty} P^{n_1+n+n_0}(y, y) \\ &\geq P^{n_1}(y, x)P^{n_0}(x, y)\sum_{n=1}^{\infty} P^n(x, x) \\ &= \infty \end{aligned}$$

since x is recurrent. Now if y were transient this would contradict 28.2(a), hence y must be recurrent.

We have now concluded that y must be recurrent and that $y \to x$. Applying the above arguments with x and y inverted then implies that $\rho_{xy} = 1$.

∎

28.18. Definition.

A collection of states $C \subset \mathcal{S}$ is **closed** *if, whenever $x \in C$ and $y \notin C$ then $\rho_{xy} = 0$. Alternatively, C is closed if $x \in C$ and $x \rightarrow y$ implies $y \in C$.*

28.19. Definition.

A collection of states $C \subset \mathcal{S}$ is **irreducible** *if, whenever $x, y \in C$ then $x \rightarrow y$.*

28.20. Proposition.

Let $C \subset \mathcal{S}$ be a closed, irreducible collection of recurrent states. Then for all $x, y \in C$
(a) $\rho_{xy} = 1$;
(b) $P_x(N(y) = \infty) = 1$; and
(c) $\mathfrak{G}\,(\mathbf{x}, \mathbf{y}) = \infty$.

28.21. Proposition.

Let $C \subset \mathcal{S}$ be a closed, irreducible and finte collection of states. Then every state in C is recurrent.

28.22. Example.

Suppose that a Markov Chain has transition matrix

$$\begin{array}{c|cccccc} & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{0} & 1 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 \\ \mathbf{2} & 0 & \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & 0 & \frac{1}{5} \\ \mathbf{3} & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \\ \mathbf{4} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \mathbf{5} & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

Find the recurrent and transient states and decompose the recurrent states into closed, irreducible sets.

Solution. We can analyze the chain by constructing a matrix with a “$+$”in the (x, y) position if $x \longrightarrow y$ and a zero otherwise. The resulting matrix is

$$\begin{array}{c|cccccc} & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{0} & + & 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & + & + & + & + & + & + \\ \mathbf{2} & + & + & + & + & + & + \\ \mathbf{3} & 0 & 0 & 0 & + & + & + \\ \mathbf{4} & 0 & 0 & 0 & + & + & + \\ \mathbf{5} & 0 & 0 & 0 & + & + & + \end{array}$$

From this, we see that $\{0\}$ is absorbing and hence also closed and irreducible. The states $\{1, 2\}$ are exactly the transient states since they both lead to 0. Finally, $\{3, 4, 5\}$ is another closed, irreducible collection of recurrent states.

■

This illustrates the final theorem of this section.

28.23. Theorem.

Suppose that the collection of recurrent states $\mathcal{S}_R$ is non-empty. Then there is a collection $\{C_i\}$ of disjoint, closed, irreducible sets such that

$$\mathcal{S}_R = \cup_i C_i.$$

Proof. Let $x \in \mathcal{S}_R$ and set

$$C_x = \{y \in \mathcal{S}_R \ : \ x \to y\}.$$

Since x is recurrent, $\rho_{xx} = 1$ and so $x \in C_x$.

We first will show that C_x is irreducible and closed. Suppose that $y \in C_x$ and that $y \to z$. Since $y \in \mathcal{S}_R$ it follows that $z \in \mathcal{S}_R$. Also, $x \to y$ and $y \to z$ implies $x \to z$, hence $z \in C_x$. Thus C_x is closed.

To see that C_x is irreducible, suppose that $y, z \in C_x$. Since x is recurrent and $x \to y$ it follows that $y \to x$. Since $z \in C_x$, we know that $x \to z$. Thus, $y \to x$ and $x \to z$, from which we may conclude that $y \to z$. Thus C_x is irreducible.

Thus for each $x \in \mathcal{S}_R$ we can find a closed, irreducible set C_x containing x. Since $\mathcal{S}$ is countable, there are at most countably many such sets C_x.

To complete the proof, suppose that $x, y \in \mathcal{S}_R$ and that C_x and C_y are two closed, irreducible sets constructed as above. It will suffice to show that either $C_x = C_y$ or $C_x \cap C_y = \phi$.

Suppose that $C_x \cap C_y \neq \phi$, so we can select $u \in C_x \cap C_y$. If v is any element in C_x then $u \to v$ since C_x is irreducible. Since C_y is closed and $u \in C_y$, it follows that $v \in C_y$. Since $v \in C_x$ was arbitrary, this shows that $C_x \subseteq C_y$. The reverse containment is deduced in exactly the same manner.

∎

28. Transient and Recurrent States: Problems.

1. Show that $\rho_{xy} > 0$ if and only if $P^n(x, y) > 0$ for some integer n.

2. Show that if $x \to y$ and $y \to z$ then $x \to z$.

3. Consider the Markov Chain on

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{8} & \frac{7}{8} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{1}{5} & 0 & \frac{1}{5} & \frac{1}{5} & \frac{2}{5} \end{pmatrix} \end{array}$$

(a) Detetermine which states are transient and which are recurrent.

(b) Find $\rho_{\{0,1\}}(x)$ for $x = 0, 1, \cdots, 5$.

29. Absorption Probabilities

The last theorem in the previous section decomposes the state space into the transient states and a collection of closed, irreducible sets of recurrent states

$$\mathcal{S} = (\cup_i C_i) \bigcup \mathcal{S}_T.$$

Since the transient states can only be visited finitely many times, if the state space itself is finite then a chain starting in a transient state must eventually enter one of the closed irreducible sets C_i. Since each C_i is closed, once in C_i the chain can never leave. Hence such a chain starting in a transient state is eventually absorbed into one of the sets C_i. The current section examines various settings in which it is possible to calculate the probability of such absorption.

29.1. Definition.

Let $C \subseteq \mathcal{S}$ be a closed, irreducible set of recurrent states. For $x \in \mathcal{S}$ we define the **absorption probability** *of C relative to x as*

$$\rho_C(x) = P_x(T_C < \infty).$$

29.2. Theorem.

Suppose that $C \subseteq \mathcal{S}$ be a closed, irreducible set of recurrent states and that the collection of transient states $\mathcal{S}_T$ is finite. Then the system of equations

$$f(x) = \sum_{y \in C} P(x,y) + \sum_{y \in \mathcal{S}_T} P(x,y) f(y) \tag{29.1}$$

has the unique solution

$$f(x) = \rho_C(x)$$

for $x \in \mathcal{S}_T$.

29.3. Example.

Suppose that a Markov Chain $\{X_n\}$ has transition matrix

$$\begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & 4 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 3 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 4 & 0 & 0 & \frac{3}{5} & 0 & \frac{2}{5} \end{array}$$

(a) *Find the transient, recurrent and absorbing states.*
(b) *For each transient state $x \in \mathcal{S}_T$ find $\rho_1(x)$.*

Solution. Clearly $\{1\}$ is an absorbing state. To find the transient states, we proceed as before an place a "$+$"in the (x, y) position if $x \longrightarrow y$ and a 0 otherwise. The resulting matrix is

$$\begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & 4 \\ 0 & + & + & + & + & + \\ 1 & 0 & + & 0 & 0 & 0 \\ 2 & 0 & 0 & + & 0 & + \\ 3 & 0 & + & + & + & + \\ 4 & 0 & 0 & + & 0 & + \end{array}$$

Thus $\mathcal{S}_T = \{0, 3\}$ and

$$\mathcal{S}_C = \{1\} \cup \{2, 4\} \tag{29.2}$$

where (29.2) is the decomposition of $\mathcal{S}_R$ into closed, irreducible, disjoint sets of recurrent states guaranteed by Theorem 28.23.

Now by Theorem 29.2,

$$f(0) = P(0,1) + P(0,0)f(0) + P(0,3)f(3)$$
$$f(3) = P(3,1) + P(3,0)f(0) + P(3,3)f(3)$$

or equivalently

$$f(0) = 0 + \frac{1}{2}f(0) + \frac{1}{2}f(3)$$
$$f(3) = \frac{1}{4} + 0 + \frac{1}{4}f(3).$$

From this $f(3) = 1/3$ and $\rho_1(3) = 1/3$. Similarly $\rho_1(0) = 1/3$.

Note that we could apply the Theorem to calculate $\rho_{\{2,4\}}(x)$. However clearly

$$\rho_{\{2,4\}}(x) + \rho_1(x) = 1$$

for each transient state $x \in \mathcal{S}_T$ since a transient state must be absorbed into one of the two closed irreducible sets $\{2,4\}$ and $\{1\}$. Thus

$$\rho_{\{2,4\}}(0) = \rho_{\{2,4\}}(3) = \frac{2}{3}.$$

∎

We now prove Theorem 29.2.

Proof. If (29.2) holds, then for each $y \in \mathcal{S}_T$

$$f(y) = \sum_{z \in C} P(y,z) + \sum_{z \in \mathcal{S}_T} P(y,z) f(z)$$

If one substitutes the above into (29.2), then

$$\begin{aligned} f(x) &= \sum_{y \in C} P(x,y) + \sum_{y \in \mathcal{S}_T} P(x,y) f(y) \\ &= \sum_{y \in C} P(x,y) + \sum_{y \in \mathcal{S}_T} P(x,y) \left(\sum_{z \in C} P(y,z) + \sum_{z \in \mathcal{S}_T} P(y,z) f(z) \right) \\ &= \sum_{y \in C} P(x,y) + \sum_{y \in \mathcal{S}_T} \sum_{z \in C} P(x,y)P(y,z) + \sum_{y \in \mathcal{S}_T} \sum_{z \in \mathcal{S}_T} P(x,y)P(y,z) f(z) \end{aligned}$$

Now notice that

$$\sum_{y \in C} P(x,y) = P_x(T_C = 1)$$

while

$$\begin{aligned} \sum_{y \in \mathcal{S}_T} \sum_{z \in C} P(x,y)P(y,z) &= \sum_{z \in C} \sum_{y \in \mathcal{S}_T} P(x,y)P(y,z) \\ &= P_x(T_C = 2) \end{aligned}.$$

For the last inequality, note that one can go from a transient state x to a reccurent state $z \in C$ in exactly two steps only by first passing through a transient state. Thus

$$\begin{aligned} f(y) &= P_x(T_C = 1) + P_x(T_C = 2) + \sum_{z \in \mathcal{S}_T} \sum_{y \in \mathcal{S}_T} P(x,y)P(y,z)f(z) \\ &= P_x(T_C \leq 2) + \sum_{z \in \mathcal{S}_T} \sum_{y \in \mathcal{S}_T} P(x,y)P(y,z)f(z) \\ &= P_x(T_C \leq 2) + \sum_{z \in \mathcal{S}_T} P^2(x,z)f(z). \end{aligned}$$

For the last equality, note that the only way one can get from a state x to a transient state z is by passing through another transient state, for a recurrent state can only lead to another recurrent state. Thus

$$\sum_{y \in \mathcal{S}_T} P(x,y)P(y,z)f(z) = P^2(x,z)f(z).$$

From this

$$f(x) = P_x(T_C \leq 2) + \sum_{y \in \mathcal{S}_T} P^2(x,y)f(y)$$

from which an easy induction yields

$$f(x) = P_x(T_C \leq n) + \sum_{y \in \mathcal{S}_T} P^n(x,y)f(y).$$

Now, letting $n \to \infty$,

$$\begin{aligned} f(x) &= P_x(T_C < \infty) + \lim_{n \to \infty} \sum_{y \in \mathcal{S}_T} P^n(x,y)f(y) \\ &= P_x(T_C < \infty) + \sum_{y \in \mathcal{S}_T} \lim_{n \to \infty} P^n(x,y)f(y) \qquad (29.3.) \end{aligned}$$

where the interchange of the limit and the sum is justified by the fact that $\mathcal{S}_T$ is assumed to be finite. Since y is transient, we know that

$$\mathfrak{G}\,(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{i}=1}^{\infty} \mathbf{P}^{\mathbf{n}}(\mathbf{x},\mathbf{y}) < \infty$$

and hence that

$$\lim_{n \to \infty} P^n(x,y) = 0.$$

The result now follows immediately from (29.3).

∎

martingales

We next turn to the topic of martingales. Historically, the study of martingales arose in connection with the study of fair gambling games. The idea was that a game would be "fair" if the expected accumulated winnings of the gambler did not change from play to play. Mathematically, a martingale is defined to be a stochastic process $\{X_n\}$ in which

$$E(X_{n+1} \big| X_0 = x_0, X_1 = x_1, \cdots, X_n = x_n) = x_n \tag{29.4}$$

where X_n represents the accumulated winnings of the gambler after playing the n^{th} game. Clearly (29.4) is not sufficient to guarantee that $\{X_n\}$ is a Markov Chain, nor does every Markov Chain satisfy (29.4). There is, however, a simple condition which assures that a Markov Chain is a martingale.

29.4. Proposition.

Let $\{X_n\}$ be a Markov chain defined on a finite state space $\mathcal{S} = \{0, 1, \cdots d\}$. Then (29.4) holds if

$$\sum_{y=0}^{d} yP(x, y) = x \tag{29.5}$$

for each $x \in \mathcal{S}$.

Proof. This is immediate from the definition of expectation and the Markov Property. ■

29.5. Proposition.

Let $\{X_n\}$ be a Markov chain for which (29.5) holds. Then the states 0 and d are absorbing and the states $\{1, 2, \cdots, d-1\}$ are transient.

Proof. Taking $x = 0$ we see that

$$\sum_{y=0}^{d} yP(0, y) = 0$$

from which it follows that

$$P(0, 1) = P(0, 2) = \cdots = P(0, d) = 0$$

which in turn implies that $P(0,0) = 1$, hence 1 is absorbing.

To see that d is also absorbing,

$$\begin{aligned} d &= \sum_{y=0}^{d} yP(d,y) \\ &= \sum_{y=1}^{d} yP(d,y) \\ &= \sum_{y=1}^{d-1} yP(d,y) + dP(d,d) \\ &= \sum_{y=1}^{d-1} yP(d,y) + d\left(1 - \sum_{y=0}^{d-1} P(d,y)\right) \end{aligned}$$

which implies

$$0 = \left(\sum_{y=1}^{d-1} (y-d)P(d,y)\right) - P(d,0)$$

and hence that $P(d,y) = 0$ for $d = 0, 1, \cdots, d-1$.

We leave to the exercises the proof that the states $\{0, 1, \cdots, d-1\}$ are transient. ∎

29.6. Theorem.

Let $\{X_n\}$ be a Markov chain for which (29.5) holds. Then for every transient state $x \in \{1, 2, \cdots, d-1\}$

$$\rho_{x,d} = \frac{x}{d} \text{ and } \rho_{x,0} = 1 - \frac{x}{d}.$$

Proof. Note that

$$\begin{aligned} E_x(X_n) &= \sum_{y=0}^{d} yP_x(X_n = y) \\ &= \sum_{y=0}^{d} yP^n(x,y) \\ &= \sum_{y=1}^{d-1} yP^n(x,y) + dP^n(x,d) \\ &= \sum_{y=1}^{d-1} yP^n(x,y) + dP_x(T_d \leq n). \end{aligned}$$

Since the states $y = 1, 2, \cdots, d-1$ are transient,

$$\lim_{n\to\infty} P^n(x,y) = 0$$

and hence

$$\lim_{n\to\infty} dP_x(T_d < n) = \lim_{n\to\infty} E_x(X_n) = d\rho_{xd}.$$

On the other hand (see the exercises)

$$E(X_n) = E(X_{n-1}) = \cdots = E(X_0)$$

from which $E_x(X_n) = x$ and the conclusions follow.

∎

Next we turn to absorption probabilities for the Birth and Death Chain. We begin by recalling the basic definitions. Recall that for a birth and death chain there are non-negative numbers q_x, r_x and p_x with the properties that $q_x + r_x + p_x = 1$ and

$$P(x,y) = \begin{cases} q_x & \text{if } y = x-1 \\ r_x & \text{if } y = x \\ p_x & \text{if } y = x+1 \end{cases}$$

In addition $q_0 = 0$ and, if $\mathcal{S} = \{0, \cdots, d\}$ is finite, then $p_d = 0$.

29.7. Theorem.

Let $\{X_n\}$ be a Birth and Death Chain and suppose that $0 < q_x$ for $x > 0$ and that $0 < p_x$ for all x if $\mathcal{S} = \{0, 1, \cdots\}$ or, if $\mathcal{S} = \{0, 1, \cdots, d\}$ is finite, then $0 < p_x$ for $x = 0, 1, \cdots, d-1$. Define γ_y by $\gamma_0 = 1$ and, for $y > 0$,

$$\gamma_y = \frac{q_1 \cdots q_y}{p_1 \cdots p_y}.$$

Suppose that a, b and x are states with $a < x < b$. Then

$$P_x(T_a < T_b) = \frac{\sum_{y=x}^{b-1} \gamma_y}{\sum_{y=a}^{b-1} \gamma_y}$$

while

$$P_x(T_b < T_a) = \frac{\sum_{y=a}^{x-1} \gamma_y}{\sum_{y=a}^{b-1} \gamma_y}.$$

Proof. Fix states a and b with $a < b$ and, for $a \le x \le b$ define

$$g(x) = \begin{cases} 1 & \text{if } x = a \\ P_x(T_a < T_b) & \text{if } a < x < b \\ 0 & \text{if } x = b \end{cases}$$

Note for $a < y < b$ the chain can from state y in one step only to $y-1$, y or $y+1$. Thus

$$g(y) = q_y g(y-1) + r_y g(y) + p_y g(y+1).$$

Since $r_y = 1 - q_y - p_y$ it follows that

$$\begin{aligned} g(y) &= q_y g(y-1) + (1 - q_y - p_y) g(y) + p_y g(y+1) \\ &= g(y) + q_y (g(y-1) - g(y)) + p_y (g(y+1) - g(y)) \end{aligned}$$

which implies

$$g(y+1) - g(y) = \frac{q_y}{p_y}(g(y) - g(y-1)).$$

Re-writing this gives

$$g(y+1)-g(y)=\frac{\gamma_y}{\gamma_{y-1}}(g(y)-g(y-1)).$$

An easy induction now gives

$$\begin{aligned} g(y+1)-g(y) &= \frac{\gamma_{a+1}}{\gamma_a}\cdots\frac{\gamma_y}{\gamma_{y-1}}(g(a+1)-g(a)) \\ &= \frac{\gamma_y}{\gamma_a}(g(a+1)-g(a)) \qquad \textbf{(29.6.)} \end{aligned}$$

or

$$g(y)-g(y+1)=\frac{\gamma_y}{\gamma_a}(g(a)-g(a+1)).$$

Now sum from $y=a$ to $y=b-1$ and use the fact that $g(a)=1$ and $g(b)=0$ and that the left-hand-side telescopes to obtain

$$1=(g(a)-g(a+1))\sum_{y=a}^{b-1}\frac{\gamma_y}{\gamma_a}.$$

this implies

$$\frac{g(a)-g(a+1)}{\gamma_a}=\frac{1}{\sum_{y=a}^{b-1}\gamma_y}.$$

Now substitute this into (29.6)

$$g(y)-g(y+1)=\frac{\gamma_y}{\sum_{u=a}^{b-1}\gamma_u}.$$

Finally summing the above from $y=x$ to $y=b-1$ gives

$$g(x)=\frac{\sum_{y=x}^{b-1}\gamma_y}{\sum_{y=a}^{b-1}\gamma_y}$$

which is the desired formula for $g(x)=P_x(T_a<T_b)$. The formula for $P_x(T_b<T_a)$ follows from calculating $1-g(x)$.

∎

29.8. Example.

A gambler plays roulette with a sequence of \$1 bets. The probability of winning \$1 is $9/19$ *and the probability of losing \$1 is* $10/19$. *Suppose that the gamblers begins with \$10 and decides to quit as soon as his net winnings are \$25 or he goes broke.*

(a) Find the probability that when the gambler quits he has won \$25.

(b) Find the expected winnings – or loss.

Solution. Let X_n be the accumulated capital after n bets, so $X_0 = 10$. The gambler will quit when either $X_n = 0$ or $X_n = 35$. This is then a birth and death chain on $\mathcal{S} = \{0, \cdots, 35\}$ with

$$p_x = \frac{9}{19}$$

and

$$q_x = \frac{10}{19}$$

for $0 < x < 35$, with 0 and 35 being absorbing states.

In the context of the preceding theorem with $a = 0$, $x = 10$ and $b = 35$ means that we seek

$$P_{10}(T_{35} < T_0).$$

Note that

$$\gamma_y = \left(\frac{10}{9}\right)^y$$

for $0 \leq y \leq 34$. From this

$$\begin{aligned} P_{10}(T_{35} < T_0) &= \frac{\sum_{y=0}^{9} \left(\frac{10}{9}\right)^y}{\sum_{y=0}^{34} \left(\frac{10}{9}\right)^y} \\ &= \frac{(10/9)^{10} - 1}{(10/9)^{35} - 1} \\ &= 0.047 \end{aligned}$$

completing (a). For (b), the expected capital is

$$0 \cdot (0.953) + 35 \cdot (0.047) = 35 \cdot (0.047).$$

Since the gambler started with \$10, the expected loss is

$$10 - 35 \cdot (0.047) = 8.36.$$

∎

For an irreducible chain either every state is recurrent or every state is transient. If an irreducible chain has a finite number of elements in the state space, then it is necessarily recurrent. In general it is difficult to ascertain whether or not an irreducible chain with an infinite number of states is recurrent or transient; however the above lets us do so for birth and death chains.

29.9. Theorem.

Suppose that $\{X_n\}$ is an irreducible birth and death chain on $\mathcal{S} = \{0, 1, \cdots\}$. Then $\{X_n\}$ is recurrent if and only if

$$\sum_{x=1}^{\infty} \gamma_x = \infty.$$

Proof. As a special case of the preceding theorem,

$$P_1(T_0 < T_n) = 1 - \frac{1}{\sum_{y=0}^{n-1} \gamma_y}. \tag{29.7}$$

Now a birth and death chain starting at $x = 1$ can move to the right at most one step at a time, thus

$$1 \leq T_1 \leq T_2 \leq \cdots \leq T_n.$$

From this, the events $\{T_0 < T_n\}$ constitute a non-decreasing sequence, from which

$$\lim_{n\to\infty} P_1(T_1 < T_n) = P_1(T_1 < T_m \text{ for some } m > 0)$$

via Theorem 5.5. Further, it must be the case that $T_m \geq m$, hence $T_m \to \infty$ as $m \to \infty$ and thus the event

$$\{T_1 < T_m \text{ for some } m > 0\}$$

occurs if and only if the event

$$\{T_0 < \infty\}$$

occurs. From this

$$\lim_{n\to\infty} P_1(T_1 < T_n) = P_1(T_0 < \infty).$$

Thus (29.7) implies

$$P_1(T_0 < \infty) = 1 - \frac{1}{\sum_{y=0}^{\infty} \gamma_y}. \tag{29.8}$$

Now if $\{X_n\}$ is recurrent, then $P_1(T_0 < \infty) = 1$ and hence

$$\sum_{y=0}^{\infty} \gamma_y = \infty.$$

For the converse,

$$P_0(T_0 < \infty) = P(0,0) + P(0,1)P_1(T_0 < \infty).$$

Now if $\sum_y \gamma_y = \infty$ then $P_1(T_0 < \infty) = 1$ and so

$$P_0(T_0 < \infty) = P(0,0) + P(0,1) = 1$$

so 0 is recurrent. Since we have assumed the chain is irreducible, this implies the chain is recurrent.

∎

29.10. Example.

Suppose that a birth and death chain on $\mathcal{S} = \{0, 1, \cdots\}$ has

$$p_x = \frac{x+2}{2(x+1)} \text{ and } q_x = \frac{x}{2(x+1)}.$$

Then the chain is transient.

Solution. Since $q_x/p_x = x/(x+2)$, it follows that

$$\begin{aligned}\gamma_x &= \frac{q_1 \cdots q_x}{p_1 \cdots p_x} \\ &= \frac{1 \cdot 2 \cdots x}{3 \cdot 4 \cdots (x+2)} \\ &= \frac{2}{(x+1)(x+2)} \\ &= 2\left(\frac{1}{x+1} - \frac{1}{x+2}\right).\end{aligned}$$

So

$$\begin{aligned}\sum_x \gamma_x &= \sum_x 2\left(\frac{1}{x+1} - \frac{1}{x+2}\right) \\ &= 2\left(\frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \frac{1}{4} \cdots + - \cdots\right) \\ &= 1 \\ &< \infty\end{aligned}$$

and so the chain is transient

∎

29. Absorption Probabilities: Problems.

1. Suppose that $\{X_n\}$ is a Markov chain on $\{0, 1, 2, 3, 4, 5\}$ having transition matrix

$$\begin{array}{c|cccccc} & 0 & 1 & 2 & 3 & 4 & 5 \\ \hline 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 1 & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2} & 0 & 0 \\ 2 & 0 & \frac{1}{6} & \frac{2}{3} & 0 & \frac{1}{6} & 0 \\ 3 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 4 & 0 & 0 & 0 & \frac{1}{5} & \frac{4}{5} & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

(a) Decompose the state space $\mathcal{S}$ into

$$\mathcal{S} = \mathcal{S}_T \cup C_1 \cup C_2$$

where $\mathcal{S}_T$ consists of the transient states and C_1 and C_2 are closed, irreducible, disjoint collections of recurrent states.

(b) Calculate $\rho_5(i)$ for $i = 0, 1, 2, 3, 4, 5$.

2. Consider a Markov Chain with state space $\mathcal{S} = \{0, 1, 2, 3, 4, 5, 6\}$ and transition matrix

$$\begin{array}{c|ccccccc} & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 0 & \frac{1}{2} & 0 & \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 5 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{array}$$

(a) Determine which states are transient and which are recurrent.

(b) Find $\rho_{0,y}$ for $y = 0, 1, \cdots, 6$.

3. Let $\{X_n\}$ be a Markov chain whose state space is a subset of $\{0, 1, \cdots\}$ and with a transition function that satisfies

$$\sum_y yP(x, y) = \alpha x + \beta$$

for all $x \in \mathcal{S}$ and for some constants α and β.

(a) Show that $E(X_{n+1}) = \alpha E(X_n) + \beta$.

(b) If $\alpha \neq 1$ then show that

$$E(X_n) = \frac{\beta}{1+\alpha} + \alpha^n \left(E(X_0) - \frac{\beta}{1-\alpha} \right).$$

4. Let $\{X_n\}$ be the Ehrenfest chain. Use the previous exercise to calculate $E_x(X_n)$.

5.

6. Consider the Gambler's Ruin chain on $\mathcal{S} = \{0, 1, \cdots, d\}$. For $0 < x < d$ find

$$P_x(T_0 < T_d).$$

Let $\{X_n\}$ be a Markov chain on $\mathcal{S} = \{0, \cdots, d\}$ satisfying (29.5). Suppose further that $\{X_n\}$ has no absorbing states other than 0 and d. Show that each of the states $x = 1, 2, \cdots, d-1$ leads to zero and hence must be transient.

7. Consider a birth and death chain on the non-negative integers for which $p_x >$ and $q_x > 0$ for $x \geq 1$.

(a) If $\sum_y \gamma_y = \infty$ then show that $\rho_{x0} = 1$ for $x \geq 1$.

(b) If $\sum_y \gamma_y < \infty$ then show that

$$\rho_{x0} = \frac{\sum_{y=x}^{\infty} \gamma_y}{\sum_{y=0}^{\infty} \gamma_y}.$$

8. Consider the gambler's ruin chain on $\mathcal{S} = \{0, 1, \cdots\}$.

(a) If $q \geq p$ then show that $\rho_{x0} = 1$ for $x \geq 1$.

(b) If $q < p$ then show that $\rho_{x0} = (q/p)^x$ for $x \geq 1$.

9. Consider an irreducible birth and death chain on the non-negative integers and suppose that $p_x \leq q_x$ for $x \geq 1$. Show that the chain is recurrent.

This section we discuss the branching chain introduced in section 26.12. Recall that in the branching chain we have a collection $\{\xi_i\}$ of independent, identically distributed random variables with values contained in the non-negative integers and having common density function f. The branching chain models the evolution of a population through distinct generations, with the random variable ξ_i representing the number of offspring of the i^{th} member of the population surviving to the next generation. Thus the transition matrix is

$$P(x, y) = \mathfrak{Pr}\,(\xi_1 + \cdots + \xi_x = y)$$

for $x, y > 0$ and $P(0, 0) = 1$. If $f(1) = 1$ then the branching chain is degenerate and every state is an absorbing state. Thus in the sequel we will assume that $f(1) < 1$.

If we start with a population of one, it is possible that eventually the population drops to zero, i.e., all of the descendents of the initial member of the population have died off or become extinct. Thus the absorption probabilities to zero have particular meaning in the context of the branching chain. This motivates the next definition.

30.1. Definition.

Let $\{X_n\}$ be a branching chain. The **extinction probability** *ρ for the chain is defined to be*

$$\rho = P_1(T_0 < \infty).$$

30.2. Proposition.

Let $\{X_n\}$ be a branching chain with extinction probability ρ. Then

$$\rho_{x0} = P_x(T_0 < \infty) = \rho^x$$

.

Proof. This follows immediately from the independence of the random variables $\{\xi_i\}$ and the Markov Property.

∎

Our results in this section rely on the probability generating function Φ for a random variable ξ having density function f. Recall that

$$\Phi(t) = E(t^\xi) = f(0) + \sum_{x=1}^{\infty} f(x)t^x.$$

Our main result for this section is the following theorem.

30.3. Theorem.

Let $\{X_n\}$ be a branching chain and let ξ be a random variable having the same distribution as the $\{\xi_i\}$. Let $\mu = E(\xi)$ where we permit $\mu = \infty$ and suppose that $\mathfrak{Pr}\left(\xi = 1\right) < 1$.

(a) If $\mu \le 1$ then $\rho = 1$.

(b) If $\mu > 1$ and if $\Phi(t)$ is the probability generating function for ξ, then the equation

$$\Phi(t) = t$$

has a unique root in $[0, 1)$ and ρ is equal to this unique root.

30.4. Example.

Suppose that the male of a certain species always has exactly three offspring, each of which has exactly equal chances of being male or female. Let $\{X_n\}$ be the number of males in the n^{th} generation and suppose that $\{X_n\}$ is a branching chain. Find the probability that a male line becomes extinct.

Solution. The density function for the number of male offspring is binomial with $n = 3$ and $p = 1/2$ so

$$f(0) = 1/8,\ f(1) = 3/8,\ f(2) = 3/8,\ f(3) = 1/8.$$

Note that $\mu = 3/2$, so ρ is the unique solution to the equation

$$\Phi(t) = t$$

in $[0, 1)$. This equation becomes

$$\frac{1}{8} + \frac{3}{8}t + \frac{3}{8}t^2 + \frac{1}{8}t^3 = t$$

or equivalently

$$t^3 + 3t^2 - 5t + 1 = 0.$$

This factors to

$$(t-1)(t^2 + 4t - 1) = 0$$

and hence the roots are $t = 1$, $t = -\sqrt{5} - 2$ and $t = \sqrt{5} - 2$. Thus $\rho = \sqrt{5} - 2$.

∎

In order to prove our main result about branching chains, we deduce a slightly more general result about probability generating functions.

30.5. Theorem.

Let ξ be random variable having values in the non-negative integers and having density function f. Let $\mu = E(\xi)$ where we permit $\mu = \infty$. If $\mu \leq 1$ and $\mathfrak{Pr}\,(\xi = 1) < 1$ then the equation

$$\Phi(t) = t \tag{30.1}$$

has no solutions in $[0, 1)$. If $\mu > 1$ then (30.1) has a unique solution ρ in $[0, 1)$.

Proof. Recall that

$$\Phi(t) = f(0) + f(1)t + f(2)t^2 + \cdots$$

from which

$$\Phi'(t) = f(1) + 2f(2)t + 3f(3)t^2 + \cdots$$

from which

$$\Phi(0) = f(0),\ \Phi(1) = 1,\ \Phi'(1) = \mu.$$

We first establish that the equation

$$\Phi(t) = t$$

has no solutions in $[0, 1)$ when $\mu \leq 1$. We begin by showing that

$$\Phi'(t) < 1 \text{ for } 0 \leq t < 1. \tag{30.2}$$

We can distinguish two cases.
Case 1. $\mu < 1$.

In this case

$$\lim_{t \uparrow\ 1} \Phi'(t) = \Phi'(1) = \mu < 1.$$

Now for $0 \leq t < 1$, $\Phi''(t) \geq 0$ and so $\Phi'(t)$ is non-decreasing in t. Thus $\Phi'(t) < 1$, establishing (30.2) in case 1.

Case 2. $\mu = 1$.

In this case, there is some $n \geq 2$ with $f(n) > 0$. If this were not the case, then

$$\mu = 1f(1) = \mathfrak{Pr}\left(\xi = 1\right) < 1,$$

a contradiction. This implies that $\Phi'(t) > 0$ for $0 < t$, and thus $\Phi(t)$ is strictly increasing on $[0, 1)$. But since

$$\lim_{t \uparrow\ 1} \Phi'(t) = \mu = 1$$

and hence $\Phi'(t) < 1$ for $0 \leq t < 1$ in case two as well.

Now via (30.2),

$$\frac{d}{dt}\left(\Phi(t) - t\right) < 0$$

for $0 \leq t < 1$, and hence $\Phi(t) - t$ is strictly decreasing on $[0, 1)$. Since $\Phi(1) - 1 = 0$, it follows that $\Phi(t) - t > 0$ on $[0, 1)$ and hence has no roots.

Now suppose that $\mu > 1$; we will show that $\Phi(t) - t$ has exactly one root in $[0, 1)$.

For existence, note that

$$\lim_{t \uparrow\ 1} \Phi'(t) = \mu > 1.$$

This implies that there is some t_0 with $0 < t_0 < 1$ and $\Phi'(t) > 1$ for $t_0 \leq t < 1$. The Mean Value Theorem then implies that

$$\frac{1 - \Phi(t_0)}{1 - t_0} = \frac{\Phi(1) - \Phi(t_0)}{1 - t_0} > 1$$

which in turn implies that

$$\Phi(t_0) - t_0 < 0.$$

Now $\Phi(t)$ is continuous and $\Phi(0) \geq 0$, so there is some t_1 with $0 \leq t_1 \leq t_0$ and

$$\Phi(t_1) - t_1 = 0$$

via the intermediate value theorem. This establishes the existence of a solution in $[0, 1)$.

For uniqueness, suppose that there are two such roots, t_1 and $\tilde{t_1}$, i.e.,

$$\Phi(t_1) - t_1 = \Phi(\tilde{t_1}) - \tilde{t_1} = \Phi(1) - 1 = 0.$$

Thus, from Rolle's Theorem, Φ' has at least two roots in $[0, 1)$, which in turn implies that Φ'' has a root in $[0, 1)$. However, since $\mu > 1$ there is necessarily an $n > 1$ for which $f(n) > 0$. This in turn implies that if $0 < t < 1$ then

$$\Phi''(t) = 2f(2) + 3!f(3)t + \cdots > 0$$

if $0 < t < 1$. Thus Φ'' cannot have a root in $(0, 1)$ and so there cannot be two solutions. ∎

Proof of Theorem 30.3 We begin by verifying that the extinction probability ρ must solve the equation

$$\Phi(\rho) = \rho.$$

To see this,

$$\begin{aligned}\rho &= \rho_{10}\\ &= P(1,0) + \sum_{y=1}^{\infty} P(1,y)\rho_{y,0}\\ &= f(0) + \sum_{y=1}^{\infty} f(y)\rho^{y}\\ &= \Phi(\rho)\end{aligned}$$

If $\mu \leq 1$ then $\Phi(t) = t$ has no solution in $[0,1)$, from which $\rho = 1$.

On the other hand, if $\mu > 1$, then $\Phi(t) = t$ has a root t_1 in $[0,1)$. Of course we also know that $\Phi(1) = 1$. Thus the proof will be complete if we can verify that the extinction probability ρ is t_1 as opposed to 1.

Since the particles act independently,

$$P_x(T_0 \leq n) = \left(P_1(T_0 \leq n)\right)^x.$$

so, for $n \geq 0$,

$$\begin{aligned}P_1(T_0 \leq n+1) &= P(1,0) + \sum_{y=1}^{\infty} P(1,y)P_y(T_0 \leq n)\\ &= P(1,0) + \sum_{y=1}^{\infty} P(1,y)\left(P_1(T_0 \leq n)\right)^y\\ &= f(0) + \sum_{y=1}^{\infty} f(y)\left(P_1(T_0 \leq n)\right)^y\\ &= \Phi\left(P_1(T_0 \leq n)\right)\end{aligned}$$

i.e.

$$P_1(T_0 \leq n+1) = \Phi\left(P_1(T_0 \leq n)\right).$$

We next assert that

$$P_1(T_0 \leq n) \leq t_1$$

for all n. Note that $P_1(T_0 \leq 0) = 0 \leq t_1$. Proceeding by induction, if follows that

$$\begin{aligned} P_1(T_0 \leq n+1) &= \Phi(P_1(T_0 \leq n) \\ &\leq t_1 \end{aligned}$$

via the inductive assumption. Thus for all n

$$P_1(T_0 \leq n) \leq t_1$$

Now letting $n \to \infty$,

$$\begin{aligned} \rho &= P_1(T_0 < \infty) \\ &= \lim_{n\to\infty} P_1(T_0 \leq n) \\ &=\leq t_1 \end{aligned}$$

Since ρ must be either t_1 or 1, this implies that $\rho = t_1$.

∎

Theorem 30.5 turns out to also enable us to determine whether or not the queuing chain is recurrent or transient. While the conclusions are strikingly different from those for the branching chain, the techniques are quite similar.

Recall that in a queuing chain, ξ_n represents the number of customers arriving in the n^{th} time interval, where $\{\xi_i\}$ are independent and identically distributed random variables having common density function f. In any time interval, if the queue is non-empty then exactly one customer will be served. Thus the transition function is

$$P(x, y) = \begin{cases} f(y) & \text{if } x = 0 \\ f(y - x + 1) & \text{if } x \geq 1 \end{cases}$$

Now if the average number of newly arrived customers in each time interval is greater than one, then it makes sense that the queue never empties and, in fact, that $X_n \to \infty$ as $n \to \infty$. On the other hand, if on average fewer than one customer arrives per unit, then it makes sense that the queue eventually empties since one customer is always served if the queue is non-empty. This would imply that state 0 is recurrent, so if the chain is irreducible then the chain is recurrent. The following theorem formalizes these intuitive ideas, and also handles the less intuitive case when the average number of new customers is exactly one.

30.6. Theorem.

Let $\{X_n\}$ be an irreducible queuing chain and let ξ be a random variable having the same distribution as the $\{\xi_i\}$. Let $\mu = E(\xi)$. If $\mu > 1$ then the chain is transient, while if $\mu \leq 1$ then the chain is recurrent.

Proof. First note that

$$P(0, y) = P(1, y)$$

and that

$$\rho_{00} = \rho_{10}.$$

We set $\rho = \rho_{00} = \rho_{10}$. We begin by showing that

$$\Phi(\rho) = \rho. \tag{30.3}$$

If 0 is recurrent, then $\rho = 1$ and (30.3) follows from $\Phi(1) = 1$. Thus without loss we may assume that 0 is not recurrent.

We will first establish that

$$\rho_{x0} = \rho^x. \tag{30.4}$$

Suppose that the queuing chain starts at $y > 0$. Then the event

$$\{T_{y-1} = n\}$$

occurs if and only if

$$\begin{aligned} n &= \min\{m > 0 : \underbrace{y + (\xi_1 - 1) + \cdots + (\xi_m - 1)}_{\text{m time periods}} = y - 1\} \\ &= \min\{m > 0 : \xi_1 + \cdots + \xi_m = m - 1\} \end{aligned}$$

which implies that

$$P_y(T_{y-1} = n)$$

is independent of y. Thus

$$\rho_{y,y-1} = \rho_{y-1,y-2} = \cdots = \rho_{1,0} = \rho.$$

Now since the number in the queue can be reduced by at most one at each time interval, it follows that in order to go from y to 0 customers in the queue we must pass through $y - 1$, $y - 2$, one step at a time, all the way to 0. Thus via independence and the Markov Property

$$\rho_{y0} = \rho_{y,y-1}\rho_{y-1,y-2} \cdots \rho_{1,0} = \rho^y$$

establishing (30.4).

Thus

$$\begin{aligned}\rho_{00} &= P(0,0) + \sum_{x=1}^{\infty} P(0,x)\rho_{x,0} \\ &= f(0) + \sum_{x=1}^{\infty} f(x)\rho_{x,0} \\ &= f(0) + \sum_{x=1}^{\infty} f(x)\rho^{x}\end{aligned}$$

which implies that

$$\Phi(\rho) = \rho.$$

Now suppose that the chain is irreducible and $\mu \leq 1$. Since the chain is irreducible, it must be the case that $f(1) < 1$. Thus the equation $\Phi(t) = t$ has no solutions in $[0, 1)$ and hence $\rho = 1$. From the definition of ρ, this implies that 0 is recurrent. Since the chain is assumed to be irreducible, this implies the chain is recurrent.

On the other hand, suppose that $\mu > 1$. Since we know that the equation $\Phi(t) = t$ has exactly one solution t_1 in $[0, 1)$ in this case (recall that in the case $\mu > 1$ we did not hypothesize that $f(1) < 1$). We will show that $\rho = t_1$.

Similar to the proof for the branching chain,

$$\begin{aligned}P_1(T_0 \leq n+1) &= P(1,0) + \sum_{y=1}^{\infty} P(1,y)P_y(T_y \leq n) \\ &= f(0) + \sum_{y=1}^{\infty} f(y)P_y(T_y \leq n)\end{aligned}$$

We next argue that

$$P_y(T_0 \leq n) \leq (P_1(T_0 \leq n))^y\,. \tag{30.5}$$

Using the same reasoning as we did to deduce that $\rho_{0x} = \rho^x$, it follows that

$$P_y(T_0 \leq n) \leq P_y(T_{y-1} \leq n)P_{y-1}(T_{y-2} \leq n)\cdots P_1(T_0 \leq n).$$

Since

$$P_z(T_{z-1} \leq n) = P_1(T_0 \leq n)$$

it follows that

$$P_y(T_0 \leq n) \leq (P_1(T_0 \leq n))^y\,.$$

From this

$$P_1(T_0 \leq n+1) \leq f(0) + \sum_{y=1}^{\infty} f(y)\,(P_1(T_0 \leq n))^y$$

which establishes (30.5).

Using an argument identical to the one for the branching chain, it now follows that

$$P_1(T_0 \leq n) \leq t_1.$$

Letting $n \to \infty$ gives

$$P_1(T_0 < \infty) \leq t_1.$$

Thus if $\mu > 1$ it must be the case that $\rho = t_1$ and so 0 is transient. Since the chain is irreducible, all states must be transient.

∎

If $\mu > 1$ and if the chain is not irreducible, it is still possible to deduce that the chain is transient; see the exercises.

30. Extinction Probabilities: Problems.

1. Consider a branching chain with $f(0) = f(3) = 1/2$. Find the probability ρ of extinction.

2. Consider a branching chain with

$$f(x) = p(1-p)^x$$

for $x \geq 0$, where $0 < p < 1$. Show that $\rho = 1$ if $p \geq 1/2$ and that

$$\rho = \frac{p}{1-p}$$

if $p < 1/2$.

3. Let $\{X_n\}$ be the queuing chain.
(a) If either $f(0) = 0$ or if $f(0) + f(1) = 1$, show that the chain is not irreducible.
(b) If $f(0) > 0$ and $f(0) + f(1) < 1$, then show that the chain is irreducible.

4. Let $\{X_n\}$ be the queuing chain which is not irreducible. Determine which states are absorbing, recurrent and transient by considering the following four cases.
(a) $f(1) = 1$;
(b) $f(0) > 0$, $f(1) > 0$, and $f(0) + f(1) = 1$;
(c) $f(0) = 1$;
(d) $f(0) = 0$ and $f(1) = 1$.

5. Let $\{X_n\}$ be a queuing chain that is not irreducible and suppose that $\mu > 1$. Show that (d) of the previous exercise applies and hence that the chain is transient.

31. Stationary Distributions: Definitions and Examples

When we examined the 2-state chain we saw there was an important connection between the asymptotic behavior of the chain

$$\lim_{n\to\infty} X_n$$

and the stationary distribution π that satisfied

$$\pi P = \pi$$

where π was written as a row-vector. In this section we begin to explore this relationship in greater detail. We start by recalling the definition of a stationary distribution.

31.1. Definition.

Let $\{X_n\}$ be a Markov Chain with transition function P. A distribtution π is a **stationary distribution** *provided that*

$$\sum_x \pi(x) P(x, y) = \pi(y)$$

for all y.

31.2. Proposition.

If π is a stationary distribution, then

$$\sum_x \pi(x) P^n(x, y) = \pi(y)$$

for all y.

Proof. Note that

$$\begin{aligned}\sum_x \pi(x)P^2(x,y) &= \sum_x \pi(x)\sum_z P(x,z)P(z,y)\\ &= \sum_z\sum_x \pi(x)P(x,z)P(z,y)\\ &= \sum_z\left(\sum_x \pi(x)P(x,z)\right)P(z,y)\\ &= \sum_z \pi(z)P(z,y)\\ &= \pi(y).\end{aligned}$$

A simple induction verifies the conclusion. ∎

The following corollary is immediate.

31.3. Corollary.

If X_0 has a stationary distribution π, then

$$\mathfrak{Pr}\,(X_n = y) = \pi(y)$$

and in particular the distribution of X_n is independent of n. Conversely, if the distribution of X_n is independent of n then X_0 has a stationary distribution.

The next result gives conditions under which the asymptotic behavior of the chain can be deduced from the behavior of the stationary distribution. This is analogous to the result we deduced for the two-state chain.

31.4. Proposition.

Supoose that π is a stationary distribution and that for each y

$$\lim_{n\to\infty} P^n(x,y) = \pi(y). \tag{31.1}$$

Then for all y

$$\lim_{n\to\infty} \mathfrak{Pr}\,(X_n = y) = \pi(y).$$

Proof. We know that

$$\mathfrak{Pr}\,(X_n = y) = \sum_x \pi(x) P^n(x,y)$$

so

$$\begin{aligned}
\lim_{n\to\infty} \mathfrak{Pr}\,(X_n = y) &= \lim_{n\to\infty} \sum_x \pi_0(x) P^n(x,y) \\
&= \sum_x \pi_0(x) \lim_{n\to\infty} P^n(x,y) \\
&= \sum_x \pi_0(x)\pi(y) \\
&= \pi(y).
\end{aligned}$$

∎

In the next section, we will devote some effort to deciding when (31.1) holds and hence when the conclusions of the preceding proposition are valid.

31.5. Example.

Suppose that a Markov Chain on $\{0, 1, 2\}$ *has transition matrix*

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \end{pmatrix} \end{array}$$

Find the (unique) stationary distribution.

Solution. The stationary distribution π must satisfy

$$\begin{aligned} \frac{\pi(0)}{3} + \frac{\pi(1)}{4} + \frac{\pi(2)}{6} &= \pi(0) \\ \frac{\pi(0)}{3} + \frac{\pi(1)}{2} + \frac{\pi(2)}{3} &= \pi(1) \\ \frac{\pi(0)}{3} + \frac{\pi(1)}{4} + \frac{\pi(2)}{2} &= \pi(2) \end{aligned}$$

and also

$$\pi(0) + \pi(1) + \pi(2) = 1.$$

This reduces to $\pi(0) = 6/25$, $\pi(1) = 2/5$ and $\pi(2) = 9/25$.

∎

31.6. Example.

Let $\{X_n\}$ be the Ehrenfest chain on $\{0, 1, 2, 3\}$, so that the transition matrix is

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array}\begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Find the (unique) stationary distribution π.

Solution. We obtain

$$\begin{aligned} \frac{1}{3}\pi(1) &= \pi(0) \\ \pi(0) + \frac{2}{3}\pi(2) &= \pi(1) \\ \frac{2}{3}\pi(1) + \pi(3) &= \pi(2) \\ \frac{1}{3}\pi(2) &= \pi((3) \end{aligned}$$

and also

$$\pi(0) + \pi(1) + \pi(2) + \pi(3) = 1.$$

This reduces to

$$\pi(0) = \frac{1}{8}, \ \pi(1) = \frac{3}{8}, \ \pi(2) = \frac{3}{8} \text{ and } \pi(3) = \frac{1}{8}.$$

∎

Note that in the preceding example $P^n(x, x) = 0$ for odd values of n and hence it is not possible for

$$\lim_n P^n(x, y) = \pi(y)$$

for all y. This shows that the elementary conclusions we were able to obtain for the two-state chain will need some refinement.

31.7. Example. Birth and Death Chains.

In this example we deduce necessary and sufficient conditions for an irreducible birth and death chain to have a stationary distribution.

Solution. Let $\{X_n\}$ be a birth and death chain. We will permit the state space to be either finite

$$\mathcal{S} = \{0, 1, \cdots, d\}$$

or infinite

$$\mathcal{S} = \{0, 1, 2, \cdots\}.$$

In the case that the state space is finite, the assumption that the chain is irreducible reduces to

$$\begin{aligned} p_x &> 0 \qquad (0 \le x < d) \\ q_x &> 0 \qquad (0 < x \le d) \end{aligned}$$

while in the case the state space is infinite this assumption becomes

$$\begin{aligned} p_x &> 0 \qquad (0 \le x < \infty) \\ q_x &> 0 \qquad (0 < x < \infty). \end{aligned}$$

Now a necessary condition for π to be a stationary distribution is

$$\sum_x \pi(x)P(x,y) = \pi(y)$$

for all $y \in \mathcal{S}$. For $y = 0$ this becomes

$$\pi(0)r_0 + \pi(1)q_1 = \pi(0)$$

while for $y > 0$ we obtain

$$\pi(y-1)p_{y-1} + \pi(y)r_y + \pi(y+1)q_{y+1} = \pi(y).$$

Since $p_y + r_y + q_y = 1$, the case $y = 0$ is equivalent to

$$-p_0\pi(0) + \pi(1)q_1 = 0$$

and the case $y > 0$ is equivalent to

$$q_{y+1}\pi(y+1) - p_y\pi(y) = q_y\pi(y) - p_{y-1}\pi(y-1).$$

Applying this relationship repeatedly we obtain

$$\begin{aligned} q_{y+1}\pi(y+1) - p_y\pi(y) &= q_y\pi(y) - p_{y-1}\pi(y-1) \\ &= q_{y-1}\pi(y-1) - p_{y-2}\pi(y-2) \\ &\ \ \vdots \\ &= \pi(1)q_1 - p_0\pi(0) \\ &= 0. \end{aligned}$$

In particular then

$$q_{y+1}\pi(y+1) - p_y\pi(y) = 0$$

or

$$\pi(y+1) = \frac{p_y}{q_{y+1}}\pi(y).$$

Applying this formula recursively,

$$\pi(x) = \frac{p_0 \cdots p_{x-1}}{q_1 \cdots q_x}\pi(0).$$

Now if we define

$$\pi_x = \begin{cases} 1 & x = 0 \\ \frac{p_0 \cdots p_{x-1}}{q_1 \cdots q_x} & x \geq 1 \end{cases}$$

Then a necessary condition for $\pi(x)$ to be a stationary distribution is that

$$\pi(x) = \pi_x\pi(0).$$

Another necessary condition is that

$$\sum_x \pi(x) = 1.$$

Since a stationary distribution must satsify

$$\pi(x) = \pi_x\pi(0)$$

it follows that

$$1 = \sum_x \pi_x \pi(0)$$

or $(\pi(0))^{-1} = \sum_x \pi_x$. In particular, in order for a stationary distribution to exist, we must also have

$$\sum_x \pi_x < \infty.$$

Thus in order for a stationary distribution to exist we must have

$$\sum_x \pi_x < \infty$$

in which case the stationary distribution must be given by

$$\pi(x) = \frac{\pi_x}{\sum_y \pi_y} \qquad (31.2).$$

It is a routine matter to show that $\pi(x)$ defined by (31.2) is in fact a stationary distribution (see the exercises).

▮

Our next example proposes one possible model for the number of calls active in telephone switch. Many of the assumptions in the model, such as that new calls arrive according a Poisson distribution, are supported by empirical data. We will return to refinements of this model in laster sections.

31.8. Example. Telephone Switch Model.

In this example, we will suppose that X_n represents the number of calls active in a phone switch at time interval n. We will suppose that call is active at time $n+1$ in exactly one of two ways:

1. *The call was active at time n and not terminated during the time interval between n and $n+1$; or*
2. *The call was newly initiated at time $n+1$.*

Thus

$$X_{n+1} = R(X_n) + \xi_n$$

where

$$R(X_n) = \begin{cases} \text{\# of calls in progress at the start of interval } n \\ \text{that are still in progress at the start of interval } n+1 \end{cases}$$

and

$$\xi_n = \text{\# of new calls initiated in interval } n.$$

We further assume that the random variables $\{\xi_i\}$ are independent Poisson variables with parameter λ, and that the $\{\xi_i\}$ are independent of $R(X_n)$. Moreover, we suppose that the duration of a paricular call is independent of when it was initiated (n) or how many other calls are in the switch (the value of X_n) or how many new calls are iniitiated. We also suppose that in any time period there is a fixed probability q that a call in progress will terminate before the start of the next period. Equivalently, a call has probability $p = 1 - q$ of continuing into the next interval.

We will verify that this is a Markov Chain, find a stationary distribution for the chain and determine the value of $\lim_n X_n$.

Solution. First note that

$$\mathfrak{Pr}\left(R(X_n) = z \middle| X_n = x\right) = \binom{x}{z} p^z (1-p)^{x-z}$$

for $0 \le z \le x$, i.e, $R(X_n)\big|_{X_n=x}$ is binomial. Further, by assumption

$$\mathfrak{Pr}\left(\xi_n = z\right) = \frac{\lambda^z e^{-\lambda}}{z!}.$$

Since

$$\begin{aligned}\mathfrak{Pr}\left(X_{n+1}=y\middle|X_n=x\right)&=\sum_{z=0}^{\min\{x,y\}}\mathfrak{Pr}\left(R(X_n)=z,\,\xi_{n+1}=y-z\middle|X_n=x\right)\\&=\sum_{z=0}^{\min\{x,y\}}\mathfrak{Pr}\left(\xi_{n+1}=y-z\right)\mathfrak{Pr}\left(R(X_n)=z\middle|X_n=x\right)\\&=\sum_{z=0}^{\min\{x,y\}}\frac{\lambda^{y-z}e^{-\lambda}}{(y-z)!}\binom{x}{z}p^z(1-p)^{x-z}\\&=P(x,y)\end{aligned}$$

Thus $P(x,y)>0$ and hence the chain is irreducible. Coincidentally, this verifies that

$$\mathfrak{Pr}\left(X_{n+1}=y\middle|X_n=x\right)$$

does not depend on n.

Next we establish the following

Claim. If X_n has a Poisson distribution with parameter t then $R(X_n)$ has a Poisson distribution with parameter pt.

In order to establish the claim we compute:

$$\begin{aligned}
\mathfrak{Pr}\left(R(X_n = y\right) &= \sum_{x=y}^{\infty} \mathfrak{Pr}\left(X_n = x,\, R(X_n) = y\right) \\
&= \sum_{x=y}^{\infty} \mathfrak{Pr}\left(X_n = x\right) \mathfrak{Pr}\left(R(X_n) = y \middle| X_n = x\right) \\
&= \sum_{x=y}^{\infty} \frac{t^x e^{-t}}{x!} \binom{x}{y} p^y (1-p)^{x-y} \\
&\quad \text{(applying the assumption that } X_n \text{ is Poisson)} \\
&= \sum_{x=y}^{\infty} \frac{t^x e^{-t}}{y!(x-y)!} p^y (1-p)^{x-y} \\
&= \frac{(pt)^y e^{-t}}{y!} \sum_{x=y}^{\infty} \frac{t^{x-y}}{(x-y)!} (1-p)^{x-y} \\
&= \frac{(pt)^y e^{-t}}{y!} \sum_{m=0}^{\infty} \frac{t^m}{m!} (1-p)^m \\
&= \frac{(pt)^y e^{-t} e^{t(1-p)}}{y!} \\
&= \frac{(pt)^y e^{-tp}}{y!}
\end{aligned}$$

verifying the claim.

We now can verify that a stationary distribution for this chain exists and has a Poisson distribution with parameter $t = \lambda/q = \lambda/(1-p)$.

To see this, we start by supposing that X_0 has a Poisson distribution with parameter t. Then $R(X_0)$ has a Poisson distribtuion with parameter pt and so

$$X_1 = \xi_1 + R(X_0)$$

is the sum of two Poisson random variables, one with parameter λ and the other with parameter pt. Thus X_1 has a Poisson distribution with parameter $\lambda + pt$.

Applying this argument recursively, one can show (see the exercises) that if X_0 has a Poisson distribution with parameter t then X_n has a Poisson distribution with parameter

$$tp^n + \frac{\lambda}{q}(1 - p^n). \tag{31.3}$$

We will use this fact in the sequel.

From the above, X_1 and X_0 will have the same distribution if and only if

$$t = \lambda + pt$$

or equivalently

$$t = \frac{\lambda}{1-p} = \frac{\lambda}{q}.$$

Finally, we examine the limiting behavior of X_n under the assumption that X_0 has a Poisson distribution.

Claim. *If X_0 has a Poisson distribution with parameter t then*

$$\lim_{n\to\infty} P^n(x,y) = \frac{e^{-\lambda/q}\left(\frac{\lambda}{q}\right)^y}{y!}$$

so in particular

$$\lim_{n\to\infty} \mathfrak{Pr}\left(X_n = y\right) = \frac{e^{-\lambda/q}\left(\frac{\lambda}{q}\right)^y}{y!}.$$

To establish this, suppose that X_0 has a Poisson distribution with parameter t. Then applying (31.3),

$$\begin{aligned}
&\sum_{x=0}^{\infty} \frac{e^{-t}t^x}{x!} P^n(x,y) = \mathfrak{Pr}\left(X_n = y\right) \\
&\quad = \frac{\left[tp^n + \frac{\lambda}{q}(1-p^n)\right]^y}{y!} \exp\left(-\left[tp^n + \frac{\lambda}{q}(1-p^n)\right]\right) \\
&\quad = \frac{e^{-\lambda/q(1-p^n)}}{y!} \left(\sum_{j=0}^{y} \binom{y}{j} (tp^n)^j \left(\frac{\lambda}{q}(1-p^n)\right)^{y-j}\right) \left(\sum_{j=0}^{\infty} \frac{(-tp^n)^j}{(j)!}\right)
\end{aligned}$$

applying binomial expansion to the first factor and Taylor's series to the second. From this we have two expressions for $\mathfrak{Pr}\left(X_n = y\right)$, one a power series

$$\sum_{x=0}^{\infty} \frac{e^{-t}t^x}{x!} P^n(x,y)$$

and the other a product of two power series:

$$\frac{e^{-\lambda/q(1-p^n)}}{y!}\left(\sum_{j=0}^{y}\binom{y}{j}(tp^n)^j\left(\frac{\lambda}{q}(1-p^n)\right)^{y-j}\right)\left(\sum_{j=0}^{\infty}\frac{(-tp^n)^j}{(j)!}\right)$$

This means we are in a position to apply the following fact about power series:
Claim. *If*

$$\sum_x c_x t^x = \left(\sum_x a_x t^x\right)\left(\sum_x b_x t^x\right)$$

and the power series have positive radius of convergence, then

$$c_x = \sum_{z=0}^{x} a_z b_{x-z}.$$

In particular, if $a_z = 0$ *for* $z > y$ *then*

$$c_x = \sum_{z=0}^{\min\{x,y\}} a_z b_{x-z}.$$

This claim is plausible since

$$\begin{aligned}\sum_x c_x t^x &= \left(\sum_x a_x t^x\right)\left(\sum_x b_x t^x\right)\\ &= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} b_x a_y t^{x+y}\\ &= \sum_{x=0}^{\infty}\sum_{z=x}^{\infty} b_x a_{z-z} t^z\\ &= \sum_{z=0}^{\infty}\sum_{x=0}^{z} b_x a_{z-z} t^z\\ &= \sum_{z=0}^{\infty} t^z\left(\sum_{x=0}^{z} b_x a_{z-z}\right).\end{aligned}$$

If the power series have positive radius of convergence, we can equate the terms and the result follows.

The second conclusion of the claim is applicable to $\mathfrak{Pr}\left(X_n = y\right)$, from which we obtain

$$\frac{P^n(x,y)}{x!} = \frac{e^{-\lambda(1-p^n)/q}}{y!} \sum_{z=0}^{\min\{x,y\}} \binom{y}{z} p^{nz} \left(\frac{\lambda}{q}(1-p^n)\right)^{y-z} \frac{(1-p^n)^{x-z}}{(x-z)!}$$

This simplifies (slightly!) to

$$P^n(x,y) = e^{-\lambda(1-p^n)/q} \sum_{z=0}^{\min\{x,y\}} \binom{x}{z} p^{nz} (1-p^n)^{x-z} \frac{\left[\frac{\lambda}{q}(1-p^n)\right]^{y-z}}{(y-z)!}$$

Now if we let $n \to \infty$, then $(1-p^n) \to 1$ and $p^{nz} \to 0$. Thus the only term in the sum that does not vanish is the $z = 0$ term. This implies that

$$\lim_{n\to\infty} P^n(x,y) = \frac{e^{-\lambda/q}\left(\frac{\lambda}{q}\right)^y}{y!}$$

as desired.

∎

31. Stationary Distributions: Definitions and Examples: Problems.

1. Consider a Markov Chain having the state space $\{0,1,2\}$ and transition matrix

$$P = \begin{array}{c c} & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \end{array} & \begin{pmatrix} 0.4 & 0.4 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.4 & 0.4 \end{pmatrix} \end{array}.$$

Show that the chain has a unique stationary distribution π and find it.

2. Let $\{X_n\}$ be a Markov Chain on the finite state space $\{0,1,\cdots,d\}$ and having transition matrix $P=(a_{ij})$. Let λ be an eigenvalue for P, i.e., a solution to the equation

$$\det(P-\lambda I)=0.$$

Note that λ is an eigenvalue for P if and only if there is a non-zero vector v such that $(P-\lambda I)v=0$.

(a) Let $P^T=(a_{ji})$ be the transpose of P. Show that P and P^T have the same eigenvalues.

(b) Show that 1 is a eigenvalue for P.

(c) Show that if λ is an eigenvalue for P then $|\lambda|\le 1$.

3. Let $\{X_n\}$ be the chain described in example 31.8. Show that if X_0 has a Poisson distribtuion with parameter t then X_n has a Poisson distribtuion with parameter

$$tp^n+\frac{\lambda}{q}(1-p^n).$$

4. Let $\{X_n\}$ be the chain described in example 31.8. Show that

$$E_x(X_n)=xp^n+\frac{\lambda}{q}(1-p^n).$$

32. Stationary Distributions: Results

As we have already noted, there are chains for which

$$\lim_n P^n(x,y)$$

does not exist. For example, in the birth-and-death chain with $r_x = 0$ for all x, then $P^n(x,x) = 0$ for odd values of n. However it turns out that

$$\lim_n \frac{1}{n} P^n(x,y)$$

has more regular behavior, and so we examine the above limit instead.

We begin by recalling some definitions.

32.1. Definition.

For numbers x and y we define the **indicator funtion** *of $\{y\}$ to be*

$$\mathbf{1}_y(x) = \begin{cases} 1 & \text{if x=y} \\ 0 & \text{otherwise} \end{cases}$$

Note that

$$E_x(\mathbf{1}_y(X_n)) = \mathfrak{Pr}\left(X_n = y \middle| X_0 = x\right) = P^n(x,y).$$

32.2. Definition.

For a non-negative integer n and for $y \in \mathcal{S}$ we define

$$\mathfrak{N}_n(y) = \sum_{m=1}^{n} \mathbf{1}_y(X_m)$$

so that $\mathfrak{N}_n(y)$ counts the number of visits to state y in the first n time intervals.

32.3. Definition.

For a non-negative integer n and for $x, y \in \mathcal{S}$ we define

$$\mathfrak{G}_{\mathbf{n}}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{m}=1}^{\mathbf{n}} \mathbf{P}^{\mathbf{m}}(\mathbf{x}, \mathbf{y}).$$

Clearly

$$E_x(\mathfrak{N}_n(y)) = \mathfrak{G}_{\mathbf{n}}(\mathbf{x}, \mathbf{y}).$$

Also note that if $y \in \mathcal{S}$ is transient then

$$\lim_{n\to\infty} \mathfrak{N}_n(y) = N(y) < \infty$$

with probability one and

$$\lim_{n\to\infty} \mathfrak{G}_{\mathbf{n}}(\mathbf{x}, \mathbf{y}) = \mathfrak{G}(\mathbf{x}, \mathbf{y}) < \infty.$$

Thus

$$\lim_{n\to\infty} \frac{\mathfrak{N}_n(y)}{n} = 0$$

with probability one and

$$\lim_{n\to\infty} \frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x}, \mathbf{y})}{\mathbf{n}} = 0$$

with probability one.

From the definitions

$$\frac{\mathfrak{N}_n(y)}{n} = \text{\% of time the chain is in state } y \text{ in } 1^{st}\ n \text{ states.}$$

and

$$\frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x}, \mathbf{y})}{\mathbf{n}}$$

is the expected value of $\mathfrak{N}_n(y)/n$.

32.4. Definition.

For $y \in \mathcal{S}$ we define

$$\text{ш}_y = \begin{cases} E_y(T_y) & \textit{if } E_y(T_y) < \infty \\ \infty & \textit{otherwise} \end{cases}.$$

We further define the random variable $\mathbf{1}_{\{T_y<\infty\}}$ to be

$$\mathbf{1}_{\{T_y<\infty\}} = \begin{cases} 1 & \textit{if } T_y < \infty \\ 0 & \textit{otherwise} \end{cases}$$

Our first main result uses the Strong Law of Large Numbers.

32.5. Theorem. Strong Law of Large Numbers.

Let $\{\xi_i\}$ be independent and identically distributed random variables having finite mean μ. Then with probability one

$$\lim_{n\to\infty} \frac{\xi_1 + \cdots \xi_n}{n} = \mu.$$

If $\xi \geq 0$ and $\mu = \infty$ then the above limits still holds.

This is considerably stronger than the Weak Law proved earlier (and in fact implies the Weak Law). An elementary, if somewhat technical proof, can be found in *An Elementary Proof of the Strong Law of Large Numbers*, E. Etemadi, **Zeitschrift fiir Wahrscheinlichkeitstheorie und verwandte Gebiete**, volume 51, number 1, pages 119-122, Springer-Verlag 1981.

32.6. Theorem.

Let y be a recurrent state. Then

(a)

$$\lim_{n\to\infty} \frac{\mathfrak{N}_n(y)}{n} = \frac{\mathbf{1}_{\{T_y<\infty\}}}{\text{Ш}_y}$$

with probability one.

(b)

$$\lim_{n\to\infty} \frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x},\mathbf{y})}{\mathbf{n}} = \frac{\rho_{xy}}{\text{Ш}_y}$$

for all $x \in \mathcal{S}$.

Proof. Let y be a recurrent state and, for $r \geq 1$, set

$$T_y^r = \min\{n \geq 1 : \mathfrak{N}_n(y) = r\}$$

so that

$$T_y^r = \text{time of the } r^{th} \text{ visit to state } y.$$

Next we define a sequence of random variables $\{\xi_y^r\}$ by

$$\xi_y^1 = T_y^1 = T_y$$

and, for $r \geq 2$,

$$\xi_y^r = T_y^r - T_y^{r-1}$$

so that

$$\xi_i^r = \text{waiting time between } (r-1)^{st} \text{ and } r^{th} \text{ visit.}$$

Clearly

$$\sum_{k=1}^{r} \xi_y^k = T_y^r.$$

Now since $\{X_n\}$ is a Markov Chain with stationary transition probabilities, $\{\xi_y^1, \cdots, \xi_y^r\}$ are independent and identically distributed random variables having common mean

$$E_y(\xi_y^1) = E_y(T_y) = \text{Ш}_y\,.$$

The Strong Law of Large Numbers then implies that

$$\lim_{k\to\infty} \frac{\xi_y^1 + \cdots \xi_y^k}{k} = \text{щ}_y$$

with probability one. This in turn implies that

$$\lim_{k\to\infty} \frac{T_y^k}{k} = \text{щ}_y$$

with probability one.

Next observe that

$$\underbrace{T_y^{\mathfrak{N}_n(y)}}_{\mathfrak{N}_n(y)^{th} \text{ visit occurs before } n} \leq n \leq \underbrace{T_y^{\mathfrak{N}_n(y)+1}}_{(\mathfrak{N}_n(y)+1)^{st} \text{ visit occurs after } n}.$$

This implies that

$$\frac{T_y^{\mathfrak{N}_n(y)}}{\mathfrak{N}_n(y)} \leq \frac{n}{\mathfrak{N}_n(y)} \leq \frac{T_y^{\mathfrak{N}_n(y)+1}}{\mathfrak{N}_n(y)}$$

with probability one. Now since $\mathfrak{N}_n(y) \to \infty$ as $n \to \infty$, this implies that

$$\text{щ}_y \leq \lim_{n\to\infty} \frac{n}{\mathfrak{N}_n(y)} \leq \text{щ}_y$$

with probability one. Equivalently, with probability one,

$$\lim_{n\to\infty} \frac{\mathfrak{N}_n(y)}{n} = \frac{\mathbf{1}_{\{T_y<\infty\}}}{\text{щ}_y}.$$

Taking expectations completes the proof.

∎

32.7. Corollary.

Let $C \subseteq \mathcal{S}$ be a closed, irreducible collection of recurrent states. Then

$$\lim_{n\to\infty} \frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x},\mathbf{y})}{\mathbf{n}} = \frac{1}{\text{Ш}_y}$$

for $x, y \in \mathcal{C}$. If in addition $\mathfrak{Pr}\left(X_n \in C\right) = 1$ then

$$\lim_{n\to\infty} \frac{\mathfrak{N}_n(y)}{n} = \frac{1}{\text{Ш}_y}$$

with probability one.

32.8. Definition.

A recurrent state $y \in \mathcal{S}$ is said to be **null recurrent** *if $\text{Ш}_y = \infty$.*

Via 32.6, if y is null recurrent then

$$\lim_{n\to\infty} \frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x},\mathbf{y})}{\mathbf{n}} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^n(x,y) = 0$$

for all $x \in \mathcal{S}$.

32.9. Definition.

A recurrent state $y \in \mathcal{S}$ is said to be **positive recurrent** *if $\text{Ш}_y < \infty$.*

Applying 32.6 again, we see that if y is positive recurrent then

$$\lim_{n\to\infty} \frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x},\mathbf{y})}{\mathbf{n}} = \frac{1}{\text{Ш}_y} > 0.$$

32.10. Theorem.

Let $x \in \mathcal{S}$ be positive recurrent and suppose that x leads to y. Then y is positive recurrent and y leads to x.

Proof. Since x is recurrent and x leads to y, it follows that y is recurrent and y leads to x. Thus there are integers n_1 and n_2 such that

$$P^{n_1}(y,x) > 0 \quad \text{and} \quad P^{n_2}(x,y) > 0.$$

Now for any integer m,

$$P^{n_1+m+n_2}(y,y) \geq P^{n_1}(y,x)P^m(x,x)P^{n_2}(x,y).$$

Thus, upon summing from $m = 1$ to $m = n$ we obtain

$$\begin{aligned}
\frac{\mathfrak{G}_{\mathbf{n_1+n+n_2}}(\mathbf{y},\mathbf{y})}{\mathbf{n}} - \frac{\mathfrak{G}_{\mathbf{n_1+n_2}}(\mathbf{y},\mathbf{y})}{\mathbf{n}} &= \frac{1}{n}\sum_{m=1}^{n_1+n+n_2} P^m(y,y) - \frac{1}{n}\sum_{m=1}^{n_1+n_2} P^m(y,y) \\
&= \frac{1}{n}\sum_{m=n_1+n_2+1}^{n_1+n+n_2} P^m(y,y) \\
&\geq P^{n_1}(y,x)P^{n_2}(x,y)\frac{1}{n}\sum_{m=1}^{n} P^m(x,x) \\
&= P^{n_1}(y,x)p^{n_2}(x,y)\frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x},\mathbf{y})}{\mathbf{n}}
\end{aligned}$$

Now as $n \to \infty$, the left-hand-side of the above inequality goes to $1/\text{Ш}_y$ while the right-hand-side goes to

$$P^{n_1}(y,x)p^{n_2}(x,y)\frac{1}{\text{Ш}_x} > 0.$$

Thus $0 < 1/\text{Ш}_y$ and so $\text{Ш}_y < \infty$, implying that y is positive recurrent. ∎

32.11. Corollary.

If $C \subseteq \mathcal{S}$ is a closed, irreducible collection of states, then either
(a) *every state in C is transient; or*
(b) *every state in C is null recurrent; or*
(c) *every state in C is positive recurrent.*

32.12. Corollary.

If $C \subseteq \mathcal{S}$ is finite and closed, then there is an $x \in \mathcal{S}$ such that x is positive recurrent.

Proof. Suppose for contradiction that every $x \in \mathcal{S}$ is either transient or null recurrent. Since C is closed,

$$\sum_{y \in C} P^m(x, y) = 1$$

for all $x \in C$ and for all integers m. Summing this from $m = 1$ to $m = n$ and dividing by n, we see that

$$\begin{aligned} 1 &= \frac{1}{n} \sum_{y \in C} \sum_{m=1}^{n} P^m(x, y) \\ &= \sum_{y \in C} \frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x}, \mathbf{y})}{\mathbf{n}}. \end{aligned}$$

But we have assumed that C has only transient and null recurrent states, and so

$$\lim_{n \to \infty} \frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x}, \mathbf{y})}{\mathbf{n}} = 0$$

for all $x, y \in C$. Since C is finite, we can interchange the summmation and the limit and so obtain

$$\begin{aligned} 1 &= \lim_{n \to \infty} \frac{1}{n} \sum_{y \in C} \sum_{m=1}^{n} P^m(x, y) \\ &= \sum_{y \in C} \lim_{n \to \infty} \frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{x}, \mathbf{y})}{\mathbf{n}} \\ &= 0 \end{aligned}$$

a contradiction. Thus C must have at least one positive recurrent state.

∎

32.13. Theorem.

If $C \subseteq \mathcal{S}$ is closed, irreducible and finite, then every state in C is positive recurrent.

32.14. Corollary.

An irreducible Markov Chain having a finite state space is positive recurrent.

32.15. Corollary.

A Markov Chain having a finite number of states has no null recurrent states.

Proof. If $y \in \mathcal{S}$ is recurrent, then there is a closed, irreducible set $C \subseteq \mathcal{S}$ with $y \in C$. But a closed, irreducible, finite set must have a postive recurrent state x. Since C is closed and irreducible, y must also be postive recurrent.

∎

Our next results require the use of a special case of an advanced theorem in analysis.

32.16. Theorem. Bounded Convergence Theorem for Sequences.

Let $\alpha(x)$, $\beta_n(x)$ and $\beta(x)$ be sequences where x ranges over some set $\mathcal{S} \subseteq \mathbb{N}$. Suppose that $\alpha(x) \geq 0$ for all x and that

$$\sum_x \alpha(x) < \infty.$$

Suppose that $|\beta_n(x)| \leq 1$ for each x and each n and that, for each fixed x,

$$\lim_{n\to\infty} \beta_n(x) = \beta(x).$$

Then

$$\begin{aligned}\lim_{n\to\infty} \sum_x \alpha(x)\beta_n(x) &= \sum_x \alpha(x) \lim_{n\to\infty} \beta_n(x)\\ &= \sum_x \alpha(x)\beta(x).\end{aligned}$$

Proof. Let $\epsilon > 0$ be an arbitrarily small number. We can choose N so large that

$$\sum_{x\geq N} \alpha(x) < \frac{\epsilon}{2}.$$

Then

$$\begin{aligned}\Big|\sum_x \beta_n(x)\alpha(x) - \sum_x \beta(x)\alpha(x)\Big| &\leq \sum_x |\beta_n(x) - \beta(x)|\,\alpha(x)\\ &= \sum_{x<N} |\beta_n(x) - \beta(x)|\,\alpha(x) + \sum_{x\geq N} |\beta_n(x) - \beta(x)|\,\alpha(x)\\ &\leq \sum_{x<N} |\beta_n(x) - \beta(x)|\,\alpha(x) + \sum_{x\geq N} \left(|\beta_n(x)| + |\beta(x)|\right)\alpha(x)\\ &\leq \sum_{x<N} |\beta_n(x) - \beta(x)|\,\alpha(x) + \sum_{x\geq N} 2\alpha(x)\\ &\leq \sum_{x<N} |\beta_n(x) - \beta(x)|\,\alpha(x) + \epsilon\end{aligned}$$

or

$$\Big|\sum_x \beta_n(x)\alpha(x) - \sum_x \beta(x)\alpha(x)\Big| \le \sum_{x<N} |\beta_n(x) - \beta(x)|\,\alpha(x) + \epsilon.$$

Now since the sum on the right-hand-side is over only finitely many terms, we may interchange the sum and the limit when we let $n \to \infty$. The sum on the right-hand-side vanishes as $n \to \infty$, so we can conclude that

$$\limsup_{n\to\infty}\left|\sum_x \beta_n(x)\alpha(x) - \sum_x \beta(x)\alpha(x)\right| \le \epsilon.$$

Since $\epsilon > 0$ was arbitrary, this proves the result.

∎

32.17. Theorem.

Let $\{X_n\}$ be a Markov Chain having stationary distribution π. If $x \in \mathcal{S}$ is a transient or null recurrent state, then $\pi(x) = 0$.

Proof. If z is transient or null recurrent we have shown that

$$\lim_{n\to\infty} \frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{y})}{\mathbf{n}} = 0$$

for all $z \in \mathcal{S}$. Since π is a stationary distribution,

$$\sum_{z\in\mathcal{S}} \pi(z)P^m(z,x) = \pi(x)$$

for all $x \in \mathcal{S}$. Summing this from $m = 1$ to $m = n$ and dividing by n gives

$$\sum_{z\in\mathcal{S}} \pi(z)\frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{x})}{\mathbf{n}} = \pi(x).$$

But since

$$0 \le \frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{x},\mathbf{y})}{\mathbf{n}} = \frac{1}{n}\sum_{m=1}^{n} P^m(x,y) \le 1$$

the bounded convergence theorem implies

$$\begin{aligned}\pi(x) &= \lim_{n\to\infty}\sum_{z\in\mathcal{S}}\pi(z)\frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{x})}{\mathbf{n}}\\ &= \sum_{z\in\mathcal{S}}\pi(z)\lim_{n\to\infty}\frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{x})}{\mathbf{n}}\\ &= 0\end{aligned}$$

∎

32.18. Theorem.

Let $\{X_n\}$ be an irreducible, closed Markov Chain. Then $\{X_n\}$ has a unique stationary distribution π given by

$$\pi(x) = \frac{1}{\text{Щ}_x}.$$

Proof. We know that

$$\lim_{n\to\infty}\frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{x})}{\mathbf{n}} = \frac{1}{\text{Щ}_y}$$

for all $x, y \in \mathcal{S}$.

First suppose that π is a stationary distribution. As before,

$$\sum_z \pi(z)\frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{x})}{\mathbf{n}} = \pi(x).$$

So, taking limits, and again applying the Bounded Convergence Theorem,

$$\begin{aligned}\pi(x) &= \lim_{n\to\infty}\sum_z\pi(z)\frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{x})}{\mathbf{n}}\\ &= \sum_z\pi(z)\lim_{n\to\infty}\frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{x})}{\mathbf{n}}\\ &= \frac{1}{\text{Щ}_x}\sum_z\pi(z)\\ &= \frac{1}{\text{Щ}_x}.\end{aligned}$$

Thus if π is a stationary distribution, then it must satisfy $\pi(x) = 1/\text{Щ}_x$.

For the converse set

$$\pi(x) = \frac{1}{\text{Щ}_x}.$$

It will suffice to show that

(a) $\sum_x \frac{1}{\text{Щ}_x} = 1$; and

(b)

$$\sum_x \frac{1}{\text{Щ}_x} P(x,y) = \frac{1}{\text{Щ}_y}.$$

The Bounded Convergence Theorem no longer helps with this part of the proof. However, if $\mathcal{S}$ is finite, the result is fairly direct. First note that

$$\sum_{x \in \mathcal{S}} P^m(z,x) = 1$$

and so

$$\sum_{x \in \mathcal{S}} \frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{y})}{\mathbf{n}} = 1.$$

If $\mathcal{S}$ is finite, we can let $n \to \infty$ and interchange the limit and the sum to get

$$\sum_{x \in \mathcal{S}} \frac{1}{\text{Щ}_x} = 1.$$

Similarly, since

$$\sum_{x \in \mathcal{S}} P^m(z,x) P(x,y) = P^{m+1}(z,y)$$

it follows that

$$\sum_{x \in \mathcal{S}} \frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{x})}{\mathbf{n}} P(x,y) = \frac{\mathfrak{G}_{\mathrm{n}}(\mathbf{z},\mathbf{y})}{\mathbf{n}} - \frac{P(z,y)}{n}.$$

Once again, $\mathcal{S}$ is finite, we can let $n \to \infty$ and interchange the limit and the sum to obtain

$$\sum_{x \in \mathcal{S}} \frac{1}{\text{Щ}_x} P(x,y) = \frac{1}{\text{Щ}_y}.$$

We can deduce the case when $\mathcal{S}$ is not finite by expanding on the above reasoning. Suppose that $\mathcal{S}_1$ is a finite subset of $\mathcal{S}$. Then

$$\sum_{x \in \mathcal{S}_1} P^m(z,x) \le 1$$

and so

$$\sum_{x\in\mathcal{S}_1}\frac{\mathfrak{G}_{\mathbf{n}}(\mathbf{z},\mathbf{y})}{\mathbf{n}}\le 1.$$

Since $\mathcal{S}_1$ is finite, we can let $n\to\infty$ and conclude that

$$\sum_{x\in\mathcal{S}_1}\frac{1}{\text{Щ}_x}\le 1.$$

Since this must be true for any finite subset $\mathcal{S}_1\subseteq\mathcal{S}$, it follows that

$$\sum_{x\in\mathcal{S}}\frac{1}{\text{Щ}_x}\le 1.$$

By exactly similar reasoning,

$$\sum_{x\in\mathcal{S}}\frac{1}{\text{Щ}_x}P(x,y)\le\frac{1}{\text{Щ}_y}.$$

Suppose next that there is some y for which the above inequality is strict. Then

$$\begin{aligned}\sum_{y\in\mathcal{S}}\frac{1}{\text{Щ}_y}&>\sum_{y\in\mathcal{S}}\sum_{x\in\mathcal{S}}\frac{1}{\text{Щ}_x}P(x,y)\\&=\sum_{x\in\mathcal{S}}\sum_{y\in\mathcal{S}}\frac{1}{\text{Щ}_x}P(x,y)\\&=\sum_{x\in\mathcal{S}}\frac{1}{\text{Щ}_x}\end{aligned}$$

which is a contradiction. Thus

$$\sum_{x\in\mathcal{S}}\frac{1}{\text{Щ}_x}P(x,y)=\frac{1}{\text{Щ}_y}.$$

Then if we set

$$\lambda=\frac{1}{\sum_x\frac{1}{\text{Щ}_x}}$$

and

$$\pi(x)=\frac{\lambda}{\text{Щ}_x}$$

it follows that $\pi(x)$ is a stationary distribution. But from the first part of the theorem we must then have

$$\pi(x) = \frac{1}{\text{Щ}_x}$$

or $\lambda = 1$, from which

$$\sum_x \frac{1}{\text{Щ}_x} = 1$$

as desired.

∎

There are several consequences of this important result.

32.19. Corollary.

An irreducible Markov Chain is positive recurrent if and only if it has a stationary distribution.

32.20. Example.

Let $\{X_n\}$ be an irreducible birth and death chain on the non-negative integers. When is the chain positive recurrent, null recurrent and transient?

Solution. We have previously seen that such a chain has a stationary distribution if and only if

$$\sum_{x=1}^{\infty} \frac{p_0 \cdots p_{x-1}}{q_1 \cdots q_x} < \infty$$

which is therefore a necessary and sufficient condition for the chain to be positive recurrent.

Similarly, we previously showed that such a chain is transient if and only if

$$\sum_{x=1}^{\infty} \frac{q_1 \cdots q_x}{p_1 \cdots p_x} < \infty.$$

From this, such a chain is null recurrent if and only if both

$$\sum_{x=1}^{\infty} \frac{p_0 \cdots p_{x-1}}{q_1 \cdots q_x} = \infty$$

and

$$\sum_{x=1}^{\infty} \frac{q_1 \cdots q_x}{p_1 \cdots p_x} = \infty.$$

∎

32.21. Corollary.

Let $\{X_n\}$ be an irreducible Markov chain having a finite state space. Then the chain has a unique stationary distribtution.

32.22. Corollary.

Let $\{X_n\}$ be an irreducible, positive recurrent Markov Chain having stationary distribution π. Then with probability one

$$\lim_{n\to\infty} \frac{\mathfrak{N}_n(x)}{n} = \pi(x)$$

for all $x \in \mathcal{S}$.

32.23. Theorem.

Let $C \subseteq \mathcal{S}$ be a closed, irreducible collection of positive recurrent states. Then there is a unique stationary distribution π that vanishes outside of C and, for $x \in C$ is given by

$$\pi(x) = \frac{1}{\text{Ш}_x}.$$

Proof. Since C is closed and irreducible, if $\{X_n\}$ starts in C then it never leaves C. Thus the embedded chain given by

$$Y_n = \begin{cases} X_n & \text{if } X_n \in C \\ 0 & \text{otherwise} \end{cases}$$

has exactly C for its state space. Further the transition function for $\{Y_n\}$ agrees with the transition function for $\{X_n\}$ restricted to C, $\{Y_n\}$ is a closed, irreducible Markov Chain and the theorem applies.

∎

32.24. Corollary.

Let $\mathcal{S}_p$ be the collection of positive recurrent states for a Markov Chain.

(i) If $\mathcal{S}_p = \emptyset$ then the chain does not have a stationary distribution.

(ii) If $\mathcal{S}_p \neq \emptyset$ and is irreducible, then the chain has a unique stationary distribution.

(iii) If $\mathcal{S}_p \neq \emptyset$ but is not irreducible, then the chain has an inifinte number of stationary distributions.

Proof. For (iii), if $\mathcal{S}_p \neq \emptyset$ but is not irreducible, we can find disjoint, closed, irreducible sets C_1 and C_2. If π_1 is a stationary distribution concentrated on C_1 and π_2 is a stationary distribution concentrated on C_2 and if $0 < \lambda < 1$ then

$$\lambda\pi_1 + (1-\lambda)\pi_2$$

is a stationary distribution.

∎

32.25. Example.

Suppose that $\{X_n\}$ is a Markov chain on $\mathcal{S} = \{0, 1, 2, 3, 4, 5\}$ having transition function

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/5 & 2/5 & 1/5 & 0 & 1/5 \\ 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/4 & 0 & 3/4 \end{pmatrix} \end{array}$$

If the chain starts in state 4, how long, on average, before it returns to state 4?

Solution. It suffices to find щ_4. Notice that

$$C_1 = \{0\}$$

and

$$C_3 = \{3, 4, 5\}$$

are closed, irreducible sets, while

$$C_2 = \{1, 2\}$$

is the set of transient states. If we find a stationary distribution concentrated on C_3, then the relationship

$$\pi(x) = \frac{1}{\text{щ}_x}$$

will answer the question.

A stationary distribution on C_3 must satisfy

$$\pi(3) + \pi(4) + \pi(5) = 1$$

and

$$\begin{array}{llll} \pi(3)/6 & +\pi(4)/2 & +\pi(5)/4 & = \pi(3) \\ \pi(3)/3 & & & = \pi(4) \\ \pi(3)/2 & +\pi(4)/2 & +3\pi(5)/4 & = \pi(5) \end{array}$$

This implies that

$$\pi(3) = 1/4 \quad \pi(4) = 1/12 \quad \pi(5) = 2/3.$$

Thus a chain that starts in state four will, on average, return to state four in tweleve steps. ∎

32. Stationary Distributions: Results: Problems.

1. Suppose that $\{X_n\}$ is a Markov chain on $\mathcal{S} = \{0, 1, 2, 3, 4, 5\}$ having transition function

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \!\!\begin{array}{c} \begin{array}{cccccc} 0 & \;\;1 & \;\;2 & \;\;3 & \;\;4 & \;\;5 \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 1/3 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/8 & 0 & 7/8 & 0 \\ 1/4 & 1/4 & 0 & 0 & 1/4 & 1/4 \\ 0 & 0 & 3/4 & 0 & 1/4 & 0 \\ 0 & 1/5 & 0 & 1/5 & 1/5 & 2/5 \end{pmatrix} \end{array}$$

(a) Decompose the state space into transient states and irreduciable, close, recurrent states.
(b) For each closed, irreducible set of states C, find the stationary distribution concentrated on C.
(c) Find $\rho_C(5)$ for each of the closed, irreducible collections of states C.
(c) Find

$$\lim_{n\to\infty} \frac{\boldsymbol{\mathfrak{G}}_{\mathbf{n}}(\mathbf{5},\mathbf{0})}{\mathbf{n}}$$

33. Convergence to the Stationary Distribution

Thus far we have been able to deduce an “averaged”asymptotic relationship between the transition probabilities and the stationary distribution:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^n(x,y) = \lim_{n\to\infty} \frac{\mathfrak{G}\left(\mathbf{x},\mathbf{y}\right)}{\mathbf{n}} = \pi(y).$$

This section will examine the relationship between

$$\lim_{n\to\infty} P^n(x,y)$$

and the stationary distribution $\pi(y)$. This relationship is necessarily more complex as the following two examples show.

33.1. Example.

Let $\{X_n\}$ be the Ehrenfest chain on $\{0,1,2,3\}$, so that the transition function is

$$P = \begin{array}{c c} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

This chain has a stationary distribution given by

$$\pi(0) = 1/8 \quad \pi(1) = 3/8 \quad \pi(2) = 3/8 \quad \pi(3) = 1/8.$$

However, it is impossible to get from a state x back to state x in an odd number of steps, i.e.,

$$P^n(x,x) = 0$$

if n is odd. Indeed, by finding the Jordan Form of P we can calculate P^n directly as

$$\mathrm{P}^{\mathrm{n}} = \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & 1/8 & -1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & -3/8 & 3/8 \end{pmatrix} \begin{pmatrix} (-1)^n & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & (1/3)^n & 0 \\ 0 & 0 & 0 & (-1/3)^n \end{pmatrix} \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & 1/8 & -1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & -3/8 & 3/8 \end{pmatrix}^{-1}$$

where

$$\begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & 1/8 & -1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & -3/8 & 3/8 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -3 & 3 & -1 \\ 1 & 3 & 3 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Thus for n large and even

$$\begin{aligned} P^n &\approx \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & 1/8 & -1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & -3/8 & 3/8 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & 1/8 & -1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & -3/8 & 3/8 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 1/4 & 0 & 3/4 & 0 \\ 0 & 3/4 & 0 & 1/4 \\ 1/4 & 0 & 3/4 & 0 \\ 0 & 3/4 & 0 & 1/4 \end{pmatrix} \end{aligned}$$

while for n large and odd

$$\begin{aligned} P^n &\approx \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & 1/8 & -1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & -3/8 & 3/8 \end{pmatrix} \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & 1/8 & -1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & -3/8 & 3/8 \end{pmatrix}^{-1} . \\ &= \begin{pmatrix} 0 & 3/4 & 0 & 1/4 \\ 1/4 & 0 & 3/4 & 0 \\ 0 & 3/4 & 0 & 1/4 \\ 1/4 & 0 & 3/4 & 0 \end{pmatrix} \end{aligned}$$

From this we can see that the chain exhibits a periodicity of period two asymptotically. ■

33.2. Example.

We can modify the Ehrenfest Chain as so that the periodicity of the preceeding example is not present. In this modification, we first randomly select a ball, remove the ball from its urn, then randomly choose which of the two urns in which to place the ball.

It is easy to verify (see the exercises) that the transition matrix is

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & \;\;1 & \;\;2 & \;\;3 \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/6 & 1/2 & 1/3 & 0 \\ 0 & 1/3 & 1/2 & 1/6 \\ 1/2 & 1/2 & 1/2 & 1/2 \end{pmatrix} \end{array}$$

The stationary distribution is the same as for the unmodified Ehrenfest chain:

$$\pi(0) = 1/8 \quad \pi(1) = 3/8 \quad \pi(2) = 3/8 \quad \pi(3) = 1/8.$$

In this case, however,

$$\mathrm{P}^{\mathrm{n}} = \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & -1/8 & 1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & 3/8 & -3/8 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & (1/3)^n & 0 \\ 0 & 0 & 0 & (2/3)^n \end{pmatrix} \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & -1/8 & 1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & 3/8 & -3/8 \end{pmatrix}^{-1}$$

from which one can show that

$$\lim_{n\to\infty} P^n(x, y) = \pi(y)$$

for all x and y.

∎

In order to understand the asymptotic behavior of $P^n(x, y)$ it is necessary to expand on the intuitive notion of periodicity for chains.

33.3. Definition.

If $J \subseteq \mathbb{N}$ is a collection of positive integers, then an integer d is a **divisor** *of J if n/d is an integer whenever $n \in J$. The* **greatest common divisor** *of a set J is the number*

$$\mathfrak{gcd}\,(J) = \max\{d \in \mathbb{N} \,:\, d \text{ is a divisor of } J\}.$$

For example, if $J = \{4, 8, 12, 16, \cdots\}$ then 1, 2 and 4 are all divisors of J, but $\mathfrak{gcd}\,(J) = 4$.

33.4. Definition.

Let $\{X_n\}$ be a Markov Chain having state space $\mathcal{S}$ and let $x \in \mathcal{S}$. If $P^n(x, x) > 0$ for some n then we define the **period** *of x to be*

$$d_x = \mathfrak{gcd}\,\{n \in \mathbb{N} \,:\, P^n(x, x) > 0\}.$$

33.5. Proposition.

Let $\{X_n\}$ be a Markov Chain having state space $\mathcal{S}$ and let $x, y \in \mathcal{S}$ with d_x and d_y defined (i.e., $P^{n_1}(x,x) > 0$ and $P^{n_2}(y,y) > 0$ for some integers n_1 and n_2). If x leads to y and y leads to x, then $d_x = d_y$.

Proof. We may choose n_1 and n_2 so that

$$P^{n_1}(x,y) > 0 \quad \text{and} \quad P^{n_2}(y,x) > 0.$$

But this implies that

$$P^{n_1+n_2}(x,x) \geq P^{n_1}(x,y)P^{n_2}(y,x) > 0$$

and so d_x must divide $n_1 + n_2$, i.e.,

$$n_1 + n_2 = k_1 d_x$$

for some integer k_1.

Further, if $P^n(y,y) > 0$ then

$$P^{n_1+n+n_2}(x,x) \geq P^{n_1}(x,y)P^n(y,y)P^{n_2}(y,x) > 0$$

and so d_x must be a divisor of $n_1 + n + n_2$, i.e.,

$$n_1 + n + n_2 = k_2 d_x$$

for some integer k_2. But this implies that

$$n = (k_2 - k_1)d_x$$

or that d_x divides n. Since d_y is the largest divisor of such n, it follows that

$$d_x \leq d_y.$$

An exactly symmetric argument shows that $d_y \leq d_x$ and hence that $d_x = d_y$. ∎

33.6. Corollary.

Let $\{X_n\}$ be an irreducible Markov Chain. Then all states have the same period.

33.7. Definition.

Let $\{X_n\}$ be an irreducible Markov Chain having period d. If $d = 1$ we say that the chain $\{X_n\}$ is **aperiodic** *while of $d > 1$ we say that the chain $\{X_n\}$ has period d.*

We are now ready to state the main result of this section.

33.8. Theorem.

Let $\{X_n\}$ be an irreducible, positive recurrent Markov Chain.

(a) If $\{X_n\}$ is aperiodic, then

$$\lim_{n\to\infty} P^n(x,y) = \pi(y)$$

for all $x, y \in \mathcal{S}$.

(b) If $\{X_n\}$ is periodic with period d, then for every pair $x, y \in \mathcal{S}$ there is a number r with $0 \leq r < d$ such that

$$P^n(x,y) = 0$$

unless $n = md + r$ for some integer $m \in \mathbb{N}$. Further,

$$\lim_{m\to\infty} P^{md+r}(x,y) = d\pi(y).$$

Conclusion (a) of the theorem corresponds to the situation in example 33.2, while conclusion (b) corresponds to the situation in example 33.1 with $d = 2$. The proof of this theorem is rather complex and starts with a simple Lemma from number theory. Indeed, in example 33.1, the Theorem asserts directly without appeal to Jordan Forms that for n

large and even

$$P^n \approx \begin{pmatrix} 1/4 & 0 & 3/4 & 0 \\ 0 & 3/4 & 0 & 1/4 \\ 1/4 & 0 & 3/4 & 0 \\ 0 & 3/4 & 0 & 1/4 \end{pmatrix}$$

while for n large and odd

$$P^n \approx \begin{pmatrix} 0 & 3/4 & 0 & 1/4 \\ 1/4 & 0 & 3/4 & 0 \\ 0 & 3/4 & 0 & 1/4 \\ 1/4 & 0 & 3/4 & 0 \end{pmatrix}$$

For example, suppose n is large and even. Since it's not possible to transition from 0 to 1 in an even number of steps, $P^{2n}(0,1) = 0$; similarly, $P^{2n}(0,3) = 0$. Then applying these two observations and the theorem, we obtain the first row of the first matrix above:

$$\begin{aligned} &P^{2n}(0,0) \longrightarrow 2\pi(0) \\ &P^{2n}(0,1) = 0 \\ &P^{2n}(0,2) \longrightarrow 2\pi(2) \\ &P^{2n}(0,1) = 0 \end{aligned}$$

The other rows are similar.

Before proving the Theorem we will need an elementary result from number theory.

33.9. Lemma.

Let $J \subseteq \mathbb{N}$ be a collection of positive integers and suppose that
(a) $\mathfrak{gcd}\,(J) = 1$*; and*
(b) whenever $m, n \in J$ it follows that $n + m \in J$.
Then there is a number N such $n \in J$ for all $n \geq N$.

Proof. We begin by showing that J contains two consecutive integers. To do this, suppose the contrary. We can then choose a $k \geq 2$ so that any two integers in J differ by at least k. Moreover, we can take k to be the least such number and can find two specific integers, n_1 and $n_1 + k$ that differ by exactly k.

Since $k \geq 2$ there must be an integer $n \in J$ so that k is not a divisor of n. Thus we can write

$$n = mk + r$$

where $0 < r < k$.

Now assumption (b) of the lemma implies that if $j \in J$ then any multiple of j must also be in J. Thus we may conclude that

$$(m+1)(n_1+k) \in J$$

and

$$n + (m+1)n_1 \in J.$$

But this then implies that

$$\begin{aligned}(m+1)(n_1+k) - [n + (m+1)n_1] &= k + mk - n \\ &= k - r \\ &< k\end{aligned}$$

which says that we have found two elements of J closer to each other than k. But we selected k to be the least separation between elements of J, so this is a contradiction.

Thus we may conclude that J contains two consecutive integers, say m_1 and $m_1 + 1$. Set $N = m_1^2$ and let $n \geq N$. We can choose an integer m and a remainder r ($0 \leq r < m_1$) so that

$$n - N = n - m_1^2 = mm_1 + r.$$

This implies that

$$\begin{aligned}n &= r + mm_1 + m_1^2 \\ &= r(m_1+1) + (m_1 - r + m)m_1 \\ &\in J\end{aligned}$$

using assumption (b) of the Lemma. Thus if $n \geq N$ then $n \in J$.

■

Proof of Theorem 33.8.

We begin with the aperiodic case. For arbitrary $a \in \mathcal{S}$ set

$$J = \{n \, : \, P^n(a,a) > 0\}.$$

Since the chain is assumed to be aperiodic, $\mathfrak{gcd}\,(J) = 1$. Further, if $m, n \in J$ then

$$P^{n+m}(a,a) \geq P^n(a,a)P^m(a,a) > 0$$

and so $n+m \in J$. By the Lemma, there is an intger N so that if $n \geq N$ then $n \in J$, i.e.,

$$n \geq N \quad \Longrightarrow \quad P^n(a,a) > 0.$$

Now let $x, y \in \mathcal{S}$. Since the chain is irreducible, there are integers n_1 and n_2 so that $P^{n_1}(x,a) > 0$ and $P^{n_2}(a,y) > 0$. Thus if $n \geq N$

$$P^{n_1+n+n_2}(x,y) \geq P^{n_1}(x,a)P^n(a,a)P^{n_2}(a,y) > 0.$$

Thus for any pair $x, y \in \mathcal{S}$ there is an n_0 such that

$$n \geq n_0 \quad \Longrightarrow \quad P^n(x,y) > 0.$$

Now set $\mathcal{S} \times \mathcal{S} = \{(x,y) \, : \, x, y \in \mathcal{S}\}$ and define a new chain $\{(X_n, Y_n)\}$ having state space $\mathcal{S} \times \mathcal{S}$ and transition function

$$P_2\big((x,y),(u,v)\big) = P(x,u)P(y,v).$$

Clearly the chains $\{X_n\}$ and $\{Y_n\}$ taken separately are Markov Chains having transition function P the same as the original chain and the successive transitions of the chains are independent of one another.

We next show that the new chain $\{(X_n, Y_n)\}$ is aperiodic, irreducible and positive recurrent. If we select any pair of states $\big((x,y),(u,v)\big) \in \mathcal{S} \times \mathcal{S}$ we can find an n_0 so that

$$n > n_0 \quad \Longrightarrow \quad P^n(x,u) > 0 \quad \text{and} \quad P^n(y,v) > 0.$$

This implies that

$$P_2\big((x,y),(u,v)\big) = P^n(x,u)P^n(y,v) > 0$$

for all $n \geq n_0$, and thus the chain is irreducible and aperiodic.

In order to show the chain is positive recurrent it will suffice to show that the chain has a stationary distribution. If π is the stationary distribution for the original chain, we can define

$$\pi_2\big((x,y)\big) = \pi(x)\pi(y).$$

Then

$$\begin{aligned}\sum_{(x,y)\in\mathcal{S}\times\mathcal{S}} \pi_2\big((x,y)\big)P_2\big((x,y),(u,v)\big)& \\ &= \sum_{x\in\mathcal{S}}\sum_{y\in\mathcal{S}} \pi(x)\pi(y)P(x,u)P(y,v) \\ &= \pi(u)\pi(v) \\ &= \pi_2\big((u,v)\big).\end{aligned}$$

This shows that π_2 is a stationary distribution and hence that $\{(X_n, Y_n)\}$ is positive recurrent.

Next set

$$T = \min\{n > 0 \,:\, X_n = Y_n\}.$$

For $(a, a) \in \mathcal{S} \times \mathcal{S}$ set

$$T_{(a,a)} = \min\{n > 0 \,:\, (X_n, Y_n) = (a, a)\}.$$

Note that $T_{(a,a)} < \infty$ with probability one. Since $T \leq T_{(a,a)}$, it follows that $T < \infty$ with probability one.

Next we observe that

$$\mathfrak{Pr}\,(X_n = y \text{ and } T \leq n\} = \mathfrak{Pr}\,(Y_n = y \text{ and } T \leq n\}$$

for all $y \in \mathcal{S}$. This is intuitively reasonable since the two chains are, with probability one, the same for large n. We leave a precise argument to the exercises.

Now observe that for $y \in \mathcal{S}$

$$\begin{aligned}\mathfrak{Pr}\,(X_n = y) &= \mathfrak{Pr}\,(X_n = y,\ T \leq n) + \mathfrak{Pr}\,(X_n = y,\ T > n)\\ &= \mathfrak{Pr}\,(Y_n = y,\ T \leq n) + \mathfrak{Pr}\,(X_n = y,\ T > n)\\ &\leq \mathfrak{Pr}\,(Y_n = y) + \mathfrak{Pr}\,(T > n)\end{aligned}$$

Similarly,

$$\mathfrak{Pr}\,(Y_n = y) \leq \mathfrak{Pr}\,(X_n = y) + \mathfrak{Pr}\,(T > n)$$

which implies

$$\big|\,\mathfrak{Pr}\,(X_n = y) - \mathfrak{Pr}\,(Y_n = y)\big| \leq \mathfrak{Pr}\,(T > n).$$

But since T is finite with probabilty one, it follows that

$$\lim_{n\to\infty} \big|\,\mathfrak{Pr}\,(X_n = y) - \mathfrak{Pr}\,(Y_n = y)\big| = 0$$

with probability one.

Now let $x \in \mathcal{S}$ and let the initial distribution of $\{(X_n, Y_n)\}$ satisfy

$$\mathfrak{Pr}\,(X_0 = x) = 1 \quad \text{and} \quad \mathfrak{Pr}\,(Y_0 = y) = \pi(y)$$

for all $y \in \mathcal{S}$. From this $\mathfrak{Pr}\,(X_n = y) = P^n(x, y)$ while $\mathfrak{Pr}\,(Y_n = y) = \pi(y)$. Thus

$$\begin{aligned}0 &= \lim_{n\to\infty} \big|\,\mathfrak{Pr}\,(X_n = y) - \mathfrak{Pr}\,(Y_n = y)\big|\\ &= \lim_{n\to\infty} \big|P^n(x, y) - \pi(y)\big|\end{aligned}$$

with probability one, which completes the proof of the aperiodic case.

Before proceeding with the periodic case, we note that we have incidentally proven the following corollary.

33.10. Corollary.

Let $\{X_n\}$ be a Markov chain with state space $\mathcal{S}$ and let $C \subseteq \mathcal{S}$ be a closed, irreducible and positive recurrent set of states. Then there is a stationary distribution concentrated on C and, for $x, y \in C$

$$\lim_{n\to\infty} P^n(x,y) = \pi(y) = \frac{1}{\text{Ш}_y}.$$

Proof of Theorem 33.8 completed.

Set $Y_m = X_{md}$ for $m \in \mathbb{N}$, so $\{Y_m\}$ is a Markov Chain having transition function $Q = P^d$. Further, if $y \in \mathcal{S}$ then

$$\begin{aligned} \mathfrak{gcd}\,\{m > 0 \,:\, Q^m(y,y) > 0\} &= \mathfrak{gcd}\,\{m > 0 \,:\, P^{md}(y,y) > 0\} \\ &= \frac{1}{d}\,\mathfrak{gcd}\,\{m > 0 \,:\, P^d(y,y) > 0\} \\ &= 1 \end{aligned}$$

so $\{Y_m\}$ is an aperiodic chain.

Now if Ш_y is the expected return time to y for $\{X_n\}$ and $\tilde{\text{Ш}}_y$ is the expected return time to y for $\{Y_n\}$ then

$$\tilde{\text{Ш}}_y = \frac{\text{Ш}_y}{d}$$

and

$$\lim_{n\to\infty} Q^m(y,y) = \frac{1}{\tilde{\text{Ш}}_y} = \frac{d}{\text{Ш}_y} = d\pi(y).$$

This in turn implies that

$$\lim_{n\to\infty} P^{md}(y,y) = d\pi(y) \tag{33.1}$$

for all $y \in \mathcal{S}$.

Now let $x, y \in \mathcal{S}$ and set

$$r_1 = \min\{n > 0 \,:\, P^n(x,y) > 0\}.$$

We assert that

$$P^n(x,y) > 0 \quad \iff \quad n - r_1 = kd \quad \exists k \in \mathbb{N}.$$

To see this, choose n_1 so that $P^{n_1}(y,x) > 0$, which implies that

$$P^{n_1+r_1}(y,y) \geq P^{n_1}(y,x)P^{r_1}(x,y) > 0$$

so that

$$r_1 + n_1 = k_1 d \quad \exists k_1 \in \mathbb{N}.$$

If $P^n(x,y) > 0$ then the same argument shows that

$$n + n_1 = k_2 d \quad \exists k_2 \in \mathbb{N}.$$

This implies that

$$\begin{aligned} n - r_1 &= n + n_1 - n_1 - r_1 \\ &= k_2 d - k_1 d \\ &= (k_2 - k_1)d \end{aligned}$$

so $n - r_1$ is a multiple of d as desired.

Next we divide r_1 by d to obtain an m_1 and an r ($0 \leq r < d$) so that

$$r_1 = m_1 d + r.$$

But then

$$\begin{aligned} P^n(x,y) > 0 &\Longrightarrow n - r_1 = kd \\ &\Longrightarrow n - md - r = kd \\ &\Longrightarrow n = (k+m)d + r \end{aligned}$$

From this we conclude that $P^n(x,y) > 0$ only if $n = md + r$ for some $m \in \mathbb{N}$.

Finally,

$$P^{md+r}(x,y) = \sum_{k=0}^{m} P_x(T_y = kd + r)P^{(m-k)d}(y,y).$$

If we set

$$\gamma_m(k) = \begin{cases} P^{(m-k)d}(y,y) & \text{if } 0 \leq k \leq m \\ 0 & \text{otherwise} \end{cases}$$

then for each fixed k equation (33.1) implies

$$\lim_{m\to\infty} \gamma_m(k) = d\pi(y).$$

Thus we can apply the Bounded Convergence Theorem to conclude that

$$\begin{aligned}\lim_{m\to\infty} P^{md+r}(x,y) &= d\pi(y)\sum_{k=0}^{\infty} P_x(T_y = kd + r)\\ &= d\pi(y)P_x(T_y < \infty)\\ &= d\pi(y)\end{aligned}$$

since $P_x(T_y < \infty) = 1$.

■

33. Convergence to the Stationary Distribution: Problems.

1. Verify the formula for P^n is example 33.1 by showing that

$$\mathrm{P} = \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & 1/8 & -1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & -3/8 & 3/8 \end{pmatrix} \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & -1/3 \end{pmatrix} \begin{pmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ -1/8 & 1/8 & 1/8 & -1/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ -1/8 & 1/8 & -3/8 & 3/8 \end{pmatrix}^{-1}$$

2. With P as in example 33.2, verify that

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/6 & 1/2 & 1/3 & 0 \\ 0 & 1/3 & 1/2 & 1/6 \\ 1/2 & 1/2 & 1/2 & 1/2 \end{pmatrix} \end{array}$$

is the transition matrix, find the stationary distribution and conclude that

$$\lim_{n\to\infty} P^n(x,y) = \pi(y)$$

for all x and y.

3. Let $(\Omega, \mathcal{E}, \mathfrak{Pr})$ be a probability space and suppose that all the sets in this problem are in $\mathcal{E}$.

(a) If $\{D_n\}$ are disjoint and $\mathfrak{Pr}(C \big| D_n) = p$ independently of n, then

$$\mathfrak{Pr}(C \big| \bigcup_n D_n) = p.$$

(b) If $\{C_n\}$ are disjoint, then

$$\mathfrak{Pr}(\bigcup_n C_n \big| D) = \sum_n \mathfrak{Pr}(C_n \big| D).$$

(c) If $\{E_n\}$ are disjoint and $\cup_n E_n = \Omega$, then

$$\mathfrak{Pr}(C \big| D) = \sum_n \mathfrak{Pr}(E_n \big| D)\, \mathfrak{Pr}(C \big| E_n \cap D).$$

(d) If $\{C_n\}$ are disjoint and $\mathfrak{Pr}(A \big| C_n) = \mathfrak{Pr}(B \big| C_n)$ for all n, then

$$\mathfrak{Pr}(A \big| \bigcup_n C_n) = \mathfrak{Pr}(B \big| \bigcup_n C_n).$$

4. Let $\{(X_n, Y_n)\}$ and T be as in the proof of 33.8.
(a) For $1 \le m \le n$ and for $z \in \mathcal{S}$ show that

$$\mathfrak{Pr}\left(X_n = y \middle| T = m,\ \ X_m = Y_m = z\right) = \mathfrak{Pr}\left(Y_n = y \middle| T = m,\ \ X_m = Y_m = z\right).$$

(b) Show that

$$\{T \le n\} = \bigcup_m \bigcup_z \{T = m,\ \ X_m = Y_m = z\}.$$

(c) Conclude that

$$\mathfrak{Pr}\left(X_n = y \text{ and } T \le n\right\} = \mathfrak{Pr}\left(Y_n = y \text{ and } T \le n\right\}.$$

5. Let $\{X_n\}$ be a Markov Chain on $\{0, 1, 2\}$ having transition matrix

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix} \end{array}.$$

(a) Show that the chain is irreducible.
(b) Find the period.
(c) Find the stationary distribution.

6. Consider a Markov Chain on $\{0, 1, 2, 3, 4\}$ having transition matrix

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & 3/4 \\ 0 & 0 & 0 & 1/4 & 3/4 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

(a) Show that the chain is irreducible.
(b) Find the period.
(c) Find the stationary distribution.
(d) Estimate $P^{7294}(0, x)$ for $x = 0, 1, 2, 3, 4$.

34. Poisson Processes

Thus far we have considered random processes $\{X_n\}$ where $n \in \mathbb{N}$. In particular the evolution through time is measured in discrete increments n. In many situations it is far more natural to allow time to vary continuously even while the state space remains discrete. For example, counting the number of cars passing through an interchange has a discrete state space $\mathbb{N}$ but the cars could arrive at any time $t \in [0, \infty)$. This is an example of a particular kind of stochastic process. Intuitively, if $X(t)$ is the number of cars that have passed by time t, then one would expect that $X(t)$ would start out equalling zero, remain so for a while, then 'jump' to $X(t) = 1$ when the first car passes at some time τ_1. The process then stays equal to one for a period until some time τ_2 when the second car passes, and so on. For these kinds of process there are actually then two random quantities: the value of $X(t)$ and the sequence of times $0 \leq \tau_1 < \tau_2 < \tau_3 < \cdots$ when the jumps occur.

While a chain X_n is defined for all n, it is possible that a jump process $X(t)$ could *explode*, i.e., that $X(t) \to \infty$ while $t_n \to t_\infty < \infty$. For example if the time between arrivals is summable

$$\tau_{n+1} - \tau_n = 2^{-n}$$

then $X(t)$ is only defined for $0 \leq t < 2$. Generally we will assume that the processes we study will be *non-explosive* or defined for all $t > 0$.

Another slightly exotic behavior that a jump process $X(t)$ could exhibit is to enter some state x_0 then instantly jump to another state instead of staying there for some period of time. In most applications of interest jump processes instead 'rest 'in a state x_0 before jumping to another state. More mathematically, the times of succeeding jumps are strictly increasing:

$$\tau_1 < \tau_2 < \tau_3 < \cdots.$$

Finally, it is of course possible that a jump process could enter a state x_0 and never leave it. We have encountered this type of state before: if this happens with probability one then x_0 is an absorbing state.

We will begin our more formal study of jump processes in the next section. In this section we will study a particular kind of jump process, one that counts the number of events that have occured. It turns out that many of the important properties of jump processes are illuminated by this particular process, the Poisson Process. We begin with some simple definitions.

34.1. Definition. Counting Processes.

A collection of random variables $\{X(t)\}_{t\in[0,\infty)}$ *defined on a probability space* $(\Omega, \mathcal{E}, \mathfrak{Pr})$ *is a* **counting process** *if*

(i) $X(0) = 0$*;*
(ii) $X(t)$ *takes on only integer values;*
(iii) $\mathfrak{Pr}(X(s) \leq X(t)) = 1$ *if* $s \leq t$*.*

34.2. Definition. Independent Increments.

The process $\{X(t)\}$ *is said to have* **independent increments** *if whenever* $s_1 < t_1 \leq s_2 < t_2 \leq \cdots \leq s_n < t_n$ *then the random variables*

$$X(t_1) - X(s_1), X(t_2) - X(s_2), \cdots, X(t_n) - X(s_n)$$

are independent, i.e., if the number of events that occur in disjoint intervals are independent.

34.3. Definition. Stationary Increments.

The process $\{X(t)\}$ *is said be* **time homogeneous** *(or have* **stationary increments***) if for all* $s, t \geq 0$

$$X(t+s) - X(t) \quad \text{and} \quad X(s) - X(0)$$

have the same distributions, i.e., if the number of events that occur in the interval $(t, t+s]$ *depends only on the length of the interval* s*.*

Intuitively, if $\{X(t)\}$ is a counting process then $X(t)$ counts the number of random events that have occured up to time t. For example, $X(t)$ might be counting the number of vehicles passing a checkpoint on a road, or the total number of calls that have arrived at a telephone exchange, or the number of emissions from a radioactive substance or the number of bacteria in a culture.

In each of the examples in the previous paragraph it is reasonable to suppose that there is an "average" arrival rate λ and so, on any fixed interval $[t, t+h)$, one would expect that the number of arrivals would be

$$X(t+h) - X(t) = h\lambda$$

If one assumes that the arrivals occur independently and cannot happen simultaneously, the result is a Poisson Process.

The particular mathematical assumptions we will make on $X(t)$ are the following.

34.4. Definition.

A counting process $\{X(t)\}$ is said to be a **Poisson Process** *if there is a $\lambda > 0$ such that for any $t \geq 0$ and any $h > 0$*

(a)

$$\lim_{h \to 0} \frac{\mathfrak{Pr}\left(X(t+h) - X(t) = 1\right)}{h} = \lambda;$$

(b)

$$\lim_{h \to 0} \frac{1 - \mathfrak{Pr}\left(X(t+h) - X(t) = 0\right)}{h} = \lambda;$$

(c) For any $k \geq 2$,

$$\lim_{h \to 0} \frac{\mathfrak{Pr}\left(X(t+h) - X(t) = k\right)}{h} = 0.$$

Assumption (a) formalizes the notion of an average arrival time. Assumption (c) formalizes the notion that events cannot occur simultaneously.

Assumption (b) is a slight strengthening of (a) and (c) taken together. Notice first that the complementary form of (b) is

$$\lim_{h \to 0} \frac{\mathfrak{Pr}\left(X(t+h) - X(t) \geq 1\right)}{h} = \lambda$$

Clearly

$$\begin{aligned}
\lim_{h \to 0} \frac{\mathfrak{Pr}\left(X(t+h) - X(t) \geq 1\right)}{h} &= \lim_{h \to 0} \sum_{k=1}^{\infty} \frac{\mathfrak{Pr}\left(X(t+h) - X(t) = k\right)}{h} \\
&= \lim_{h \to 0} \frac{\mathfrak{Pr}\left(X(t+h) - X(t) = 1\right)}{h} + \lim_{h \to 0} \sum_{k=2}^{\infty} \frac{\mathfrak{Pr}\left(X(t+h) - X(t) = k\right)}{h} \\
&= \lambda + \lim_{h \to 0} \sum_{k=2}^{\infty} \frac{\mathfrak{Pr}\left(X(t+h) - X(t) = k\right)}{h}
\end{aligned}$$

applying (a). Now if the limit and the infinite sum could be interchanged, then second term would be zero by (c). Since it is not possible to interchange infinite processes in general, it is necessary to make the assumption (b) explicitly.

It is possible to test, by gathering data, whether or not the above assumptions are reasonable in any particular setting. In each of the examples noted above there is a considerable body of evidence to support making exactly these assumptions.

This small set of assumptions is sufficient to enable us to deduce the distribution of $X(t)$ for each t. For convenience we begin with a definition.

34.5. Definition.

For $k = 0, 1, \cdots$ set

$$P_k(t) = \mathfrak{Pr}\,(X(t) = k)$$

34.6. Lemma.

With $P_0(t)$ defined as above, $P_0(t) = e^{-\lambda t}$

Proof. Using the fact that $X(t)$ is non-decreasing,

$$\begin{aligned} P_0(t+h) &= \mathfrak{Pr}\,(X(t+h) = 0) \\ &= \mathfrak{Pr}\,\left(X(t) = 0, \quad X(t+h) - X(t) = 0\right) \\ &\quad \text{(since } X(0) = 0 \text{ by assumption)} \\ &= \mathfrak{Pr}\,\left(X(t) - X(0) = 0, \quad X(t+h) - X(t) = 0\right) \\ &= \mathfrak{Pr}\,\left(X(t) - X(0) = 0\right)\mathfrak{Pr}\,(X(t+h) - X(t) = 0) \\ &= P_0(t)\,\mathfrak{Pr}\,\left(X(t+h) - X(t) = 0\right). \end{aligned}$$

Thus,

$$\begin{aligned} \lim_{h\to 0} \frac{P_0(t+h) - P_0(t)}{h} &= \lim_{h\to 0} P_0(t) \frac{\mathfrak{Pr}\,\left(X(t+h) - X(t) = 0\right) - 1}{h} \\ &= -\lambda P_0(t) \end{aligned}$$

by assumption (b) for Poisson Processes. This implies that

$$P_0'(t) = -\lambda P_0(t)$$

Since $P_0(0) = 1$, it follows that

$$P_0(t) = e^{-\lambda t}$$

$$\blacksquare$$

34.7. Theorem.

For each t

$$\mathfrak{Pr}\left(X(t) = k\right) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

i.e., $X(t)$ *has a Poisson distribution.*

Proof. As in the lemma, first compute

$$\begin{aligned}
P_k(t+h) &= \mathfrak{Pr}\left(X(t+h) = k\right) \\
&= \sum_{i=0}^{k} \mathfrak{Pr}\left(X(t) = k-i, X(t+h) - X(t) = i\right) \\
&= \sum_{i=0}^{k} \mathfrak{Pr}\left(X(t) = k-i\right) \mathfrak{Pr}\left(X(t+h) - X(t) = i\right) \\
&= \sum_{i=0}^{k} P_{k-i}(t)\, \mathfrak{Pr}\left(X(t+h) - X(t) = i\right)
\end{aligned}$$

Now re-write the sum segregating the first two terms:

$$\begin{aligned}
&P_k(t+h) = \\
&\quad = P_k(t)\, \mathfrak{Pr}\left(X(t+h) - X(t) = 0\right) + \\
&\quad + P_{k-1}(t)\, \mathfrak{Pr}\left(X(t+h) - X(t) = 1\right) + \\
&\quad + \sum_{i=2}^{k} P_{k-i}(t)\, \mathfrak{Pr}\left(X(t+h) - X(t) = i\right)
\end{aligned}$$

So

$$\begin{aligned}\lim_{h\to 0}\frac{P_k(t+h)-P_k(t)}{h} &= \\ &= \lim_{h\to 0} P_k(t)\frac{\mathfrak{Pr}\left(X(t+h)-X(t)=0\right)-1}{h}+ \\ &+ \lim_{h\to 0} P_{k-1}(t)\frac{\mathfrak{Pr}\left(X(t+h)-X(t)=1\right)}{h}+ \\ &+ \lim_{h\to 0}\sum_{i=2}^{k}\frac{P_{k-i}(t)\,\mathfrak{Pr}\left(X(t+h)-X(t)=i\right)}{h}\end{aligned}$$

Now applying (b)^c to the first term, (a) to the second and (c) to the third, we see that

$$P_k'(t) = -\lambda P_k(t) + \lambda P_{k-1}(t) + 0$$

or

$$P_k'(t) = \lambda\big(P_{k-1}(t) - P_k(t)\big) \qquad k = 1, 2, 3, \cdots \tag{34.1}$$

Thus applying the Lemma to the case $k = 1$ gives

$$\begin{aligned}P_1'(t) &= \lambda\big(P_0(t) - P_1(t)\big) \\ &= \lambda e^{-\lambda t} - \lambda P_1(t)\end{aligned}$$

Solving the differential equation gives

$$P_1(t) = \lambda t e^{-\lambda t}$$

Induction on equation (34.1) gives the result.

∎

Because we have assumed that the process has stationary increments, we can actually conclude the following.

34.8. Corollary.

For each $0 \le s < t$

$$\mathfrak{Pr}\left(X(t) = k \middle| X(s) = 0\right) = \frac{(\lambda(t-s))^k e^{-\lambda(t-s)}}{k!} \tag{34.2}$$

Indeed, (34.2) coupled with independent and stationary incements are sufficient to establish (a), (b) and (c) of definition 34.4, and so provides both a necessary and sufficient condition for a Poisson Process.

34.9. Theorem.

Let $\{X(t)\}$ be a counting process that having both independent and stationary increments. Suppose that for each $0 \leq s < t$

$$\mathfrak{Pr}\left(X(t) = k \middle| X(s) = 0\right) = \frac{(\lambda(t-s))^k e^{-\lambda(t-s)}}{k!} \tag{34.3}$$

Then $\{X(t)\}$ is a Poisson Process.

The proof simply involves verifying the limits and is left to the exercises.

The Poisson Process has another random quantity in addition to the number of events $X(t)$ that occured prior to time t. This second quantity involves the waiting time between events, or the arrival time of the first jump from the present state.

34.10. Definition.

For fixed $t > 0$, the **arrival time** *of the first jump is the random variable*

$$\tau_1(t) = +\infty \qquad \text{if } X(t+s) = X(t) \text{ for all } s \geq 0$$

and

$$\tau_1(t) = \inf\{s > 0 : X(t+s) \neq X(t)\}$$

otherwise.

34.11. Proposition.

If $\{X(t)\}$ is a time-homogeneous counting process, then the distribution of $\tau_1(t)$ is independent of t.

Proof. Note that $\tau_1(t) > u$ for some $u > 0$ and if an only if

$$X(t+s) = X(t) \qquad 0 \leq s \leq u.$$

Thus applying the time homogeneity of $\{X(t)\}$

$$\begin{aligned}\mathfrak{Pr}\left(\tau_1(t) > u\right) &= \mathfrak{Pr}\left(X(t+s) = X(t)\right) \quad \text{for} \quad 0 \le s \le u) \\ &= \mathfrak{Pr}\left(X(s) = X(0)\right) \quad \text{for} \quad 0 \le s \le u) \\ &= \mathfrak{Pr}\left(\tau_1(0) > u\right)\end{aligned}$$

showing that $\tau_1(t)$ and $\tau_1(0)$ have the same distribution for all t.

∎

From this we are able to conclude that the arrival time between events in a Poisson Process is exponentially distributed.

34.12. Corollary.

If $\{X(t)\}$ is the Poisson Process, then $\tau_1(t)$ is exponentially distributed with parameter λ.

Proof. Since $X(t)$ is non-decreasing and $X(0) = 0$

$$\begin{aligned}\mathfrak{Pr}\left(\tau_1(t) > u\right) &= \mathfrak{Pr}\left(\tau_1(0) > u\right) \\ &= \mathfrak{Pr}\left(X(s) = 0 \quad \text{for} \quad 0 \le s \le u\right) \\ &= e^{-\lambda u}\end{aligned}$$

applying Theorem 34.6

∎

This important corollary holds in more general settings as we shall see in the next section on Markov Jump Processes. It also provides another way of viewing the Poisson Process.

34.13. Theorem.

Let $\{Y_i\}$ be independent and identically distributed exponential random variables having paramenter λ and let $S_i = Y_1 + \cdots + Y_i$. Define

$$N(t) = \textit{ the \# of } S_i\textit{'s that are less than or equal to } t.$$

Then $\{N(t)\}$ is a Poisson Process.

Note that $N(t) = n$ if and only if

$$Y_1 + \cdots + Y_n \leq t < Y_1 + \cdots + Y_n + Y_{n+1}.$$

Thus $N(t)$ counts the number of occurances of a random event (such as arrivals in a queue) where the time between the n^{th} and $(n+1)^{st}$ event is exponentially distributed. We leave the proof of this result to the problems as well.

1. Let $\{X(t)\}$ be a Poisson Process.

(a) For $0 < t < s$ show that

$$\mathfrak{Pr}\left(X(t) = 1 \middle| X(s) = k\right) = k\left(\frac{t}{s}\right)\left(1 - \frac{t}{s}\right)^{k-1}$$

(b) For $0 < t < s$ show that

$$\mathfrak{Pr}\left(X(t) = m \middle| X(s) = k\right) = \binom{k}{m}\left(\frac{t}{s}\right)^{m}\left(1 - \frac{t}{s}\right)^{k-m}$$

2. Let $\{X(t)\}$ be a process that having both independent and stationary increments. Suppose that for each $0 \leq t < s$

$$\mathfrak{Pr}\,\Big(X(s-t) = k\Big) = \frac{(\lambda(s-t)^{k}e^{-\lambda(s-t)}}{k!}$$

Then $\{X(t)\}$ is a Poisson Process.

3. Let $\{X(t)\}$ be a Poisson process with paramter λ. Suppose that each arrival is registered by a sensor with probability p independent of other arrivals. Let $\{Y(t)\}$ count the number of registered arrivals. Show that $\{Y(t)\}$ is a Poisson process with parameter $p\lambda$. *Hint: consider*

$$\mathfrak{Pr}\left(Y(t) = k\right) = \sum_{m=0}^{\infty}\mathfrak{Pr}\left(Y(t) = k \; and \, X(t) = m + k\right)$$

and use Corollary 34.8 and Theorem 34.9.)

4. Let $\{Y_i\}$ be independent and identically distributed exponential random variables having paramenter λ and let $S_i = Y_1 + \cdots + Y_i$. Define

$$N(t) = \text{ the \# of } S_i\text{'s that are less than or equal to } t.$$

(a) Show that

$$\mathfrak{Pr}\left(N(t) = k\right) = \frac{(\lambda t)^{k}e^{-\lambda t}}{k!}$$

(b) Show that for $0 \leq r < t$

$$\mathfrak{Pr}\left(N(r) = m \text{ and } N(t) - N(r) = k\right) = \frac{(\lambda r)^{m}e^{-\lambda r}}{m!}\,\frac{[\lambda(t-r)]^{k}e^{-\lambda(t-r)}}{k!}.$$

(c) Use (b) to show that for $0 \leq r < t$

$$\mathfrak{Pr}\left(N(t) - N(r) = k \middle| N(r) = m\right) = \frac{[\lambda(t-r)]^k e^{-\lambda(t-r)}}{k!}$$

(d) Use (b) to show that for $0 \leq r < t$

$$\mathfrak{Pr}\left(N(r) - N(0) = m \,\text{and}\, N(t) - N(r) = k\right) = \mathfrak{Pr}\left(N(r) - N(0) = m\right) \mathfrak{Pr}\left(N(t) - N(r) = k\right)$$

which verifies independent increments for adjacent intervals.

(e) Use (d) to establish that if $s_1 < t_1 \leq s_2 < t_2 \leq \cdots \leq s_n < t_n$ then the random variables

$$N(t_1) - N(s_1), N(t_2) - N(s_2), \cdots, N(t_n) - N(s_n)$$

are independent. *Hint: first establish the conclusion for a pair of non-adjacent intervals by inserting additional intervals so that all are adjacent.*

35. Markov Pure Jump Processes

In general a stochastic process is a collection of random variables $\{X(t)\}$ defined on a common probability space $(\Omega, \mathcal{E}, \mathfrak{P}\mathfrak{r})$ and indexed by a real parameter $t \in \mathbb{R}$. In order to have a useful structure, we must of course make additional assumptions. Thus we assume that each of the random variables $X(t)$ shares the same range or state space $\mathcal{S}$, i.e., for each t

$$X(t) : \Omega \to \mathcal{S}.$$

35.1. Definition.

The phase space $\mathcal{T} \subseteq \mathbb{R}$ is the collection of indices $\mathcal{T} = \{t\}$. Where $\mathcal{T}$ is countable the resulting process is called a chain.

35.2. Definition. Jump Processes.

In this section we consider the case where the phase space $\mathcal{T}$ is the non-negative real numbers $[0, \infty)$ but retain the assumption that state space $\mathcal{S}$ is discrete.

The result is called a *jump process* since $X(t)$ stays constant for a time, then 'jumps' to a new value, stays there for a period of time, then jumps again.

There are many examples of jump processes. One of common experience would be the server queues in a grocery store. In such a queue, we might let $X(t)$ denote the number of persons standing in line. Persons arrive in the line at random times, so the random variable describing arrival times would be continuous. Similarly, customers leave the queue at random times – perhaps dependent upon the size or complexity of the contents of their shopping cart – so the serving times are likewise continuous random variables. Thus at any time there are a finite number of customers in the line, but the number could and does change according to random continuous random variables, namely the arrival times and serving times. This simple situation (a *birth and death* process) subsumes many other examples and, with simple assumptions on how the arrivals and departures occur, can be readily modeled.

As with chains, the Markov property is an essential assumption in analyzing jump processes. The Markov property, which roughly says that the future depends only on the present and not the past history of the system, is also a reasonable one in most applications.

35.3. Definition. The Markov Property.

A jump process $X(t)$ is said to satisfy the $Markov\ Property$ if for each

$$0 \le s_1 \le s_2 \le \cdots \le s_n \le s \le t \in \mathcal{T} \text{ and}$$
$$x_1, x_2, \cdots, x_n, x, y \in \mathcal{S}$$

it is the case that

$$\begin{aligned}&\mathfrak{Pr}\left(X(t) = y \middle| X(s_1) = x_1, X(s_2) = x_2, \cdots, X(s_n) = x_n, X(s) = x\right)\\ &\quad = \mathfrak{Pr}\left(X(t) = y \middle| X(s) = x\right).\end{aligned}$$

In general the value of

$$\mathfrak{Pr}\left(X(t+s) = y \middle| X(t) = x\right)$$

could depend on both $t+s$ and t as well as the states x and y. If this probability depends only the states x and y and on the $elapsed$ time $t-s$ then the process is said to be $time\text{-}homogeneous$. More precisely:

35.4. Definition. Time Homogeneous Processes.

A Markov jump process $\{X(t)\}$ is said to be time homogeneous (or have stationary increments) if

$$X(t+s) - X(t) \quad \text{and} \quad X(s) - X(0)$$

are identically distributed for all $s, t \ge 0$.

In this case the future state of the process depends only on the present state and not on the time at which that state is attained. Thus a time-homogenous Markov process can be thought of 'starting over' again for each t and the process $\tilde{X}(t) = X(t+s)$ has essentially the same behaviors as the process $X(t)$. We will suppose that all the processes we consider are time-homogeneous without further comment.

More formally, the following version of the Markov property is equivalent to the above and is the one we will use.

$$\begin{aligned}&\mathfrak{Pr}\left(X(t) = y \middle| X(s_1) = x_1, X(s_2) = x_2, \cdots, X(s_n) = x_n, X(s) = x\right)\\ &\quad = \mathfrak{Pr}\left(X(t-s) = y \middle| X(0) = x\right).\end{aligned}$$

35.5. Definition.

The transition probabilities can thus be defined as

$$p_{(x,y)}(t) = \mathfrak{Pr}\left(X(t) = y \middle| X(0) = x\right).$$

Because of stationary increments, for all $s, t \geq 0$

$$p_{(x,y)}(t) = \mathfrak{Pr}\left(X(t+s) = y \middle| X(s) = x\right).$$

Notice for each t that $\left(p_{(x,y)}(t)\right)$ is a – potentially infinite – matrix

$$\mathbf{P}(t) = \left(p_{(x,y)}(t)\right)_{(x,y)\in\mathcal{S}\times\mathcal{S}}$$

called the transition matrix.

For the Poisson Process, we showed in the last section that

$$p_{0k}(t-s) = \mathfrak{Pr}\left(X(t) = k \middle| X(s) = 0\right) = \frac{(\lambda(t-s)^k e^{-\lambda(t-s)}}{k!}$$

Since a Poisson Process has stationary increments and can only transition from x to $y > x$, we are able to deduce from this a closed form for the transition probabilities $p_{xy}(t)$–see the exercises.

However, it is generally quite difficult to write down the transition probabilities $p_{x,y}(t)$ for any particular process. It is possible to deduce a differential equation which can sometimes be solved, which is the approach that we used in the special case of the Poisson Process. One of the primary results of this section will be to write down a differential equation for general jump processes that the transition probabilities must always solve. Preliminary to that, we will show that the the waiting times between jumps are exponetially distributed.

First observe that the following analogue of (26.4) is true for jump processes.

35.6. Theorem.

For any n let $0 \leq t_0 < t_1 < \cdots < t_n$ and let $\{x_0, x_1, \cdots, x_n\}$ be states. Then

$$\begin{aligned}\mathfrak{Pr}\,(X(t_1) = x_1, \ldots, X(t_n) = x_n \big| X(t_0) = x_0) \\ = P_{x_0,x_1}(t_1 - t_0) \cdots P_{x_{n-1},x_n}(t_n - t_{n-1}).\end{aligned}$$

Proof. For example in the case $n = 2$,

$$\begin{aligned}&\mathfrak{Pr}\,(X(t_1) = x_1, X(t_2) = x_2 \big| X(0) = x_0) \\ &\qquad = \mathfrak{Pr}\,(X(t_1 = x_1 \big| X(0) = x_0)\,\mathfrak{Pr}\,(X(t_2) = x_2 \big| X(t_1) = x_1, X(0) = x_0) \\ &\qquad = P_{x_0,x_1}(t_1)\,\mathfrak{Pr}\,(X(t_1 = x_1 \big| X(0) = x_0)\,\mathfrak{Pr}\,(X(t_2) = x_2 \big| X(0) = x_0) \\ &\qquad = P_{x_0,x_1}(t_1 - t_0) P_{x_1,x_2}(t_2 - t_1)\end{aligned}$$

The general case follws by an easy induction.

∎

The transition matrix has the following important property.

35.7. Lemma.

The matrix $\mathbf{P}(t)$ is a semigroup, i.e., for $s, t \geq 0$

$$\mathbf{P}(t + s) = \mathbf{P}(t)\mathbf{P}(s) \tag{35.1}$$

where the product is understood to be matrix multiplication.

Proof. This follows immediately from the following calculation

Chapman-Kolmogorov equation

$$\begin{aligned}
p_{(x,y)}(t+s) &= \mathfrak{Pr}\left(X(t+s)=y \middle| X(0)=x\right)\\
&= \sum_{u\in\mathcal{S}} \mathfrak{Pr}\left(X(t+s)=y,\ \ X(t)=u \middle| X(0)=x\right)\\
&= \sum_{u\in\mathcal{S}} \mathfrak{Pr}\left(X(t+s)=y \middle| X(t)=u\right)\mathfrak{Pr}\left(X(t)=u \middle| X(0)=x\right)\\
&= \sum_{u\in\mathcal{S}} \mathfrak{Pr}\left(X(s)=y \middle| X(0)=u\right)\mathfrak{Pr}\left(X(t)=u \middle| X(0)=x\right)\\
&= \sum_{u\in\mathcal{S}} p_{(x,u)}(t)p_{(u,y)}(s)
\end{aligned}$$

∎

The semigroup identity (35.1) in the above lemma is one form of the *Chapman-Kolmogorov equation*.

The transition function $\mathbf{P}(t)$ can be quite irregular, as can the process $\{X(t)\}$. A modest continuity assumption on $\mathbf{P}(t)$ can help somewhat. We will assume in the sequel that

35.8. Definition. Continuity Assumption of $\mathbf{P}(t)$.

We will assume that

$$\lim_{h\downarrow 0}\mathbf{P}(h)=\mathbf{I}$$

where I is the identity matrix.

Since $\mathbf{P}(0)=\mathbf{I}$ the above assumption implies that the coordinate entries $p_{(x,y)}(t)$ are continuous at $t=0$. The Chapman-Kolmogorov equation implies continuity everywhere of the coordinates of $\mathbf{P}$ if P is continuous at $t=0$. Recall

35.9. Definition. Delta Function.

For any numbers x and y define

$$\delta_{xy} = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

35.10. Proposition.

If

$$\lim_{h \downarrow 0} p_{(x,y)}(h) = \delta(x, y)$$

Then $p_{(x,y)}(t)$ is continuous for each fixed x and y and each $t \geq 0$.

Proof. For $h \downarrow 0$ it follows readily from the Chapman-Kolmogorov equation that

$$\lim_{h \downarrow 0} \mathbf{P}(t + h) = \mathbf{P}(t)$$

where the limit is understood to mean coordinate-by-coordinate. Similarly, for $h \uparrow 0$ the conclusion follows from the above and

$$\begin{aligned} \mathbf{P}(h)(\mathbf{P}(t) - \mathbf{P}(t - h)) &= \mathbf{P}(h)\mathbf{P}(t) - \mathbf{P}(h)\mathbf{P}(t - h) \\ &= \mathbf{P}(h + t) - \mathbf{P}(t) \end{aligned}$$

applying the Chapman-Kolmogorov equation.

■

While the paths $\{X(t)\}$ can be quite irregular, the above assumptions imply that they have a property called *stochastic continuity*.

35.11. Theorem.

A Markov jump process satisfying the above assumptions is stochastically continuous, i.e.,

$$\lim_{h \downarrow 0} \mathfrak{Pr}\left(X(t + h) \neq X(t)\right) = \lim_{h \uparrow \ 0} \mathfrak{Pr}\left(X(t + h) \neq X(t)\right) = 0.$$

Proof. First observe that for fixed j

$$\begin{aligned}\mathfrak{Pr}\left(X(t+h) \neq X(t) \middle| X(t)=j\right) &= 1-\mathfrak{Pr}\left(X(t+h)=j \middle| X(t)=j\right) \\ &= 1-\mathfrak{Pr}\left(X(h)=j \middle| X(0)=j\right) \\ &= 1-p_{(j,j)}(h)\end{aligned}$$

Thus applying the Markov Property

$$\begin{aligned}&\mathfrak{Pr}\left(X(t+h) \neq X(t) \middle| X(0)=i\right) = \\ &\qquad = \sum_j \mathfrak{Pr}\left(X(t)=j \middle| X(0)=i\right) \mathfrak{Pr}\left(X(t+h) \neq X(t) \middle| X(t)=j\right) \\ &\qquad = \sum_j p_{(i,j)}(t)(1-p_{(j,j)}(h))\end{aligned}$$

Now we can apply the Dominated Convergence Theorem to let $h \downarrow 0$ and obtain

$$\lim_{h\downarrow 0} \mathfrak{Pr}\left(X(t+h) \neq X(t) | X(0)=i\right) = 0.$$

A second application of the same theorem yields

$$\lim_{h\downarrow 0} \sum_i \mathfrak{Pr}\left(X(t+h) \neq X(t) \middle| X(0)=i\right) \mathfrak{Pr}\left(X(0)=i\right) = 0$$

from which

$$\lim_{h\downarrow 0} \mathfrak{Pr}\left(X(t+h) \neq X(t)\right) = 0.$$

For $\lim_{h\uparrow\ 0}$, note that if $h < 0$

$$\begin{aligned}\mathfrak{Pr}\left(X(t+h) = X(t) \middle| X(0)=i\right) &= \sum_j p_{(i,j)}(t+h) p_{(j,j)}(h) \\ &\geq \sum_{j\in S} p_{(i,j)}(t+h) p_{(j,j)}(h)\end{aligned}$$

where S is any finite subset of the state space. As $h \downarrow 0$ continuity implies that

$$p_{(i,j)}(t+h) \to p_{(i,y)}(t).$$

Thus we conclude that as

$$\lim_{h\uparrow\ 0} \mathfrak{Pr}\left(X(t+h) = X(t) \middle| X(0)=i\right) \geq \sum_{j\in S} p_{(i,j)}(t) p_{(j,j)}(h).$$

The right hand side goes to one as S increases to include the entire state space, which shows that

$$\lim_{h\uparrow\ 0} \mathfrak{Pr}\left(X(t+h) = X(t)\right) = 1$$

completing the proof.

∎

The transition probabilities give us the probability of jumping from state x to state y but do not tell us when that jump might occur. It turns out that the Markov property actually tells us quite a bit about the distributions of the 'jump' times.

35.12. Definition.

For fixed $t > 0$ set

$$\tau_1(t) = +\infty \qquad \text{if } X(t+s) = X(t) \text{ for all } s \geq 0$$

and

$$\tau_1(t) = \inf\{s > 0 : X(t+s) \neq X(t)\}$$

otherwise.

Note that $\tau_1(t)$ is the waiting time from time t until the first jump to a new state, so the jump itself occurs at time $T_1 = t + \tau_1(t)$. Sometimes $\tau_1(t)$ is called the *sojourn time*, since it reflects the waiting time between jumps, not the time of the next jump.

To emphasize that $\tau_1(t)$ is a random variable defined on a probability space $(\Omega, \mathcal{E}, \mathfrak{Pr})$, we should technically write $\tau_1(t)(\omega)$ in the above definition. However, the underlying probability space $(\Omega, \mathcal{E}, \mathfrak{Pr})$ rarely plays an explicit role in our calculations and so for convenience we will most often write

$$\tau_1(t) = \inf\{s > 0 : X(t+s) \neq X(t)\}$$

where it is understood that the formulae involve random variables defined on Ω.

The following important theorem says that the waiting times between jumps are exponentially distributed.

35.13. Theorem.

For each state $x \in \mathcal{S}$ there is a $\lambda_x \in [0, \infty]$ such that for each $t \geq 0$

$$\mathfrak{Pr}\left(\tau_1(t) > u \middle| X(t) = x\right) = e^{-\lambda_x u}, \qquad u \geq 0$$

where it is understood that if $\lambda_x = \infty$ then $e^{-\lambda_x u} \equiv 0$.

Before proving this theorem, we first prove two preliminary lemmae.

35.14. Lemma.

If $\tau_1(t) > u$ then for $0 \leq s \leq u$

$$\tau_1(t+s) = \tau_1(t) - s.$$

This makes intuitive sense. If T_1 is the actual time of the next jump, then at time t the time remaining to T_1 is $\tau_1(t) = T_1 - t$. If $0 \leq s \leq u < \tau_1(t)$, then at time $t+s$, the next jump hasn't yet occured, and won't until time T_1. Thus the time remaining to T_1 from time $t+s$ is

$$\begin{aligned}\tau_1(t+s) &= T_1 - (t+s)\\ &= T_1 - t - s\\ &= \tau_1(t) - s\end{aligned}$$

More formally:

Proof. If $\tau_1(t) > u$ then for $0 \leq s \leq u$ it follows that

$$X(t+s) = X(t).$$

Now fix s_0 with $0 \leq s_0 \leq u$, so that

$$\tau_1(t+s_0) = \inf\{s > 0 : X(t+s_0+s) \neq X(t+s_0)\}.$$

For this choice of s_0,

$$X(t+s_0) = X(t)$$

and so

$$\begin{aligned}\tau_1(t+s_0) &= \inf\{s > 0 : X(t+s_0+s) \neq X(t+s_0)\}\\ &= \inf\{s > 0 : X(t+s_0+s) \neq X(t)\}\\ &= \inf\{s > s_0 : X(t+s) \neq X(t)\} - s_0\\ &= \inf\{s > 0 : X(t+s) \neq X(t)\} - s_0\\ &= \tau_1(t) - s_0\end{aligned}$$

∎

35.15. Corollary.

For $u, v > 0$ the events

$$E_1 = \{\tau_1(t) > u + v\}$$

and

$$E_2 = \{\tau_1(t) > u \quad \text{and} \quad \tau_1(t+u) > v\}$$

are the same.

Proof. We first show $E_1 \subseteq E_2$. If $\omega \in E_1$, then clearly $\tau_1(t) > u + v > u$. Further

$$\tau_1(t+u) = \tau_1(t) - u > u + v - u = v$$

so $\omega \in E_2$.

For the reverse inclusion, if $\omega \in E_2$, then $\tau_1(t) > u$ implies that

$$\tau_1(t+u) = \tau_1(t) - u.$$

From this we have

$$\begin{aligned} v &< \tau_1(t+u) \\ &= \tau_1(t) - u \end{aligned}$$

and ence $E_2 \subseteq E_1$.

$\blacksquare$

Proof. Since the process $X(t)$ is time-homogeneous the conditional probability in question is independent of t, a fact that we shall use later in the proof.

Fix $x \in \mathcal{S}$ and set

$$\phi(u) = \mathfrak{Pr}\left(\tau_1(t) > u \middle| X(t) = x\right).$$

Note that

$$1 - \phi(u) = \mathfrak{Pr}\left(\tau_1(t) \leq u \middle| X(t) = x\right)$$

and hence $1 - \phi$ is a probability distribution function. Thus $\phi(u+)$ and $\phi(u-)$ both exist and ϕ is right-continuous.

If $\phi(0) = 0$ then we may take $\lambda_x = +\infty$ and the result is immediate. Thus we assume without loss of generality that $\phi(0) > 0$.

Now for $u, v > 0$ the sets

$$E_1 = \{\omega \in \Omega : \tau_1(t) > u + v\}$$

and

$$E_2 = \{\omega \in \Omega : \tau_1(t) > u \quad \& \quad \tau_1(t+u) > v\}$$

are the same, so

$$\begin{aligned}
\phi(u+v) &= \mathfrak{Pr}\left(\tau_1(t) > u + v \middle| X(t) = x\right) \\
&= \mathfrak{Pr}\left(\tau_1(t) > u, \tau_1(t+u) > v \middle| X(t) = x\right) \\
&= \mathfrak{Pr}\left(\tau_1(t) > u \middle| X(t) = x\right) \mathfrak{Pr}\left(\tau_1(t+u) > v \middle| X(t) = x, \quad \tau_1(t) > u\right) \\
&= \phi(u) \, \mathfrak{Pr}\left(\tau_1(t+u) > v \middle| X(t) = x, \quad \tau_1(t) > u\right) & *
\end{aligned}$$

Now if $X(t) = x$ and $\tau_1(t) > u$, then $X(t+u) = x$. Then by the Markov Property

$$\begin{aligned}
\mathfrak{Pr}\left(\tau_1(t+u) > v \middle| X(t) = x, \quad \tau_1(t) > u\right) &= \mathfrak{Pr}\left(\tau_1(t+u) > v \middle| X(t+u) = x\right) \\
&= \phi(v)
\end{aligned}$$

the last equality following from the fact that $X(t)$ is time-homogeneous.

Substituting this into $(*)$ gives

$$\phi(u+v) = \phi(u)\phi(v). \qquad **$$

But now taking $u = v = 0$ in $(**)$ and using $\phi(0) > 0$ we see that $\phi(0) = 1$. Clearly

$$\lim_{u \to \infty} \phi(u) = 0.$$

Further, $(**)$ implies for any $u > 0$ and for any natural numbers n, m that

$$\phi(nu) = (\phi(u))^n \quad \text{and} \quad \phi(u) = \left(\phi\left(\frac{u}{m}\right)\right)^m.$$

We next claim that $0 < \phi(1) \leq 1$. If $\phi(1) = 0$ then

$$\phi(1/m) = (\phi(1))^{\frac{1}{m}} = 0$$

and so via right-continuity $\phi(0) = 0$, contradicting $\phi(0) > 0$.

Since $0 < \phi(1) \leq 1$ we can set

$$\lambda_x = -\ln(\phi(1))$$

so that

$$\phi(1) = e^{-\lambda_x}.$$

and $0 \leq \lambda_x < \infty$. Now for any rational number $r = p/q$ it follows that

$$\begin{aligned} \phi(r) &= \phi\left(\frac{p}{q}\right) \\ &\phi\left(\frac{1}{q}\right)^p \\ &= \phi(1)^{\frac{p}{q}} \\ &= e^{-\lambda_x \frac{p}{q}} \\ &= e^{-\lambda_x r} \end{aligned}$$

Since this is true for all rational numbers r the conclusion follows from the right-continuity of ϕ.

■

In view of the above theorem, all states fall into one of three classes.

35.16. Definition.

A state $x \in \mathcal{S}$ is
(a) absorbing if $\lambda_x = 0$;
(b) stable if $0 < \lambda_x < \infty$; and
(c) instantaneous if $\lambda_x = \infty$.

We can summarize these categories intuitively as follows. If x is an absorbing state, once $X(t) = x$ then $X(t+s) = x$ for all $s \geq 0$. If $X(t) = x$ where x is a stable state, then X remains equal to x for some period of time that is exponentially distributed with parameter λ_x. If $X(t) = x$ and x is an instantaneous state, then X instantly moves to another state. While instantaneous states are of theoretical interest, they will not arise in the applications we will consider and so we will assume that all states are either absorbing or stable. Because of this assumption, our original intuition that $\tau_i < \tau_{i+1}$ for all i is now confirmed. The following theorems summarizes our progress so far.

35.17. Theorem.

Let $X(t)$ be a time homogeneous, non-explosive Markov jump process and having no instantaneous states and suppose that $X(t) = x$. If x is not absorbing then there are a sequence of times $0 = T_0 < T_1 < T_2 \cdots < T_n \leq t$ and a sequence of states $x_0, \dots, x_n \in \mathcal{S}$ with the sojourn times

$$\tau_n = T_n - T_{n-1}$$

that are exponentially disributed with parameter λ_{x_n} and with $X(T_n) = x_n$.

35.18. Theorem.

Let $X(t)$ be a time homogeneous, non-explosive Markov jump process and having no instantaneous states. For each non-absorbing state x there are **transition probabilities** *Q_{xy} with $Q_{xx} = 0$ and $Q_{xy} = \mathfrak{Pr}\left(X(\tau_1) = y \middle| X(0) = x\right)$. Further,*

$$\sum_{y \in \mathcal{S}} Q_{xy} = 1$$

and has the property that

$$\begin{aligned}
\mathfrak{Pr}\big(X(t) = x, T_n - T_{n-1} > u \big| X(T_0) = x_0, x(T_1) = x_1, \cdots, X(T_n) = y,&\\
\text{and given times } T_0, T_1, \cdots T_n\big)&\\
= Q_{y,x} e^{-\lambda_y u}&
\end{aligned}$$

The proof of the second theorem – we omit the details which are tedious – relies on the Markov Property and on the observation that

$$\begin{aligned}
&\mathfrak{Pr}\left(X(t) = x, \tau_1 > u \middle| X(0) = y\right)\\
&\qquad = \mathfrak{Pr}\left(\tau_1 > u \middle| X(0) = y\right) \mathfrak{Pr}\left(X(t) = x \middle| \tau_1 > u, X(0) = y\right)\\
&\qquad = e^{-\lambda_y u}\, \mathfrak{Pr}\left(X(t + \tau_1) = x \middle| X(s) = y \text{ for } s \leq u\right)\\
&\qquad = e^{-\lambda_y u}\, \mathfrak{Pr}\left(X(u + \tau_1) = x \middle| X(u) = y\right)\\
&\qquad = e^{-\lambda_y u}\, \mathfrak{Pr}\left(X(\tau_1) = x \middle| X(0) = y\right)\\
&\qquad = Q_{y,x} e^{-\lambda_x u}
\end{aligned}$$

Nex we can deduce the following integeral equation, the last step before deducing the differential equation that the transition function must satisfy.

35.19. Theorem.

Let $X(t)$ be a time homogeneous, non-explosive Markov jump process and having no instantaneous states. Then for all $t \geq 0$ and for all states x

$$p_{xy}(t) = \delta_{xy} e^{-\lambda_x t} + \int_0^t \lambda_x e^{-\lambda_x s} \left(\sum_{z \neq x} Q_{xz} p_{zy}(t-s) \right) ds$$

where

$$\delta_{xy} = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

Proof. Note that if x is absorbing, $Q_{xz} = 0$ and so the above reduces to the obvious fact that

$$p_{xy}(t) = \delta_{xy}.$$

Thus we assume that x is not absorbing.

Now for x not absorbing, the event

$$\{\tau_1 \leq t, X(\tau_1) = z \text{ and } X(t) = y\}$$

occurs only if the first jump occurs at some time $s \leq t$ and takes the process to some state z and the process then goes from z to y in the remainin $(t-s)$ time period. Thus

$$P_x(\tau_1 \leq t, X(\tau_1) = z \text{ and } X(t) = y) = \int_0^t \lambda_x e^{-\lambda_x s} Q_{xz} p_{zy}(t-s)\, ds$$

This implies that

$$\begin{aligned} P_x(\tau_1 \leq t \text{ and } X(t) = y) &= \sum_{z \neq x} P_x(\tau_1 < t, X(\tau_1) = z \text{ and } X(t) = y) \\ &= \sum_{z \neq x} \int_0^t \lambda_x e^{-\lambda_x s} Q_{xz} p_{zy}(t-s)\, ds \end{aligned}$$

We have already noted that

$$\begin{aligned} P_x(\tau_1 > t \text{ and } X(t) = y) &= \delta_{xy} P_x(\tau_1 > t) \\ &= \delta_{xy} e^{-\lambda_x t} \end{aligned}$$

Thus

$$\begin{aligned} p_{xy}(t) &= P_x(X(t) = y) \\ &= P_x(\tau_1 > t \text{ and } X(t) = y) + P_x(\tau_1 \le t \text{ and } X(t) = y) \\ &= \delta_{xy} e^{-\lambda_x t} + \int_0^t \lambda_x e^{-\lambda_x s} \left(\sum_{z \neq x} Q_{xz} p_{zy}(t-s) \right) ds \end{aligned}$$

as desired.

■

A simple change of variables yeilds

35.20. Corollary.

Let $X(t)$ be a time homogeneous, non-explosive Markov jump process and having no instantaneous states. Then for all $t \ge 0$ and for all states x

$$p_{xy}(t) = \delta_{xy} e^{-\lambda_x t} + \lambda_x e^{-\lambda_x t} \int_0^t e^{-\lambda_x s} \left(\sum_{z \neq x} Q_{xz} p_{zy}(s) \right) ds \tag{35.2}$$

35.21. Corollary.

With $X(t)$ as above, $p_{xy}(t)$ is continuous.

35.22. Definition. Infinitessimal Generators

Let $X(t)$ be a time homogeneous, non-explosive Markov jump process and having no instantaneous states. Then the infinitessimal generators of $X(t)$ are the numbers

$$q_{xy} = p'_{xy}(0).$$

35.23. Theorem. The Backward Equation.

Let $X(t)$ be a time homogeneous, non-explosive Markov jump process and having no instantaneous states and infitessimal generators q_{xy}. Then

$$q_{xy} = \begin{cases} -\lambda_x & y = x \\ \lambda_x Qxy & y \neq x \end{cases}$$

Further

$$\sum_{y \neq x} q_{xy} = \lambda_x = -q_{xx}$$

and for $t \geq 0$

$$p'_{xy}(t) = -\lambda_x p_{xy}(t) + \lambda_x \sum_{z \neq x} Q_{xz} p_{zy}(t).$$

Further

$$p'_{xy}(0) = -\lambda_x \delta_{xy} + \lambda_x Q_{xy}$$

Proof. We can differentiate (35.2) to obtain

$$p'_{xy}(t) = -\lambda_x p_{xy}(t) + \lambda_x \sum_{z \neq x} Q_{xz} p_{zy}(t).$$

In particular,

$$\begin{aligned} p'_{xy}(0) &= -\lambda_x p_{xy}(0) + \lambda_x \sum_{z\neq x} Q_{xz} p_{zy}(0) \\ &= -\lambda_x \delta_{xy} + \lambda_x \sum_{z\neq x} Q_{xz} \delta_{zy} \\ &= -\lambda_x \delta_{xy} + \lambda_x Q_{xy} \end{aligned}$$

and the rest of the results follow.

▮

There is a similar forward equation; we again omit the proof as it is somewhat more complex.

35.24. Theorem. Forward Equation.

Let $X(t)$ be as above. Then

$$p'_{xy}(t) = \sum_z p_{xz}(t) p'_{zy}(0).$$

In the next section we will apply these results to the birth and death process, which in turn models various queuing processes.

35. Markov Pure Jump Processes: Problems.

1. Suppose that customers arrive at queue according to a Poisson Process with arrival rate λ. Suppose that the customers all belong to one of two types, type 0 and type 1. (For example, the customers might be in a store and one type pays with cash and the other type represents all other payment types.) If $X(t)$ describes the *type* of the last customer, i.e., if

$$X(t) = \begin{cases} 1 & \text{if the last customer is type one} \\ 0 & \text{if there is no customer or the customer is not type one} \end{cases}$$

Suppose that customer type is independent of past history and that the probability of type one is p.

(a) Show that the chain $X(t)$ described above has transition matrix (with $q = 1 - p$)

$$P(t) = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} p + qe^{-\lambda t} & q - qe^{-\lambda t} \\ p - pe^{-\lambda t} & q + pe^{-\lambda t} \end{pmatrix} \end{array}$$

Hint: Find and solve the system of differential equations represented by the backward equation.

(b) For the above chain, apply 35.6 to conclude that

$$\mathfrak{Pr}\left(\tau_1 > t \middle| X(0) = 0\right) = \lim_{n\to\infty} \left(\mathfrak{Pr}\left(X(t/n) = 0 \middle| X(0) = 0\right)\right)^n$$

(c) Use (b) to show that $\mathfrak{Pr}\left(\tau_1 > t \middle| X(0) = 0\right) = e^{-\lambda t}$.

2.

(a) Find $p_{xy}(t)$ for the Poisson Process.

(b) Find the infinitessimal generators q_{xy} of the Poisson process.

A birth and death process is one in which it is possible to transition from state x only to state $x-1$ or state $x+1$. More formally

36.1. Definition.

A jump process $X(t)$ *having infinitessimal generators* q_{xy} *is a* **birth and death process** *if*

$$q_{xy} = 0 \quad \text{whenever} \quad |x-y| > 1.$$

In this case, the birth rate is the number

$$\lambda_x = q_{x,x+1}$$

and the death rate is the number

$$\mu_x = q_{x,x-1}.$$

Recall from the prior section that for each non-absorbing state x there are two random quantities: the time of the first jump, $\tau_1(x)$; and the value the process assumes at time τ_1, y. In the last section, we showed that $\tau_1(x)$ is exponentially distributed. In this section, we will denote the parameter of that exponential distribution by q_x.

36.2. Proposition.

Let $X(t)$ be a birth and death process having infinitessimal generators q_{xy}. Set $q_{xx} = -q_x$, where q_x is the parameter of the exponential distribution for $\tau_1(x)$, the time to the first jump. Then

$$-q_{xx} = q_x = q_{x,x+1} + q_{x,x-1} = \lambda_x + \mu_x.$$

Moreover, if x is not absorbing, i.e., $\mu_x \neq 0 \neq \lambda_x$, then

$$Q_{x,x+1} = \frac{\lambda_x}{\lambda_x + \mu_x}$$

and

$$Q_{x,x-1} = \frac{\mu_x}{\lambda_x + \mu_x}$$

Proof. It follows from the definitions and $-q_{xx} = \sum_{y \neq x} q_{xy}$ that

$$-q_{xx} = q_x = q_{x,x+1} + q_{x,x-1} = \lambda_x + \mu_x.$$

Similarly

$$\begin{aligned} \lambda_x &= q_{x,x+1} \\ &= q_x Q_{x,x+1} \\ &= (\lambda_x + \mu_x) Q_{x,x+1} \end{aligned}$$

Rearranging gives

$$Q_{x,x+1} = \frac{\lambda_x}{\lambda_x + \mu_x}.$$

The second conclusion follows in a similar manner.

∎

Birth and Death processes have simplified forward and backward equations which can often be solved.

36.3. Proposition.

Let $X(t)$ be a birth and death process having infinitessimal generators q_{xy}. Then the backward equation is

$$p'_{xy}(t) = \lambda_x p_{x+1,y}(t) - (\lambda_x + \mu_x)p_{xy}(t) + \mu_x p_{x-1,y}(t)$$

and the forward equation is

$$p'_{xy}(t) = \lambda_{y-1} p_{x,y-1}(t) - (\lambda_y + \mu_y)p_{xy}(t) + \mu_{y+1} p_{x,y+1}(t)$$

Proof. For the forward equation

$$\begin{aligned}
p'_{xy}(t) &= \sum_z p_{xz}(t)p'_{zy}(0) \\
&= \sum_z p_{xz}(t)q_{zy} \quad \text{(via 35.22)} \\
&= p_{x,y-1}(t)q_{y-1,y} + p_{x,y}(t)q_{y,y} + p_{x,y+1}(t)q_{y+1,y} \\
&= \lambda_{y-1} p_{xy}(t) - (\lambda_y + \mu_y)p_{xy}(t) + \mu_{y+1} p_{x-1,y}(t)
\end{aligned}$$

The proof for the forward equation is similar.

∎

In order to solve the forward equation in various examples, we will use the variation of constants formula below.

36.4. Proposition.

Suppose for some functions f and g and some constant α that

$$f'(t) = -\alpha f(t) + g(t).$$

Then

$$f(t) = f(0)e^{-\alpha t} + \int_0^t e^{-\alpha(t-s)} g(s)\, ds.$$

Proof. We can write

$$f'(s) + \alpha f(s) = g(s)$$

so, multiplying by $e^{\alpha s}$,

$$f'(s)e^{\alpha s} + \alpha e^{\alpha s} f(s) = e^{\alpha s} g(s)$$

which implies

$$\frac{d}{ds}\left(f(s)e^{\alpha s}\right) = e^{\alpha s} g(s).$$

Then integrating from $s = 0$ to $s = t$ and applying the fundamental theorem of calculus gives

$$f(t)e^{\alpha t} - f(0) = \int_0^t e^{\alpha s} g(s)\, ds.$$

Rearranging gives the result. ∎

36.5. Example.

Let $X(t)$ be a birth and death process and suppose that the state space is $\mathcal{S} = \{0, 1\}$. Suppose that both 0 and 1 are not absorbing. We will find $p_{xy}(t)$ and $\mathfrak{Pr}\left(X(t) = 0\right)$ and $\mathfrak{Pr}\left(X(t) = 1\right)$.

Solution. Note that $\mu_0 = 0 = \lambda_1$ so that the λ_0 and μ_1 determine the evolution of the process. For convenience we will write $\lambda \equiv \lambda_0$ and $\mu \equiv \mu_1$. The backward equation for $(x, y) = (0, 0)$ becomes

$$\begin{aligned} p'_{xy}(t) = p'_{00}(t) &= \lambda_x p_{1,0}(t) - (\lambda_0 + \mu_0)p_{00}(t) + \mu_0 p_{-1,0}(t) \\ &= \lambda p_{10}(t) - \lambda p_{00}(t) + 0 \\ &= -\lambda(p_{00}(t) - p_{10}(t)) \end{aligned}$$

and at $(x, y) = (1, 0)$

$$\begin{aligned} p'_{xy}(t) = p'_{10}(t) &= \lambda_1 p_{1,0}(t) - (\lambda_1 + \mu_1)p_{10}(t) + \mu_1 p_{0,0}(t) \\ &= 0 - \mu p_{10}(t) + \mu p_{00}(t) \\ &= \mu(p_{00}(t) - p_{10}(t)). \end{aligned}$$

Thus upon substracting the two equations

$$\frac{d}{dt}\left(p_{00}(t)-p_{10}(t)\right) = -(\lambda+\mu)\left(p_{00}(t)-p_{10}(t)\right).$$

This implies that

$$p_{00}(t)-p_{10}(t) = e^{-(\lambda+\mu)t}.$$

Sustituting back into the forumula for $p'_{00}(t)$ gives

$$\begin{aligned} p'_{00}(t) &= -\lambda\left(p_{00}(t)-p_{10}(t)\right) \\ &= -\lambda e^{-(\lambda+\mu)t}. \end{aligned}$$

Thus

$$\begin{aligned} p_{00}(t) &= p_{00}(0) + \int_0^t p'_{00}(s)\,ds \\ &= 1 - \frac{\lambda}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)t}\right) \end{aligned}$$

and so collecting terms

$$p_{00}(t) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t}$$

In exactly the same manner

$$p_{10}(t) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t}$$

Since $p_{01}(t) = 1 - p_{00}(t)$ and $p_{11}(t) = 1 - p_{10}(t)$ we can conclude that the transition matrix is given by

$$P = \begin{array}{c c} & \begin{array}{cc} \quad 0 & \quad\quad 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left(\begin{array}{cc} \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} & \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} \\ \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t} & \frac{\mu}{\lambda+\mu} - \frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t} \end{array}\right) \end{array}$$

■

In general a birth and death process need not be non-explosive. While we will not pursue this topic here, a simple condition that is sufficient to guarantee that birth and death process be nonexplosive is

$$\lambda_x \le Ax + B$$

for some constants $A, B > 0$. This condition is fulfilled in the examples considered in this section. The following theorem collects some facts that will be useful in understanding more complex birth and death processes.

36.6. Theorem.

Let $\{\xi_1, \cdots, \xi_n\}$ be a collection of independent exponentially distributed random variables having parameters $\{\alpha_1, \cdots, \alpha_n\}$. Then

(a) the random variable $\min\{\xi_1, \cdots, \xi_n\}$ is an exponentially distributed random variable having parameter $\alpha_1 + \cdots + \alpha_n$;

(b) For each $k = 1, \ldots, n$

$$\mathfrak{Pr}\left(\xi_k = \min\{\xi 1, \cdots, \xi_n\}\right) = \frac{\alpha_k}{\alpha_1 + \cdots + \alpha_n};$$

(c) with probability one the random variables $\{\xi_1, \cdots, \xi_n\}$ take on n distinct values.

Proof. For (a) observe that

$$\begin{aligned}
\mathfrak{Pr}\left(\min\{\xi_1, \cdots, \xi_n\} > t\right) &= \mathfrak{Pr}\left(\xi_n > t, \xi_2 > t, \cdots, \xi_n > t\right) \\
&= \mathfrak{Pr}\left(\xi_n > t\right) \cdots \mathfrak{Pr}\left(\xi_n > t\right) \\
&= e^{-\alpha_1 t} \cdots e^{-\alpha_n t} \\
&= e^{-(\alpha_k + \cdots + \alpha_n)t}
\end{aligned}$$

from which (a) follows.

For (b) fix k with $1 \le k \le n$ and set

$$\eta_k = \min_{j \ne k}\{\xi_j\}$$

and

$$\beta_k = \sum_{j \ne k} \alpha_j.$$

Then by (a) η_k is exponentially distributed with parameter β_k. Further since ξ_k and η_k are independent

$$\begin{aligned}
\mathfrak{Pr}\left(\xi_k = \eta_k\right) &= \mathfrak{Pr}\left(\xi_k \leq \eta_k\right) \\
&= \int_0^\infty \int_z^\infty \alpha_k e^{-\alpha_k u} \beta_k e^{-\beta_k v} \, dv \, du \\
&= \int_0^\infty \alpha_k e^{-\alpha_k u} e^{-\beta_k u} \, du \\
&= \frac{\alpha_k}{\alpha_k + \beta_k} \\
&= \frac{\alpha_1}{\alpha_1 + \dots + \alpha_n}
\end{aligned}$$

which establishes (b).

For (c) it is sufficient to show that $\mathfrak{Pr}\left(\xi_i \neq \xi_j\right) = 1$ if $i \neq j$. However since ξ_i and ξ_j are continuous random variables having a continuous joint density,

$$\mathfrak{Pr}\left(\xi_i \neq \xi_j\right) = \int \int_{\left\{(x,y) \middle| x \neq y\right\}} f(x,y) \, dx \, dy = 1.$$

∎

36.7. Example. Branching Process.

The branching process is similar to the branching chain. We consider a population of particles where each particle survives for a period of time that is exponentially distributed with parameter q. At the end of the survival time, the particle either splits into two particles with probability p or vanishes completely with probability $(1-p)$. We assume that the particles behave independently of each other and of elapsed time. We let $X(t)$ count the number of particles at time t. We will find the infinitessimal parameters for this process.

Solution. Suppose that $X(0) = x$ so that there are initially x particles. Then each particle has a lifespan ξ_i:

$$\begin{aligned}\xi_1 &= \text{time that particle one splits or vanishes}\\ &\vdots\\ \xi_k &= \text{time that particle } k \text{ splits or vanishes}\\ &\vdots\\ \xi_x &= \text{time that particle } x \text{ splits or vanishes}\end{aligned}$$

and

$$\tau_1 = \min\{\xi_1, \cdots, \xi_x\}.$$

Then τ_1 is exponentially distributed with parameter xq. Further

$$Q_{x,x+1} = p \text{ and } Q_{x,x-1} = (1-p)$$

so that

$$\lambda_x = q_x Q_{x,x+1} = xqp$$

and

$$\mu_x = q_x Q_{x,x-1} = xq(1-p).$$

■

We remark in passing that a Poisson Process is a branching process with $p = 1$.

36.8. Example. Branching Process with Immigration

In this case we suppose that new particles are added to the system according to a Poisson Process with parameter λ and that all particles behave independently. We will find the infinitessimal generators.

Solution. As before, we suppose that $X(0) = x$ and let

$$\xi_k = \text{time when particle } k \text{ divides or disappears.}$$

Let η be the time that the first new particle enters the system, so that the time of the first jump is

$$\tau_1 = \min\{\xi_1, \cdots, \xi_x, \eta\}.$$

Then τ_1 is exponentially distributed with parameter $xq + \lambda$ and

$$\mathfrak{Pr}\left(\tau_1 = \eta\right) = \frac{\lambda}{xq + \lambda}.$$

Thus

$$Q_{x,x+1} = \frac{\lambda}{xq + \lambda} + \frac{xq}{xq + \lambda} p$$

while

$$Q_{x,x-1} = \frac{xq}{xq + \lambda}(1 - p).$$

This in turn implies

$$\lambda_x = q_x Q_{x,x+1} = xqp + \lambda$$

and

$$\mu_x = q_x Q_{x,x-1} = xq(1 - p).$$

$\blacksquare$

Before turning to the next example, we deduce a useful property of the Poisson Process.

36.9. Theorem.

Let $X(t)$ be a Poisson Process with parameter λ. Suppose that $X(t) = k$ and, for $i = 1, \cdots, k$, set

$$\tau_i = \textit{time that jump } i \textit{ occurs.}$$

Then the random variables

$$\tau_i\big|_{X(t)=k}$$

are uniformly distributed on $[0, t]$.

Proof. Suppose for example that $X(t) = k$ and fix a partition

$$0 = t_0 < t_1 < \cdots < t_n = t$$

If we then fix i and let

$$X_i(t) = \text{number of jumps between } t_{i-1} \text{ and } t_i$$

then each X_i is a Poisson random variable with parameter $\lambda(t_i - t_{i-1})$, the random variables X_i are independent and

$$X_1 + X_2 + \cdots + X_n = X(t).$$

In turn, $X(t)$ has a Poisson distribution with parameter λt. Thus if we select any integers $x_1, \dots, x_m$ with

$$0 < x_1 < x_2 < \cdots < x_m$$

and so that $x_1 + \cdots x_m = k$ then

$$\begin{aligned}
\mathfrak{Pr}\left(X_1 = x_1, \dots, X_m = x_m \middle| X(t) = k\right)& \\
&= \mathfrak{Pr}\left(X_1 = x_1, \dots, X_m = x_m \middle| X_1 + \cdots + X_m = k\right) \\
&= \frac{\mathfrak{Pr}\left(X_1 = x_1, \dots, X_m = x_m, X_1 + \cdots + X_m = k\right)}{\mathfrak{Pr}\left(X_1 + \cdots + X_m = k\right)} \\
&= \frac{\mathfrak{Pr}\left(X_1 = x_1, \dots, X_m = x_m\right)}{\mathfrak{Pr}\left(X_1 + \cdots + X_m = k\right)} \\
&= \frac{\prod_{i=1}^{m} \frac{\lambda(t_i - t_{i-1})^{x_i} e^{-\lambda(t_i - t_{i-1})}}{x_i!}}{\frac{(\lambda t)^k e^{-\lambda t}}{k!}} \\
&= \frac{k!}{\prod_{i=1}^{m} x_k!} \prod_{i=1}^{m} \left(\frac{t_i - t_{i-1}}{t}\right)^{x_i}
\end{aligned}$$

These multinomial probabilities are exactly those that result form choosing the k arrival times indepndently and uniformly from $[0, t]$.

■

36.10. Example. Infinite Server Queue.

Suppose that customers arrive in a queue according to the Poisson Process with parameter λ and are immediately served (so this is an M/M/∞ queue) where the service times are independent of each other and the arrivals and are exponentially distributed with parameter μ. We will consider the total number of customers being served, $X(t)$. This is a special case of the Branching Chain with immigration where $q = \mu$ and $p = 0$. Thus $\lambda_x = \lambda$ and $\mu_x = x\mu$. We will find the transition function $p_{xy}(t)$ and deduce a formula for $\lim_{t\to\infty} p_{xy}(t)$.

Solution. If a customer arrives at some time $s \in (0, t]$, the probability that the customer is still being served at time t is $e^{-\mu(t-s)}$. Thus if the arrival time is chosen uniformly from $(0, t]$, then the probability that the customer is still being served at time t is

$$p_t = \frac{1}{t}\int_0^t e^{-\mu(t-s)}\,ds = \frac{1-e^{-\mu t}}{\mu t}.$$

Now if $X_1(t)$ is the number of customers who arrived in the interval $(0, t]$ who are still being served at time t, and if $Y(t)$ is the total number of customers who arrive in the interval $(0, t]$ then if follows that

$$X_1(t)\big|_{Y(t)=k}$$

is binomially distributed with paramters k and p_t, i.e.,

$$\mathfrak{Pr}\left(X_1(t) = n \middle| Y(t) = k\right) = \binom{k}{n} p_t^n (1-p_t)^{k-n}.$$

Since $Y(t)$ has a Poisson distribution with parameter λt it follows that

$$\begin{aligned}
\mathfrak{Pr}\left(X_1(t)=n\right) &= \sum_{k=n}^{\infty} \mathfrak{Pr}\left(Y(t)=k,\, X_1(t)=n\right) \\
&= \sum_{k=n}^{\infty} \mathfrak{Pr}\left(Y(t)=k\right) \mathfrak{Pr}\left(X_1(t)=n \middle| Y(t)=k\right) \\
&= \sum_{k=n}^{\infty} \frac{(\lambda t)^k e^{-\lambda t}}{k!} \frac{k!}{n!(k-n)!} p_t^n (1-p_t)^{k-n} \\
&= \frac{(\lambda t p_t)^n e^{-\lambda t}}{n!} \sum_{k=n}^{\infty} \frac{(\lambda t(1-p_t))^{k-n}}{(k-n)!} \\
&= \frac{(\lambda t p_t)^n e^{-\lambda t}}{n!} \sum_{m=0}^{\infty} \frac{(\lambda t(1-p_t))^{m}}{m!} \\
&= \frac{(\lambda t p_t)^n e^{-\lambda t}}{n!} e^{\lambda t(1-p_t)} \\
&= \frac{(\lambda t p_t)^n e^{-\lambda t p_t}}{n!}
\end{aligned}$$

In particular, $X_1(t)$ has a Poisson distribution with parameter

$$\lambda t p_t = \frac{\lambda}{\mu}(1-e^{-\mu t}).$$

Next suppose that $X(0) = x$ be the number of customers being initially served and let $X_2(t)$ be the number of those initial customers still being served at time t. Then

$$\begin{aligned}
X(t) &= (\text{number of customers present at time } 0 \text{ still being served at time } t) + \cdots \\
&\qquad \cdots + (\text{number of customers who arrive in } (0,t] \text{ still being served at time } t) \\
&= X_2(t) + X_1(t)
\end{aligned}$$

Further, $X_1(t)$ and $X_2(t)$ are independent and $X_2(t)$ has a binomial distribution with parameters x and $e^{-\mu t}$. Since $X(t) = X_1(t) + X_2(t)$ it follows that

$$p_{xy}(t) = P_x(X(t)=y) = \sum_{k=0}^{\min(x,y)} P_x(X_2(t)=k)P(X_1(t)=y-k)$$

and so

$$p_{xy}(t) = \sum_{k=0}^{\min\{x,y\}} \Bigg[\binom{x}{k} e^{-k\mu t}(1-e^{-\mu t})^{x-k} \times$$
$$\times \frac{\left(\frac{\lambda}{\mu}(1-e^{-\mu t}\right)^{(y-k)}}{(y-k)!} \exp\left(-\frac{\lambda}{\mu}(1-e^{-\mu t}\right) \Bigg].$$

Notice that if $k \geq 1$ then all of the terms in the above sum tend to zero as $t \to \infty$. Thus

$$\lim_{t\to\infty} p_{xy}(t) = \frac{(\lambda/\mu)^y e^{-\lambda/\mu}}{y!}.$$

■

36. Birth and Death Processes: Problems.

1. Let $X(t)$ be a two-state birth and death chain and set

$$\pi(0) = \frac{\mu}{\lambda + \mu} \quad \text{and} \quad \pi(1) = \frac{\lambda}{\lambda + \mu}.$$

(a) Show that for $x = 0, 1$

$$\lim_{t \to \infty} \mathfrak{Pr}\,(X(t) = x) = \pi(x).$$

(b) Show that the distribution of $X(t)$ is independent of t if and only if the inititial distribution for $X(0)$ is π.

(c) Find $\mu(t) = E(X(t))$.

2. Consider a birth and death chain having three states $\mathcal{S} = \{0, 1, 2\}$ and birth rates such that $\lambda_0 = \mu_2$. Find $\pi_{0y}(t)$ for $y = 0, 1, 2$

3. In the infinite server queue, suppoes that there are $X(0) = x$ customers initially present. Find $E(X(t)\big|X(0) = x)$.

4. Consider a birth and death process $X(t)$ with

$$\lambda_x = x\lambda \quad \text{and} \quad \mu_x = x\mu$$

for constants $\lambda \geq 0$ and $\mu \geq 0$. Set

$$\mu_x(t) = E_x(X(t)) = \sum_{y=0}^{\infty} y p_{xy}(t).$$

(a) Write the forward equation for the process.

(b) Use the forward equation to show that

$$\mu_x'(t) = (\lambda - \mu)\mu_x(t)$$

(c) for all x Conclude that $\mu_x(t) = xe^{(\lambda - \mu)t}$.

5. Let $X(t)$ be a birth and death process for which $\lambda_x = 0$ for all$x$ (such a process is called a pure death process).

`pure death process`

(a) Write the forward equation
(b) Find $p_{xy}(t)$
(c) Solve for $p_{xy}(t)$ in terms of $p_{x,y\lnot 1}(t)$.
(d) Find $p_{x,x-1}(t)$
(e) Show that if $\mu_x = x\mu$ for some constant μ then

$$p_{xy}(t) = \binom{x}{y}(e^{-\mu t})^y(1 - e^{-\mu t})^{x-y}$$

for $0 \le y \le x$.

37. Queueing Processes

Roughly speaking, the study of queues involves the study of waiting in line. In order to describe a queue, it is generally necessary to specify several components:

(i) a description of how new customers arrive in the queue;
(ii) a description of how customers are served;
(iii) how many servers (channels) are available;
(iv) the capacity of the system (how many customers are permitted);
(v) the size of the population of customers; and
(vi) the service priority.

There are several standard possibilities for each of the above components. This classification scheme is referred to as Kendall's Notation.

37.1. Definition. Kendall's Notation

(i) *a description of how new customers arrive in the queue;*
- *(a)* **M**: *stands for* $\mathfrak{Markovian}$ *which is understood to also specify inter-arrival times that have an exponential distribution;*
- *(b)* $\mathbf{M}^{[X]}$: *Markovian where customers arrive in groups described by the random variable* $[X]$*;*
- *(c)* **D**: *for deterministic (or degenerate) arrivals;*
- *(d)* E_k: *Erlang arrivals with parameter* k*;*
- *(e)* **G**: *for general independent arrivals, sometimes written as* **GI**.

(ii) *a description of how customers are served;*
- *(a)* **M**: *stands for* $\mathfrak{Markovian}$ *which is understood to also specify service times that have an exponential distribution;*
- *(b)* $\mathbf{M}^{[X]}$: *Markovian where customers are served in groups described by the random variable* $[X]$*;*
- *(c)* **D**: *for deterministic (or degenerate) service times;*
- *(d)* E_k: *Erlang service times nwith parameter* k*;*
- *(e)* **G**: *for general independent servers, sometimes written as* **GI**.

(iii) *how many servers (channels) are available;*

(iv) *the capacity of the system (how many customers are permitted); if the queue is at capacity, then additional customers are turned away or lost;*

(v) *the size of the population of customers; and*

(vi) *the service priority.*
- *(a)* **FIFO**: *first in first out;*
- *(b)* **LIFO**: *last in first out;*
- *(c)* **SIRO**: *service in random order;*
- *(d)* *other service protocols;*

If omitted, (iv) and (v) are assumed to be infinite. However, if either (iv) or (v) are included in the specification, then they must both be included to avoid confusion. Thus a queue specified by M/M/k/n/∞ would be one with exponential arrival and service times, with k servers, with a maximum of n customers in the queue at once and with an infinite number of potential customers.

37.2. Example.

Let $X(t)$ be a queue in which the new customers arrive according a Poisson process with parameter λ, there are n servers, each having service times that are exponentially distributed with parameter μ. If one supposes that the arrivals are independent of one another and the servers and that the servers are independent of one another, then the result is an M/M/n queue.

Solution. Note that this is a birth and death chain in which

$$q_{i,i-1} = \begin{cases} i\mu & 1 \le i \le n \\ n\mu & j > n \end{cases}$$

and

$$q_{i,i+1} = \lambda.$$

∎

38. Second Order Processes

covariance
cross-covariance
autocovariance

38.1. Definition.

Let $\{X(t)\}$ be a stochastic process and suppose that $E(X(t)) < \infty$ for all t. Then the mean function for $\{X(t)\}$ is the function

$$\mu_X(t) = E(X(t)).$$

Similarly, if $E(X(s)X(t)) < \infty$ for all s, t then the **covariance function** *is*

$$r_X(s,t) = \operatorname{cov}(X(s), X(t)) = E\Big(X(s)X(t)\Big) - E(X(t))E(X(s)).$$

Recall that if U and V are random variables then the *covariance* of U and V is

$$\operatorname{cov}(U,V) = E(UV) - E(U)E(V).$$

This is sometimes referred to as the *cross-covariance*, while the covariance function for $X(t)$ is sometimes called the *autocovariance* function.

Note that $\operatorname{var}(X(t)) = \operatorname{cov}(X(t), X(t)) = r_X(t,t)$. Also the covariance function is symmetric in s and t, i.e.,

$$r_X(s,t) = r_X(t,s).$$

Also note that if we are given

times	$t_1, t_2, \ldots, t_n$	and	times	$s_1, s_2, \ldots, s_m$
and scalars	$\alpha_1, \alpha_2, \ldots \alpha_n$		and scalars	$\beta_1, \beta_2, \ldots \alpha_m$

and define the random variables U and V by

$$U = \sum_{i=1}^{n} \alpha_i X(t_i) \quad \text{and} \quad V = \sum_{j=1}^{m} \beta_j X(s_j)$$

then

$$\begin{aligned}\operatorname{cov}(U,V) &= \operatorname{cov}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i X(t_i)\beta_j X(s_j)\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j r_X(s_j,t_j)\end{aligned}$$

In particular

$$\operatorname{var}(V) = \sum_{i=1}^{m}\sum_{j=1}^{m}\beta_i\beta_j r_X(s_j,s_j) \geq 0$$

and so r_X is non-negative definite.

38.2. Definition.

A process $X(t)$ is **second order stationary** *or just* **second order** *if for every $\tau \in \mathbb{R}$ the process Y defined by*

$$Y(t) = X(t+\tau)$$

has the same mean and covariance functions as $X(t)$.

38.3. Proposition.

Suppose that $X(t)$ is a second order stationary process. Then $\mu_X(t)$ does not depend on t and $r_X(s,t)$ depends only on the difference $(t-s)$.

Proof. If $Y(t) = X(t+\tau)$ then

$$\mu_X(t) = \mu_Y(t) = \mu_X(t+\tau)$$

from which

$$\mu_X(0) = \mu_X(\tau)$$

for all $\tau \in \mathbb{R}$. Thus $\mu_X(t)$ is constant and equal to $\mu_X(0)$ for all t.

Similarly,

$$\begin{aligned}r_X(s,t) &= r_Y(s,t)\\ &= \operatorname{cov}(Y(s),Y(t))\\ &= \operatorname{cov}(X(s+\tau),X(t+\tau))\\ &= r_X(s+\tau,t+\tau)\end{aligned}$$

Since this identity holds for all $\tau \in \mathbb{R}$ it holds for $\tau = -s$ so

$$r_X(s,t) = r_X(0, t-s)$$

as desired.

∎

38.4. Definition.

If $X(t)$ is a second order stationary process, then the covariance function of $X(t)$ is usually written as

$$r_X(t) \equiv r_X(0, t-0).$$

Note that, for a second order stationary process, $\operatorname{var}(X(t)) = r_X(t,t) = r_X(0)$.

38.5. Proposition.

Let $X(t)$ be a second order stationary process. Then $|r_X(t)| \leq r_X(0)$ for all $t \in \mathbb{R}$.

Proof. Note that $\operatorname{var}(X(t)) = r_X(t,t) = r_X(0)$ for all $t \in \mathbb{R}$. Thus

$$\begin{aligned} r_X(t) &= \operatorname{cov}(X(0), X(t)) \\ &= E\Big(X(0) - E(X(0)), X(t) - E(X(t)\Big) \\ &\quad \text{(applying Cauchy-Shwarz)} \\ &\leq \sqrt{\operatorname{var}(X(0)) \operatorname{var}(X(t))} \\ &= r_X(0) \end{aligned}$$

giving the desired conclusion.

∎

38.6. Example.

Let Z_1 and Z_2 be independent, normally distributed random variables having mean 0 and standard deviation σ. For fixed $\lambda \in \mathbb{R})$ set

$$X(t) = \cos(\lambda t) Z_1 + \sin(\lambda t) Z_2.$$

We will find $\mu_X(t)$ and $r_X(s,t)$.

Solution. Note that

$$\begin{aligned} \mu_X(t) &= E\left(\cos(\lambda t) Z_1 + \sin(\lambda t) Z_2\right) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} r_X(s,t) &= \operatorname{cov}(X(s), X(t)) \\ &= E\Big(X(s)X(t)\Big) - E\Big(X(s)\Big)E\Big(X(t)\Big) \\ &= E\Big(X(s)X(t)\Big) - 0 \\ &= E\left((Z_1 \cos(\lambda t) + Z_2 \sin(\lambda t))\left(Z_1 \cos(\lambda s) + Z_2 \sin(\lambda s)\right)\right) \\ &= \sigma^2 \cos(\lambda t)\cos(\lambda s) + \sigma^2 \sin(\lambda s)\sin(\lambda t) \\ &= \sigma^2 \cos(\lambda(t-s)) \end{aligned}$$

Thus $X(t)$ is second order stationary.

■

38.7. Example.

Let $X(t)$ be the Poisson process with parameter λ. Find the mean and covariance functions for $X(t)$.

Solution. First note that $\mu_X(t) = \lambda t$ and hence the Poisson Process is not a second order stationary process.

Suppose now that $0 \le s \le t$. Then

$$X(s) - X(0) \quad \text{and} \quad X(t) - X(s)$$

are independent, so

$$\operatorname{cov}(X(s), X(t) - X(s)) = \operatorname{cov}(X(s) - X(0), X(t) - X(s)) = 0$$

since $X(0) = 0$ for the Poisson Process.

Thus

$$\begin{aligned}\operatorname{cov}(X(s), X(t)) &= \operatorname{cov}(X(s), X(s) + X(t) - X(s)) \\ &= \operatorname{cov}(X(s), X(s)) + \operatorname{cov}(X(s), X(t) - X(s))) \\ &= \lambda s\end{aligned}$$

Further, if $s < 0 < t$ then $\operatorname{cov}(X(s), X(t)) = \operatorname{cov}(X(s) - X(0), X(t) - X(0)) = 0$, again via independence. The other cases are similar, and thus

$$r_X(s,t) = \begin{cases} \lambda \min\{|s|, |t|\} & \text{if } st \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

∎

38.8. Example.

Let $X(t)$ be the Poisson process with parameter λ, and set $Y(t) = X(t+1) - X(t)$. Find the mean and covariance functions for $X(t)$.

Solution. Since $E(X(t)) = \lambda t$, it follows that

$$\begin{aligned}E(Y(t)) &= E(X(t+1) - X(t)) \\ &= \lambda(t+1) - \lambda t \\ &= \lambda,\end{aligned}$$

and hence each $Y(t)$ has the same mean, λ.

To compute the covariance function of Y, first observe that if $t - s \geq 1$, then the random variables

$$X(s+1) - X(s) \quad \text{and} \quad X(t+1) - X(t)$$

are independent since

$$s \leq s+1 \leq t \leq t+1.$$

By symmetry, if $|t-s| \geq 1$, then

$$X(s+1)-X(s) \quad \text{and} \quad X(t+1)-X(t)$$

are independent. Thus,

$$r_Y(s,t)=0 \quad \text{for} \quad |t-s| \geq 1.$$

Suppose next that $s \leq t \leq s+1$. Then

$$\begin{aligned} \operatorname{cov}(Y(s),Y(t)) &= \operatorname{cov}(X(s+1)-X(s),X(t+1)-X(t)) \\ &= \operatorname{cov}(X(t)-X(s)+X(s+1)-X(t),X(s+1)-X(t)+\cdots \\ &\quad \cdots+X(t+1)-X(s+1)). \end{aligned}$$

Now, under the assumptions on s and t,

$$\operatorname{cov}(X(t)-X(s),X(s+1)-X(t))=0$$

$$\operatorname{cov}(X(t)-X(s),X(t+1)-X(s+1))=0$$

and

$$\operatorname{cov}(X(s+1)-X(t),X(t+1)-X(s+1))=0.$$

At the same time,

$$\operatorname{cov}(X(s+1)-X(t),X(s+1)-X(t))=\operatorname{var}(X(s+1)-X(t))=\lambda(s+1-t).$$

Thus

$$\operatorname{cov}(Y(s),Y(t))=\lambda(s+1-t)$$

if $s \leq t \leq s+1$.

Applying symmetry,

$$r_Y(s,t)+\begin{cases} \lambda(1-|t-s|) & \text{|t-s|<1} \\ 0 & \text{otherwise} \end{cases}$$

∎

38.9. Theorem.

Suppose that $\mu_X(t)$ is continuous second order process and that $r_X(s,t)$ is jointly continuous in (s,t). Then $X(t)$ is continuous mean-square, i.e.,

$$\lim_{s\to t} E\Big((X(s)-X(t))^2\Big)=0$$

Proof.

$$\begin{aligned}
E\Big((X(s)-X(t))^2\Big) &= \Big(E(X(s))-E(X(t))\Big)^2 + \operatorname{var}(X(s)-X(t)) \\
&= \Big(E(X(s))-E(X(t))\Big)^2 + \cdots \\
&\quad \cdots + \operatorname{var}(X(s)) - 2\operatorname{cov}(X(s),X(t)) + \operatorname{var}(X(t)) \\
&= \Big(\mu_x(s)-\mu_x(t)\Big)^2 + \cdots \\
&\quad \cdots + r_X(s,s) - 2r_X(s,t) + r_X(t,t)
\end{aligned}$$

The latter goes to zero as $s \longrightarrow t$ by virtue of the continuity of μ_X and r_X. ∎

38.10. Definition.

A process is **Gaussian** *if, given any collection of*

$$\begin{array}{rl} \textit{times} & t_1, \cdots, t_n \\ \textit{and scalars} & \alpha_1, \cdots, \alpha_n \end{array}$$

then

$$\sum_i \alpha_i X(t_i)$$

is normally distributed.

38.11. Example.

The earlier example in this section

$$X(t) = \cos(\lambda t) Z_1 + \sin(\lambda t) Z_2.$$

where Z_1 and Z_2 are independent, normally distributed random variables having mean 0 and standard deviation σ is a Gaussian process.

38.12. Definition.

Two processes $X(t)$ and $Y(t)$ are said to have the same distribution functions if for any collection of times $t_1, \cdots, t_n$ the random vectors

$$\Big(X(t_1), \cdots, X(t_n)\Big)$$

and

$$\Big(Y(t_1), \cdots, Y(t_n)\Big)$$

have the same joint distribution functions.

Note that if $X(t)$ and $Y(t)$ are Gaussian and if $\mu_X(t) = \mu_Y(t)$ and $r_X = r_Y$ then it can be shown that $X(t)$ and $Y(t)$ have the same distribtuion functions.

38.13. Definition.

A process $X(t)$ is said to be **strictly stationary** *if for all $\tau \in \mathbb{R}$ the processes $X(t)$ and $X(t+\tau)$ have the same joint distribution functions.*

Note that a second order process that is strictly stationary will necessarily also be second order stationary. The converse is not necessarily true.

Generally it is quite difficult to write down what a Gaussian process might look like. There are several important examples of processes that turn out to be Gaussian that arise in applications. The one that we shall study in detail is the Wiener Process, also known as the Wiener-Levy Process.

38.14. Definition. Wiener Process.

Suppose that a process $W(t)$ satisifies the following conditions:

(i) W(0)=0;

(ii) If $s \leq t$ then $W(t) - W(s)$ is normally distributed with mean zero and variance $\sigma^2(t-s)$;

(iii) If $t_1 \leq t_2 \leq \cdots \leq t_n$ then

$$W(t_2) - W(t_1), \cdots, W(t_n) - W(t_{n-1})$$

are independent.

Small particles suspended in a fluid exhibit an erratic, jittery motion know as $Brownian\ Motion$. This motion was probably first reported by Jan Ingenhousz in 1765 while observing carbon dust on alcohol. However, the discovery of Brownian motion is generally credited to the botanist Robert Brown who in 1827 reported that small particles within the vacuoles of the pollen grains exhibited a jittery motion. Brown repeated his observations with particles of dust and thus ruled out the motion being due to a living organism. The cause of the motion remained a major unsolved mystery througouth the nineteenth century.

The Wiener Process was introduced (in a somewhat different form) by Einstein in 1905 to provide an explanation of Brownian motion. Einstein reasoned that if the kinetic theory of gas were correct, then a particle suspended in fluid would be subject to random collisions from the molecules of the fluid. These millions of random collisions, of random magnitude and in random directions, would thus cause the jitter Brown observed. Thus the position of the particles can be described by coordinates

$$\Big(x(t) + W_1(t), y(t) + W_2(t), z(t) + W_3(t)\Big)$$

where $W_i(t)$ is a random variable satsifying conditions that turn out to be equivalent to the above. Assumption (i) corresponds to choosing a coordinate system, assumption (ii) applies the central limit theorem to assert that, on average, the cumulative effect of the observations will be normally distributed. Assumption (iii) says that jitter over disjoint time intervals is independent. In addition Einstein related σ^2 to Avogadro's number. Jean Perrin subsequently provided experimental verification of the new Einstein model. The atomic theory of matter was still controversial at this time, and so Einstein's theory not only solved a long-standing problem in physics but, coupled with Perrin's work, ended the debate over atoms and molecules. Indeed, of the five revolutionary papers published by Einstein in 1905, it was the paper on Brownin Motion that was recognized by the Nobel Prize Committee.

sample functions

The modern mathematical theory of the process, introduced by Einstein, was developed by Norbert Weiner and Paul Levy and so the process is most frequently called the Weiner-Levy Process or the Weiner Process. Independently Louis Bachelier introduced a similar process in 1900 in connection with the study of financial markets.

Since the Wiener process $W(t)$ is a random variable, there is of course a probability space $(\Omega, \mathcal{E}, \mathfrak{Pr})$ underlying the model with

$$W(t) : \Omega \to \mathbb{R}$$

for each t. Further since $W(t)$ describes the motion of a particle, $W(t)$ cannot have instantenous jumps. This leads to another technical assumption associated with the Wiener Process, namley that the *sample functions* $W(t, \omega)$ are continuous in t for each fixed $\omega \in \Omega$. More generally, it is only necessary that the sample functions be piece-wise continuous, and so the following technical assumption is usually included in the defining axioms for a Wiener process:

(iv) For each fixed $\omega \in \Omega$ the sample functions $W(\cdot, \omega)$ are piecewise continouus.

We will use the above assumption when we consider the integration of the Wiener Process.

38.15. Definition.

A process that satisfies (iv) above is said to have **piecewise continous sample functions**.

38.16. Example.

The mean and covariance functions for the Wiener Process satisfy $\mu_W(t) =$ *and*

$$r_W(s,t) = \begin{cases} \sigma^2 \min\{|s|, |t|\} & \textit{if } st \geq 0 \\ 0 & \textit{otherwise} \end{cases}$$

Solution. This can be deduced in exactly the same fashion as the similar formulae for the Poisson Process and so is left to the exercises (see problem one at the end of this section).

∎

The following formula will be useful in later applications.

38.17. Example.

If $s \geq a$ and $t \geq a$, then

$$\boldsymbol{E\Big([W(t) - W(a)][W(s) - W(a)]\Big) = \sigma^2 \min\{(t-s),(t-a)\}}$$

Solution. Suppose, for example, that $s \geq t \geq a \geq 0$ and apply the previous example:

$$\begin{aligned} E\Big([W(t) - W(a)][W(s) - W(a)]\Big) &= E(W(t)W(s)) - E(W(t)W(a)) - E(W(a)W(s)) + E(W^2(a)) \\ &= \sigma^2 t - \sigma^2 a - \sigma^2 a + \sigma^2 a \\ &= \sigma^2(t-a) \\ &= \sigma^2 \min\{(t-a),(t-s)\} \end{aligned}$$

The other cases are similar. ∎

38. Second Order Processes: Problems.

1. Set

$$X(t) = \frac{W(t+\epsilon) - W(t)}{\epsilon} \qquad -\infty < t < \infty$$

where $\epsilon > 0$ is a constant. Show that $X(t)$ is a stationary Gaussian process having covariance function

$$r_X(t) = \begin{cases} \frac{\sigma^2}{\epsilon}\left(1 - \frac{|t|}{\epsilon}\right) & |t| < \epsilon \\ 0 & \text{otherwise} \end{cases}$$

2. Find the mean and covariance of each of the following
(a) $X(t) = W^2(t)$ for $t \geq 0$
(b) $X(t) = tW(1/t)$ for $t > 0$
(c) $X(t) = \gamma^{-1}W(\gamma^2 t)$ for $t \geq 0$
(d) $X(t) = W(t) - tW(1)$ for $0 \leq t \leq 1$

39. Integration and Differentiation of Processes

sample function

Recall that a process $X(t)$ is actually a collection of random variables indexed by t that map a common probability space $(\Omega, \mathcal{E}, \mathfrak{Pr})$ to the real numbers $\mathbb{R}$. Then for each fixed $\omega \in \Omega$ the process defines a function in t:

$$X(\cdot, \omega) : \mathbb{R} \longrightarrow \mathbb{R}$$

This is called the **sample function** since it 'samples' the random path $X(t)$ along the section defined by ω. These sample functions must have some regularity assumptions in order for integration along the random path $X(t)$ to be well-defined. For our purposes it will be sufficient for the sample functions to be piecewise continuous.

39.1. Definition.

A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is said to be piecewise continuous if
(i) for each t, $f(s)$ has a finite limit as s approaches t from the right;
(ii) for each t, $f(s) \longrightarrow f(t)$ as $s \longrightarrow t$ from the left;
(iii) on any interval $[a, b]$ the function $f(t)$ has only finitely many discontinuities.

Notice in particular that jump processes and the Weiner Process will satisfy the above conditions. Piecewise continuity is suffient to assure that the approximating sums

$$\sum_{i=1}^{n} f(t^*)(t_i - t_{i-1})$$

converge to $\int_a^b f(t)\,dt$ as the mesh $\max\{t_i - t_{i-1}\}$ tends to zero. Thus if $X(t)$ has piecewise continuous sample functions, then for each fixed $\omega \in \Omega$

$$\sum_{i=1}^{n} X(t^*, \omega)(t_i - t_{i-1})$$

converge to $\int_a^b X(t, \omega)\,dt$ as the mesh $\max\{t_i - t_{i-1}\}$ tends to zero. This implies that the random variables

$$\sum_{i=1}^{n} X(t^*)(t_i - t_{i-1})$$

converge to a random variable $\int_a^b X(t)\,dt$ as the mesh $\max\{t_i - t_{i-1}\}$ tends to zero.

In the sequel we will assume that all processes we consider have piecewise continuous sample functions. We will also assume that the expectation operator $E(\cdot)$ and the integral operator can be interchanged, so that for example

$$E\left(\int_a^b X(t)\,dt\right) = \int_a^b E(X(t))\,dt.$$

Thus if $X(t)$ has mean function $\mu_X(t)$ and covariance function $r_X(s,t)$ and if $Y = \int_a^b X(t)\,dt$, then

$$\mu_Y = E\left(\int_a^b X(t)\,dt\right) = \int_a^b \mu_X(t)\,dt$$

and

$$\mathbf{cov}\left(\int_a^b f(t)X(t)\,dt \int_c^d g(s)X(s)\,ds\right) = \int_a^b \int_c^d f(t)g(s)r_X(s,t)\,ds\,dt$$

Further note that if $X(t)$ is Gaussian, then the approximating sums

$$\sum_{i=1}^{n} X(t^*)(t_i - t_{i-1})$$

must all be normaly distributed and hence the limiting random variable

$$\int_a^b X(t)\,dt$$

must also be normally distributed.

39.2. Definition.

A process $X(t)$ is **differentiable** *if there is a second order process $Y(t)$ satisfying*

$$X(t) - X(t_0) = \int_{t_0}^{t} Y(s)\,ds.$$

The process Y is called the derivative of X and is denoted $\dot{X}(t)$, so we may write

$$X(t) - X(t_0) = \int_{t_0}^{t} \dot{X}(s)\,ds.$$

Not surprisingly one can readily deduce the following.

39.3. Proposition.

Let $X(t)$ be a differentiable process. Then

$$r_{\dot{X}}(s,t) = \frac{\partial}{\partial s\, \partial t} r_X(s,t).$$

39.4. Example.

Let $X(t)$ be a second order stationary differentiable process. Then $X(t)$ and $\dot{X}(t)$ are uncorrelated.

Proof. Note that

$$r_{X\dot{X}}(s,t) = \frac{\partial}{\partial t} r_X(t-s) = \dot{r}_X(t-s)$$

so

$$r_{X\dot{X}}(t,t) = \dot{r}_X(0).$$

Since $r_X(t) = r_X(-t)$, it follows that

$$\dot{r}_X(t) = -\dot{r}_X(-t)$$

which then implies that $r_{X\dot{X}}(t,t) = 0$ upon take $t = 0$.

∎

Since the derivative $\dot{X}(t)$ is a random variable, the difference quotients

$$\frac{X(t+h) - X(t)}{h}$$

might not converge point-wise (i.e., in terms of sampling functions) to the the derivative. However, we can conclude that the difference quotients converge mean-square.

39.5. Theorem.

Suppose that $X(t)$ is a differentiable second order process. Then the quotients

$$\frac{X(t+h)-X(t)}{h}$$

converge in mean square to $\dot{X}(t)$ as $h \to 0$.

Proof. Note that we can write

$$\begin{aligned}\frac{X(t+h)-X(t)}{h} - \dot{X}(t) &= \frac{1}{h}\int_t^{t+h} \dot{X}(s)\,ds - \dot{X}(t)\\ &= \frac{1}{h}\int_t^{t+h} \dot{X}(s) - \dot{X}(t)\,ds\end{aligned}$$

Hence

$$\begin{aligned}&E\Bigg(\Big(\frac{X(t+h)-X(t)}{h} - \dot{X}(t)\Big)^2\Bigg)\\ &\quad= E\left(\frac{1}{h}\int_t^{t+h} \dot{X}(u) - \dot{X}(t)\,du\frac{1}{h}\int_t^{t+h} \dot{X}(v) - \dot{X}(t)\,dv\right)\\ &\quad= \frac{1}{h^2}\int_t^{t+h}\int_t^{t+h} E\Big((\dot{X}(u)-\dot{X}(t))(\dot{X}(v)-\dot{X}(t))\Big)\,du\,dv\\ &\qquad\text{(applying Cauchy-Shwarz)}\\ &\quad\le \frac{1}{h^2}\int_t^{t+h}\int_t^{t+h} E\Big((\dot{X}(u)-\dot{X}(t))^2\Big)^{1/2} E\Big((\dot{X}(v)-\dot{X}(t))^2\Big)^{1/2}\,du\,dv\end{aligned}$$

Now since second order processes are continuous mean-square, it follows that given any $\epsilon > 0$ there is a $\delta > 0$ so that

$$E\Big((\dot{X}(u)-\dot{X}(t))^2\Big)^{1/2} \le \epsilon$$

and

$$E\Big((\dot{X}(v)-\dot{X}(t))^2\Big)^{1/2} \le \epsilon$$

if $|u-t| \leq \delta$ and $|v-t| \leq \delta$. Thus if $h \leq \delta$ then

$$
\begin{aligned}
&\frac{1}{h^2} \int_t^{t+h} \int_t^{t+h} E\left((\dot{X}(u) - \dot{X}(t))^2\right)^{1/2} E\left((\dot{X}(v) - \dot{X}(t))^2\right)^{1/2} du\, dv \\
&\quad \leq \frac{1}{h^2} \int_t^{t+h} \int_t^{t+h} \epsilon^2 \\
&\quad = \epsilon^2
\end{aligned}
$$

Since $\epsilon > 0$ was arbitrary, this is sufficient to prove the result. ▮

As a consequence of the above theorem, it can be shown that if $X(t)$ is a Gaussian process then $\dot{X}(t)$ is normally distributed.

We note that the Wiener process is **not** differentiable. This is readily seen from the fact that

$$E\left(\left(\frac{W(t+h) - W(t)}{h}\right)^2\right) = \frac{\sigma^2}{h}.$$

Thus while the Wiener process has continuous sample functions (since the process describes the motion of particles), the motion is not smooth! Given that the process models the 'wiggles' due to millions of random collisions with molecules, it is not surprising that the resulting paths would be continuous but not differentiable. We remark in passing that the explicit construction of continuous everywhere, differentiable nowhere functions can be a challenging but not impossible task. The earliest, and most well-known, such function was given by Weierstrauss in 1872:

$$f(x) = \sum_{n=0}^{\infty} a^n \cos(b^n \pi x)$$

where $0 < a < 1$, b is a positive odd integer and

$$ab > 1 + \frac{3}{2}\pi.$$

While the Wiener process is not differentiable, it has the following regularity property.

39.6. Theorem.

Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a differentiable function. Then

$$\lim_{\epsilon \to 0} \int_a^b f(t) \frac{W(t+\epsilon) - W(t)}{\epsilon}\, dt$$

exists and equals

$$f(b)W(b) - f(a)W(a) - \int_a^b W(t)\dot{f}(t)\, dt.$$

Note that if $\tilde{W}$ is a differentiable function, then this is exactly what one would expect to get by integrating by parts:

$$\begin{aligned}\lim_{\epsilon \to 0} \int_a^b f(t) \frac{\tilde{W}(t+\epsilon) - \tilde{W}(t)}{\epsilon}\, dt &= \int_a^b \underbrace{f(t)}_{u}\, \underbrace{\tilde{W}'(t)\, dt}_{dv} \\ &= \underbrace{f(t)}_{u}\, \underbrace{\tilde{W}(t)}_{v}\Big|_{t=a}^{b} - \int_a^b \underbrace{\tilde{W}(t)}_{v}\, \underbrace{f'(t)\, dt}_{du}\end{aligned}$$

Proof. The proof actually follows by integrating by parts:

$$\begin{aligned}&\int_a^b f(t) \frac{W(t+\epsilon) - W(t)}{\epsilon}\, dt \\ &\qquad = \int_a^b \underbrace{f(t)}_{u}\, \underbrace{\frac{d}{dt}\frac{1}{\epsilon}\int_t^{t+\epsilon} W(s)\, ds\, dt}_{dv} \\ &\qquad = \underbrace{f(t)}_{u}\, \underbrace{\frac{1}{\epsilon}\int_t^{t+\epsilon} W(s)\, ds}_{v}\Big|_{t=a}^{t=b} - \int_a^b \underbrace{\frac{1}{\epsilon}\int_t^{t+\epsilon} W(s)\, ds}_{v}\, \underbrace{f'(t)\, dt}_{du} \\ &\qquad \to f(t)W(t)\Big|_{t=a}^{t=b} - \int_a^b f'(t)W(t)\, dt \quad \text{as } \epsilon \to 0\end{aligned}$$

The last equalities follow from the fact that $W(t)$ has contininous sample functions. ∎

Because of the above, we are led to the following definition.

39.7. Definition.

Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a differentiable function. Then we define

$$\int_a^b f(t)\, dW(t)$$

to be

$$f(b)W(b) - f(a)W(a) - \int_a^b W(t)\dot{f}(t)\, dt.$$

The expression $dW(t)$ in the integrand is sometimes referred to as white noise. From the standpoint of measure theory, $dW(t)$ is a signed measure and the integral can be thought of as a Stieltjes integral.

Integrating dW provides a way to approximate the differential when integrating along a path. Thus when solving the basic motion equation

$$\dot{x} = \alpha x + F$$

where f represents an external forcing function, we can solve using variation of parameters even when the forcing function F is white noise which is not, properly speaking, a function at all but rather a signed measure. We will make this explicit in the next section.

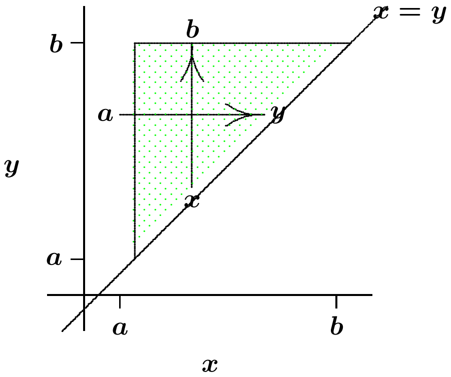
The following theorem enables one to change the order of integration when integrating dW. While it may seem intuitively obvious, notice that the integrals in question depend on the rather unusual definition above and hence this must be deduced from the definitions.

39.8. Theorem.

Let $f(x, y)$ be a jointly continuous real-valued function and let $a < b$. Then

$$\int_a^b \int_a^y f(x, y)\, dW(x)\, dy = \int_a^b \int_x^b f(x, y)\, dy\, dW(x)$$

This theorem just says that the usual change of variables formula still works when one of the iterated integrals is $dW(x)$ rather than just dx. This follows in a straightforward manner from the definition of integration $dW(x)$.



Proof. Observe that

$$\begin{aligned}\int_a^y f(x,y)\,dW(x) &= f(x,y)W(x)|_{x=a}^{y} - \int_a^y f_x(x,y)W(x)\,dx \\ &= f(y,y)W(y) - f(a,y)W(a) - \int_a^y f_x(x,y)W(x)\,dx\end{aligned}$$

Thus

$$\begin{aligned}&\int_a^b \int_a^y f(x,y)\,dW(x)\,dy = \\ &\quad= \int_a^b f(y,y)W(y) - f(a,y)W(a)\,dy - \int_a^b \int_a^y f_x(x,y)W(x)\,dx\,dy \\ &\quad= \int_a^b f(x,x)W(x) - f(a,x)W(a)\,dx - \int_a^b \int_x^b f_x(x,y)W(x)\,dy\,dx \quad (*)\end{aligned}$$

On the other hand

$$\begin{aligned}&\int_a^b \int_x^b f(x,y)\,dy\,dW(x) = \\ &\quad= \int_x^b f(x,y)\,dyW(x)|_{x=a}^{b} - \int_a^b W(x)\frac{d}{dx}\left(\int_x^b f(x,y)\,dy\right)\,dx \\ &\quad= -\int_a^b f(x,y)dy \cdot W(a) + \int_a^b W(x)\frac{d}{dx}\left(\int_x^b f(x,y)\,dy\right)\,dx \qquad (**)\end{aligned}$$

Now obsserve that

$$
\begin{aligned}
&\frac{d}{dx}\left(\int_x^b f(x,y)\,dy\right) = \\
&\quad \lim_{h\to 0}\frac{1}{h}\left(\int_{x+h}^b f(x+h,y)\,dy - \int_x^b f(x,y)\,dy\right) \\
&\quad \lim_{h\to 0} = \frac{1}{h}\left(\int_{x+h}^b f(x+h,y)\,dy \pm \int_x^b f(x+h,y)\,dy - \int_x^b f(x,y)\,dy\right) \\
&\quad \lim_{h\to 0} = \frac{1}{h}\int_{x+h}^x f(x+h,y)\,dy + \frac{1}{h}\int_x^b f(x+h,y) - f(x,y)\,dy \\
&\qquad\quad \text{(applying the fundamental theorem of calculus to the first term)} \\
&\quad = -f(x,x) + \int_x^b f_x(x,y)\,dy
\end{aligned}
$$

Substituting this into $(**)$ gives

$$\int_a^b \int_x^b f(x,y)\,dy\,dW(x) =$$

$$= -\int_a^b f(a,y)\,dy \cdot W(a) - \int_a^b W(x)\left(\int_x^b f_x(x,y)\,dy - f(x,x)\right)\,dx$$

$$= -\int_a^b f(a,x)\,dx \cdot W(a) - \int_a^b W(x)\int_x^b f_x(x,y)\,dy\,dx + \int_a^b W(x)f(x,x)\,dx$$

$$= \int_a^b f(x,x)W(x) - f(a,x)W(a)\,dx - \int_a^b \int_x^b f_x(x,y)W(x)\,dy\,dx \qquad (***)$$

Since $(*)$ and $(***)$ agree, this proves the theorem.

∎

39. Integration and Differentiation of Processes: Problems.

1. Find the covariance function for each of the following processes $X(t)$

(a)

$$X(t) = \int_0^t s\, dW(s) \quad \text{for } \; t \geq 0.$$

(b)

$$X(t) = \int_0^t \cos(ts)\, dW(s) \quad \text{for } \; -\infty < t < \infty.$$

(c)

$$X(t) = \int_{t-1}^t (t-s)\, dW(s) \quad \text{for } \; -\infty < t < \infty.$$

40.1. Theorem.

Let $\boldsymbol{f}$ and $\boldsymbol{g}$ be piece-wise differentiable functions and let $\boldsymbol{a < b}$. Then

$$\boldsymbol{E\left(\int_a^b f(t)dW(t)\int_a^b g(t)dW(t)\right) = \sigma^2\int_a^b f(t)g(t)\,dt.}$$

Proof. Applying the definition of the integral and observing

$$\begin{aligned}\int_a^b f(t)dW(t) &= f(t)W(t)|_{t=a}^b - \int_a^b f'(t)W(t)\,dt\\ &= f(b)W(b) - f(a)W(a) - f(b)W(a) + f(b)W(a) - \int_a^b f'(t)W(t)\,dt\\ &= f(b)\left(W(b) - W(a)\right) + \left(f(b) - f(a)\right)W(a) - \int_a^b f'(t)W(t)\,dt\\ &= f(b)\left(W(b) - W(a)\right) + \int_a^b f'(t)W(a)\,dt - \int_a^b f'(t)W(t)\,dt\\ &= f(b)\left(W(b) - W(a)\right) - \int_a^b f'(t)(W(t) - W(a))\,dt\end{aligned}$$

Thus, applying the above to each of the integrals inside the expectation operator, we obtain

$$\begin{aligned}&E\left(\int_a^b f(t)dW(t)\int_a^b g(t)dW(t)\right)\\ &\quad= E\left[\left(f(b)(W(b) - W(a)) - \int_a^b f'(t)(W(t) - W(a))\,dt\right)\times\cdots\right.\\ &\qquad\left.\cdots\times\left(g(b)(W(b) - W(a)) - \int_a^b g'(t)(W(t) - W(a))\,dt\right)\right]\\ &\quad= E\left[f(b)(W(b) - W(a))g(b)(W(b) - W(a))\right] + \cdots \qquad (A)\end{aligned}$$

$$\cdots - E\left[f(b)(W(b)-W(a))\int_a^b g'(t)(W(t)-W(a))\,dt\right] + \cdots \tag{B}$$

$$\cdots - E\left[g(b)(W(b)-W(a))\int_a^b f'(t)(W(t)-W(a))\,dt\right] + \cdots \tag{C}$$

$$\cdots + E\left[\left(\int_a^b f'(t)(W(t)-W(a))\,dt\right)\left(\int_a^b g'(t)(W(t)-W(a))\,dt\right)\right] \tag{D}$$

We will calculate each of the four terms (A)-(D) separately.

For (A), we recall example 38.16 which states

$$r_W(s,t) = \begin{cases} \sigma^2 \min\{|s|,|t|\} & \text{if } st \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and obtain

$$\begin{aligned} (A) &= f(b)g(b)E\big((W(b)-W(a))^2\big) \\ &= \sigma^2 f(b)g(b)(b-a) \\ &= \sigma^2 \int_a^b f(b)g(b)\,dt \end{aligned}$$

For (B), we recall 38.17 which states that if $s \geq a$ and $t \geq a$, then

$$E\Big([W(t)-W(a)][W(s)-W(a)]\Big) = \sigma^2 \min\{(t-s),(t-a)\}$$

from which we can deduce

$$
\begin{aligned}
(B) &= -E\left[f(b)(W(b)-W(a))\int_a^b g'(t)(W(t)-W(a))\,dt\right]\\
&= -f(b)\int_a^b g'(t)E\Big[(W(b)-W(a))(W(t)-W(a))\Big]\,dt\\
&= -\sigma^2 f(b)\int_a^b (t-a)g'(s)\,dt\\
&\quad \text{(integrating by parts)}\\
&= -\sigma^2 f(b)\left[(t-a)g(t)\Big|_{t=a}^{b} - \int_a^b g(t)\,dt\right]\\
&= -\sigma^2 f(b)\left[(b-a)g(b) - \int_a^b g(t)\,dt\right]\\
&= -\sigma^2 f(b)\left(\int_a^b g(b)\,dt - \int_a^b g(t)\,dt\right)\\
&= \sigma^2 \int_a^b f(b)(g(t)-g(b))\,dt
\end{aligned}
$$

In exactly the same manner,

$$
(C) = \sigma^2 \int_a^b g(b)(f(t)-f(b))\,dt.
$$

The analysis of (D) is only slightly more complex. First note that

$$
\begin{aligned}
(D) &= E\left[\left(\int_a^b f'(t)(W(t)-W(a))\,dt\right)\left(\int_a^b g'(t)(W(t)-W(a))\,dt\right)\,dt\right]\\
&= E\left[\left(\int_a^b f'(t)(W(t)-W(a))\,dt\right)\left(\int_a^b g'(s)(W(s)-W(a))\,ds\right)\right]\\
&= \int_a^b f'(t)\int_a^b g'(s)E\Big[(W(t)-W(a))(W(s)-W(a))\Big]\,ds\,dt\\
&= \sigma^2\int_a^b f'(t)\int_a^b g'(s)\min((t-a),(s-a))\,ds\,dt
\end{aligned}
$$

We can then re-write the inner integral as

$$\begin{aligned}
\int_a^b g'(s)\min((t-a),(s-a))\,ds &= \int_a^t g'(s)(s-a)\,ds + \int_t^b g'(s)(t-a)\,ds \\
&\text{(integrating the first term by parts)} \\
&= (s-a)g(s)\Big|_{s=a}^{t} - \int_a^t g(s)\,ds + (t-a)(g(b)-g(t)) \\
&= (t-a)g(b) - \int_a^t g(s)\,ds \\
&= \int_a^t (g(b)-g(s))\,ds
\end{aligned}$$

Now substitute this back into the above expression for (D):

$$\begin{aligned}
\boldsymbol{(D)} &\boldsymbol{= \sigma^2 \int_a^b f'(t) \int_a^t (g(b)-g(s))\,ds\,dt} \\
&\text{(changing the order of integration)} \\
&\boldsymbol{= \sigma^2 \int_a^b (g(b)-g(s)) \int_s^b f'(t)\,dt\,ds} \\
&\boldsymbol{= \sigma^2 \int_a^b (g(b)-g(s))(f(b)-f(s))\,ds}
\end{aligned}$$

Now, adding up the results,

$$\begin{aligned}(A)+(B)+(C)+(D) &= \sigma^2 \int_a^b f(b)g(b)\,dt + \cdots \\ &\cdots + \sigma^2 \int_a^b f(b)(g(t)-g(b))\,dt + \cdots \\ &\cdots + \sigma^2 \int_a^b g(b)(f(t)-f(b))\,dt + \cdots \\ &\cdots + \sigma^2 \int_a^b (g(b)-g(t))(f(b)-f(t))\,dt \\ &= \sigma^2 \int_a^b \Big(f(b)g(b) + \cdots \\ &\cdots + f(b)g(t) - f(b)g(b)) + \cdots \\ &\cdots + g(b)f(t) - g(b)f(b) + \cdots \\ &\cdots + (g(b)f(b) - g(b)f(t) - g(t)f(b) + g(t)f(t) \Big)\,dt \\ &= \sigma^2 \int_a^b g(t)f(t)\,dt\end{aligned}$$

■

40.2. Corollary.

If $a \le b \le c \le d$ then

$$E\left[\int_a^b f(t)\,dW(t) \int_c^d g(t)\,dW(t)\right] = 0 \qquad (40.1)$$

and

$$E\left[\int_a^b f(t)\,dW(t) \int_a^c g(t)\,dW(t)\right] = \sigma^2 \int_a^b f(t)g(t)\,dt. \qquad (40.2)$$

Proof. To verify (40.1),

$$\begin{aligned}
&E\left(\int_a^b f(t)dW(t)\ \int_c^d g(t)dW(t)\right)\\
&\qquad = E\left[\left(f(b)(W(b)-W(a))-\int_a^b f'(t)(W(t)-W(a))\,dt\right)\times\cdots\right.\\
&\qquad\quad \left.\cdots\times\left(g(c)(W(d)-W(c))-\int_c^d g'(t)(W(t)-W(c))\,dt\right)\right]\\
&\qquad = E\left[f(b)(W(b)-W(a))g(c)(W(d)-W(c))\right]+\cdots && (A)\\
&\qquad\quad \cdots - E\left[f(b)(W(b)-W(a))\int_c^d g'(t)(W(t)-W(c))\,dt\right]+\cdots && (B)\\
&\qquad\quad \cdots - E\left[g(d)(W(d)-W(c))\int_a^b f'(t)(W(t)-W(a))\,dt\right]+\cdots && (C)\\
&\qquad\quad \cdots + E\left[\left(\int_a^b f'(t)(W(t)-W(a))\,dt\right)\left(\int_c^d g'(t)(W(t)-W(d))\,dt\right)\right] && (D)
\end{aligned}$$

Because of 38.14(ii) and (iii), it follows that each term in the sum is zero.

For the second conclusion, note that

$$\int_a^c g(t)\,dW(t)=\int_a^b g(t)\,dW(t)+\int_b^c g(t)\,dW(t).$$

From this

$$\begin{aligned}
&E\left[\int_a^b f(t)\,dW(t)\int_a^c g(t)\,dW(t)\right]=\\
&\qquad = E\left[\int_a^b f(t)\,dW(t)\left(\int_a^b g(t)\,dW(t)+\int_b^c g(t)\,dW(t)\right)\right]\\
&\qquad = \sigma^2\int_a^b f(t)g(t)\,dt+0
\end{aligned}$$

applying the above theorem and the first conclusion.

■

40.3. Example.

For $t \geq 0$ and λ real constant, let $X(t)$ be the process defined by

$$X(t) = \int_0^t e^{\lambda(t-\xi)}\, dW(\xi).$$

Find the mean and covariance functions of $X(t)$.

Solution. The process clearly has zero means. For $0 \leq s \leq t$, the covariance function is

$$\begin{aligned}
E[X(s)X(t)] &= E\left[\int_0^s e^{\lambda(s-\xi)}\, dW(\xi) \int_0^t e^{\lambda(t-\xi)}\, dW(\xi)\right] \\
&= e^{\lambda(s+t)} E\left[\int_0^s e^{-\lambda\xi}\, dW(\xi) \int_0^t e^{-\lambda\xi}\, dW(\xi)\right] \\
&= \sigma^2 e^{\lambda(s+t)} \int_0^s e^{-2\lambda\xi}\, d\xi \\
&= \sigma^2 e^{\lambda(s+t)} \left(\frac{1 - e^{-2\lambda s}}{2\lambda}\right) \\
&= \frac{\sigma^2}{2\lambda}\left(e^{\lambda(s+t)} - e^{\lambda(t-s)}\right).
\end{aligned}$$

By symmetry, for $s, t \geq 0$,

$$r_X(s,t) = \frac{\sigma^2}{2\lambda}\left(e^{\lambda(s+t)} - e^{\lambda|t-s|}\right).$$

∎

40. Integration and Expectation of Processes: Problems.

1. Let $X(t)$ be defined by

$$X(t) = \int_0^t e^{\lambda(t-\xi)}\, dW(\xi),$$

and let $Y(t)$ be

$$Y(t) = \int_0^t X(s)\, ds \qquad t \geq 0.$$

(a) Show that

$$Y(t) = \int_0^t \left(\frac{e^{\lambda(t-\xi)} - 1}{\lambda} \right) dW(\xi)$$

for $t \geq 0$.

(b) Find $\mathrm{var}(Y(t))$.

41. Stochastic Differential Equations

Linear, constant coefficient differential equations of the form

$$\dot{X} + \lambda X = F \tag{41.1}$$

and

$$a\ddot{X} + b\dot{X} + cX = F \tag{41.2}$$

have many applications in engineering and the sciences. Generally these equations are used to model a physical system – such as a spring or an electrical circuit – to which an external force F has been applied. In this section we will be interested in solving these equations where the forcing function F is random in character.

Generally we will work with the integrated form of these equations for reasons that will become apparent shortly:

$$X(t) - X(t_0) + \lambda \int_{t_0}^{t} X(s)\, ds = \int_{t_0}^{t} F(s)\, ds$$

and

$$a\left(\dot{X}(t) - \dot{X}(t_0)\right) + b\left(X(t) - X(t_0)\right) + \int_{t_0}^{t} X(s)\, ds = \int_{t_0}^{t} F(s)\, ds$$

Without loss of generality we can take regularize these equations by taking $t_0 = 0$ through a simple translation. Writing

$$G(t) = \int_0^t F(s)\, ds$$

we can then re-write the above equations in a regularized, integrated form as

$$X(t) - X(0) + \lambda \int_0^t X(s)\, ds = G(t) \tag{41.3}$$

and

$$a\left(\dot{X}(t) - \dot{X}(0)\right) + b\left(X(t) - X(0)\right) + \int_0^t X(s)\, ds = G(t) \tag{41.4}$$

Now if G is any continuously differentiable function, then the integrated equations (41.3) and (41.4) are equivalent to (41.1) and (41.2) with $F = \dot{G}$. However, the integrated equations make sense even if G is *not* differentiable. Further, as we shall see, the standard variation of parameters technique for solving the non-integrated forms of the equations will carry over to the integrated case.

The particular equations that we will study arise when $G(t)$ is the Weiner process, i.e., we will study the equations

$$X(t) - X(0) + \lambda \int_0^t X(s)\, ds = W(t)$$

and

$$a\left(\dot{X}(t) - \dot{X}(0)\right) + b\left(X(t) - X(0)\right) + \int_0^t X(s)\, ds = W(t)$$

Since the Weiner process is not differentiable, the non-integrated forms of the equations are not well-defined, although you will often see the equations written as

$$\dot{X} + \lambda X = \dot{W}$$

and

$$a\ddot{X} + b\dot{X} + cX = \dot{W}$$

where $\dot{W}$ is referred to as "white noise." These latter equations are, of course, not well formed but are generally understood to mean the integrated form of the equations which are well-defined.

41.1. Lemma.

Suppose that $\phi(t)$ is the unique solution to the homogeneous differential equation

$$\dot{\phi} + \lambda\phi = 0$$

satisfying $\phi(0) = 1$. Set

$$X(t) = \int_0^t \phi(t - \xi) d\,W(\xi).$$

Then $X(t)$ is the solution to the non-homogeneous equation

$$X(t) - X(0) + \lambda \int_0^t X(s)\, ds = W(t) \tag{L_0}$$

satisfying $X(0) = 0$

In the above theorem, of course $\phi(t)$ is just

$$\phi(t) = e^{-\lambda t}.$$

The essential feature required for the proof is that $x(t)$ solves the homogeneous initial value problem. The approach we take will transfer readily to higher order differential equations.

Proof. Clearly $X(0) = 0$. Further

$$\begin{aligned}
\lambda \int_0^t X(s)\, ds &= \lambda \int_0^t \int_0^s \phi(t-\xi)\, dW(\xi)\, ds \\
&= \lambda \int_0^t \int_\xi^t \phi(s-\xi)\, ds\, dW(\xi) \\
&\quad \text{(making the change of variables } u \mapsto s-\xi\text{)} \\
&= \lambda \int_0^t \int_0^{t-\xi} \phi(u)\, du\, dW(\xi). \qquad (41.5.)
\end{aligned}$$

Since $\phi(t)$ solves the homogeneous initial value problem,

$$\phi(t) - 1 + \lambda \int_0^t \phi(u)\, du = 0$$

or equivalently

$$\lambda \int_0^{t-\xi} \phi(u)\, du = 1 - \phi(t-\xi).$$

Substituting this into (41.5) gives

$$\begin{aligned}
\lambda \int_0^t X(s)\, ds &= \int_0^t (1 - \phi(t-\xi))\; dW(\xi) \\
&= W(t) - W(0) - \int_0^t \phi(t-\xi)\, dW(\xi) \\
&= W(t) - 0 - X(t).
\end{aligned}$$

Rearranging and using the fact that $X(0) = 0$, this is equivalent to (L_0).

∎

The following theorem is immediate from the lemma and the fact that the solution to the initial value problem:

Langevin's equation

$$\dot{x} + \lambda x = 0 \quad x(0) = x_0.$$

is

$$x(t) = x_0 e^{-\lambda t}.$$

41.2. Theorem.

The initial value problem

$$X(t) - X(0) + \lambda \int_0^t X(s)\, ds = W(t) \qquad (L)$$
$$X(0) = x_0$$

has unique solution

$$x(t) + X(t)$$

where

$$X(t) = \int_0^t e^{-\lambda(t-\xi)}\, dW(\xi)$$

and $x(t)$ is the unique solution to

$$\begin{aligned} \dot{x} + \lambda x &= 0 \\ x(0) &= x_0 \end{aligned}$$

Langevin's equation is a variation on equation (L). If one takes m to be the molecular mass of the gas and γ to be the coefficient of friction, then

$$\lambda = \frac{\gamma}{m}.$$

In it's "differentiated" form, the equation becomes

$$\dot{v} + \frac{\gamma}{m} v = \frac{1}{m} dW(t).$$

Now, if $X(t)$ solves

$$\dot{v} + \frac{\gamma}{m} v = dW(t).$$

and if we set $Y(t) = \frac{1}{m}X(t)$, then $Y(t)$ solves the equation

$$\dot{Y} + \frac{\gamma}{m}Y = \frac{1}{m}dW(t).$$

In light of 40.3,

$$\begin{aligned}\operatorname{var}(Y(t)) &= \frac{1}{m^2}\frac{\sigma^2}{2\frac{\gamma}{m}}\left(1 - e^{-2\frac{\gamma}{m}t}\right)\\ &= \frac{\sigma^2}{2\gamma m}\left(1 - e^{\frac{-2\gamma t}{m}}\right).\end{aligned}$$

The Kinetic Theory of Gasses provides another way of analyzing the long-term distribution of the speed of molecules in an ideal gas. In this case, the speed can be thought of as the magnitude of a vector $N(t) = (n_x(t), n_y(t), n_z(t))$, where each component is normally distributed with mean zero and variance $\frac{kT}{m}$, where k is the Boltzman constant, T is the absolute temperature, and m is the molecular mass of the gas. Since the long-term variance from Langevin's equation must be $\frac{\sigma^2}{2\gamma m}$, we obtain

$$\frac{kT}{m} = \frac{\sigma^2}{2\gamma m}$$

or

$$\sigma^2 = 2kT\gamma.$$

Since Boltzman's constant is just

$$k = \frac{R}{N}$$

where R is the gas constant and N is Avogadro's number, these were the essential steps in relating Brownian motion to Avagadro's number. Perrin used Einstein's results to calculate Avogadro's number from Brownian motion experiments, for which he won the 1926 Nobel Prize. These calculations and the related experiments provided compelling evidence in support of the atomic theory of matter, something that was controversial prior to this work.

We can solve second-order, constant coefficient, forced differential equations (2_I) in an exactly similar fashion.

41.3. Theorem.

Suppose that ϕ solves the initial value problem

$$a\ddot{\phi} + b\dot{\phi} + c\phi = 0$$

$$\phi(0) = 0 \qquad \dot{\phi}(0) = \frac{1}{a}$$

Then the function $X(t)$ given by

$$X(t) = \int_0^t \phi(t-s)\, dW(s)$$

solves the initial value problem

$$a\left(\dot{X}(t) - \dot{X}(0)\right) + b\Big(X(t) - X(0)\Big) + c\int_0^t X(s)\, ds = W(t) \tag{41.6}$$

$$X(0) = \dot{X}(0) = 0$$

41.4. Corollary.

If $X(t)$ is defined as in the preceding theorem and if $\psi(t)$ solves the initial value problem

$$a\ddot{\psi} + b\dot{\psi} + c\psi = 0$$

$$\psi(0) = \psi_0 \qquad \dot{\psi}(0) = \psi_{00}$$

then $X(t) + \psi(t)$ solves the initial value problem

$$a\left(\dot{X}(t) - \dot{X}(0)\right) + b\Big(X(t) - X(0)\Big) + c\int_0^t X(s)\, ds = W(t)$$

$$X(0) = \psi_0 \qquad \dot{X}(0) = \psi_{00}.$$

Proof of the Theorem. First observe

$$\begin{aligned}\int_0^t \int_0^s \dot{\phi}(s-\xi)\,dW(\xi)\,ds &= \int_0^t \int_\xi^t \dot{\phi}(s-\xi)\,ds\,dW(\xi)\\ &= \int_0^t \int_0^{t-\xi} \dot{\phi}(s)\,ds\,dW(\xi)\\ &= \int_0^t \left(\phi(t-\xi)-\phi(0)\right)\,dW(\xi)\\ &= \int_0^t \phi(t-\xi)\,dW(\xi)\\ &= X(t)\end{aligned}$$

Thus X is differentiable and

$$\dot{X}(t) = \int_0^t \dot{\phi}(t-\xi)\,dW(\xi).$$

Substituting into the left-hand-side of (41.6) we see that

$$\begin{aligned}&a\left(\dot{X}(t)-\dot{X}(0)\right)+b\Big(X(t)-X(0)\Big)+c\int_0^t X(s)\,ds\\ &\qquad = a\int_0^t \dot{\phi}(t-\xi)\,dW(\xi)+b\int_0^t \phi(t-\xi)\,dW(\xi)+c\int_0^t X(s)\,ds\end{aligned}$$

Now

$$\begin{aligned}c\int_0^t X(s)\,ds &= c\int_0^t \int_0^s \phi(s-\xi)\,dW(\xi)\,ds\\ &= c\int_0^t \int_\xi^t \phi(s-\xi)\,ds\,dW(\xi)\\ &= c\int_0^t \int_0^{t-\xi} \phi(u)\,du\,dW(\xi)\end{aligned}$$

We can then further reduce the left-hand-side (41.6) to

$$\int_0^t \left[a\dot{\phi}(t-\xi)+b\phi(t-\xi)+c\int_0^{t-\xi}\phi(u)\,du\right]\,dW(\xi).$$

Now, integrating the homogenous problem and apply the fact that ϕ solves the homogenous problem with $\phi(0) = 0$ and $\dot{\phi}(0) = \frac{1}{a}$,

$$\begin{aligned}
0 &= a(\dot{\phi}(t) - \dot{\phi}(0)) + b(\phi(t) - \phi(0)) + c\int_0^t \phi(s)\,ds \\
&= a\left(\dot{\phi}(t) - \frac{1}{a}\right) + b(\phi(t) - 0) + c\int_0^t \phi(s)\,ds \\
&= -1
\end{aligned}$$

which implies that for any value of t

$$a\dot{\phi}(t) + b\phi(t) + c\int_0^t \phi(u)\,du = 1.$$

Hence, the left-hand-side of (41.6) further reduces to

$$\int_0^t 1\,dW(\xi) = W(t)$$

This shows that X solves (41.6) and completes the proof.

∎

41.5. Example.

Solve the following initial value problem:

$$\left(\dot{X}(t) - \dot{X}(0)\right) + 2\,(X(t) - X(0)) + 2\int_0^t X(s)\,ds = W(t)$$

$$X(0) = 0 \qquad \dot{X}(0) = 1.$$

Solution. First solve the homogeneous equation

$$\ddot{x} + 2\dot{x} + 2x = 0.$$

The characteristic equation is

$$\lambda^2 + 2\lambda + 2 =$$

which has complex roots

$$\lambda = -1 \pm i.$$

From this the general solution to the homogeneous equation is

$$x(t) = e^{-t}\left(C_1 \cos(t) + C_2 \sin(t)\right).$$

Thus, the function $\phi(t)$ that solves the homogenous initial value problem

$$\ddot{x} + 2\dot{x} + 2x = 0$$
$$x(0) = 0 \qquad \dot{x}(0) = 1$$

is

$$\phi(t) = e^{-t}\sin(t).$$

This lets us conclude that

$$X(t) = \int_0^t e^{-(t-\xi)} \sin(t-\xi)\, dW(\xi)$$

solves

$$\left(\dot{X}(t) - \dot{X}(0)\right) + 2\left(X(t) - X(0)\right) + 2\int_0^t X(s)\, ds = W(t)$$
$$X(0) = \dot{X}(0) = 0$$

and the general solution is in the form

$$e^{-t}\left(C_1\cos(t) + C_2\sin(t)\right) + \int_0^t e^{-(t-\xi)}\left(C_1\cos(t-\xi)C_2\sin(t-\xi)\right)\, dW(\xi).$$

Solving for C_1 and C_2 gives that

$$Y(t) = e^{-t}\sin(t) + \int_0^t e^{-(t-\xi)}\sin(t-\xi)\, dW(\xi)$$

solves the initial value problem.

∎

Note that

$$E(Y(t)) = e^{-t}\sin(t)$$

so that the expected value of a solution to the stochastic agrees with the unforced solution. Also

$$\begin{aligned}\text{Var}(Y(t)) &= \sigma^2 \int_0^t e^{-(t-\xi)^2}\sin^2(t-\xi)\, d\xi \\ &= \frac{\sigma^2}{8}\left[1 + e^{-2t}\left(\cos(wt) - \sin(2t)\right) - 2\right]\end{aligned}$$

Ornstein-Uhlenbeck process

41. Stochastic Differential Equations: Problems.

1. Let $X(t)$ be the solution to the equation

$$mX'(t) + fX(t) = W(t) \quad X(0) = x_0, X'(0) = v_0$$

for constants m and f. ($X(t)$ is called the Ornstein-Uhlenbeck process.)
(a) Express $X(t)$ in terms of $W(t)$.
(b) Express $X(t)$ in terms of the solution to Langevin's process.
(c) Find the mean and varience of the Ornstein-Uhlenbeck process.

2. Solve each of the following stochastic differential equations and find $\text{var}(X(t))$ for the solution having initial conditions $X(0) = 0 = X'(0)$.
(a) $X''(t) + X'(t) = W'(t)$
(b) $X''(t) + 3X'(t) + 2X(t) = W'(t)$
(c) $4X''(t) + 8X'(t) + 5X(t) = W'(t)$
(d) $X''(t) + 2X'(t) + X(t) = W'(t)$
(e) $X''(t) + X(t) = W'(t)$

Discrete Distributions.

Binomial Distribution	
Density	$\binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, \cdots, n$
$E(X)$	np
var(X)	$np(1-p)$
$\Phi_X(t)$	$(pt + (1-p))^n$

Negative Binomial Distribution	
Density	$\binom{k+r-1}{r-1} p^r (1-p)^k \quad k = 0, \cdots$
$E(X)$	$\frac{r(1-p)}{p}$
var(X)	$\frac{r(1-p)}{p^2}$
$\Phi_X(t)$	$\left(\frac{p}{1-(1-p)t}\right)^r$

Geometric Distribution	
Density	$p(1-p)^k \quad k = 0, \cdots, \infty$
$E(X)$	$\frac{1-p}{p}$
var(X)	$\frac{1-p}{p^2}$
$\Phi_X(t)$	$\frac{p}{1-(1-p)t}$

Hypergeometric Distribution	
Density	$\frac{\binom{r}{k}\binom{N-r}{N-k}}{\binom{N}{n}}, \quad \max\{0, r+n-N\} \le k \le \min\{n, r\}$
N, r and n satisfy $N \ge 0,\ 0 \le r \le N,\ 0 \le n \le N$	
$E(X)$	$\frac{nr}{N}$
var(X)	$\frac{n\frac{r}{N}\left(1-\frac{r}{N}\right)(N-n)}{N-1}$

Poisson Distribution

Density	$\frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0, \cdots, \infty$
$E(X)$	λ
var(X)	λ
$\Phi_X(t)$	$\exp(\lambda(t-1))$

Zipf Distribution

Generalized Harmonic Number	$H(N, s) = \sum_{k=1}^{N} \frac{1}{k^s}$
Density	$\frac{\frac{1}{k^s}}{H(N, s)} \quad k = 0, \cdots, N$
$E(X)$	$\frac{H(N, s-1)}{H(N, s)}$
$\Phi_X(t)$	$\frac{1}{H(N, s)} \sum_{n=1}^{N} \frac{nt}{n^s}$

Logarithmic Distribution

Density	$\frac{-1}{\ln(1-p)} \frac{p^k}{k} \quad k = 1, \cdots, \infty$
$E(X)$	$\frac{-1}{\ln(1-p)} \frac{p}{1-p}$
var(X)	$\frac{-p(p + \ln(1-p))}{(1-p)^2 \ln^2(1-p)}$
$\Phi_X(t)$	$\frac{\ln(1-pt)}{\ln(1-p)}$

Zeta Distribution

Zeta function	$\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}$
Density	$\frac{\frac{1}{k^s}}{\zeta(s)} \quad k = 1, \cdots, \infty$
$E(X)$	$\frac{\zeta(s-1)}{\zeta(s)} \quad s > 2$
$\Phi_X(t)$	$\frac{1}{\zeta(s)} \sum_{n=1}^{\infty} \frac{nt}{n^s}$

Continuous Distributions

Uniform Distribution	
Density	$\begin{cases} \frac{1}{b-a} & \text{if } a \leq t \leq b \\ 0 & \text{otherwise} \end{cases}$
$E(X)$	$\frac{a+b}{2}$
var(X)	$\frac{(b-a)^2}{12}$
$M_X(t)$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$

Normal Distribution	
Density	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$
$E(X)$	μ
var(X)	σ^2
$M_X(t)$	$\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$

Exponential Distribution	
Density	$\begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$
$E(X)$	$\frac{1}{\lambda}$
var(X)	$\frac{1}{\lambda^2}$
$M_X(t)$	$\frac{\lambda}{\lambda - t}$

Gamma and Beta Functions and Identities	
Gamma function	$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt, \quad x > 0$
$\Gamma(x+1) = x\Gamma(x) \text{and} \Gamma(1/2) = \sqrt{\pi}$	
Beta Function	$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt, \quad 0 < x, y$
$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$	

Cauchy Distribution	
Density	$\frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}$
$E(X)$	undefined
var(X)	undefined
$M_X(t)$	undefined

Gamma Distribution	
Density	$\frac{x^{\alpha-1}\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)}, \quad x > 0$
$E(X)$	$\frac{\alpha}{\lambda}$
var(X)	$\frac{\alpha}{\lambda^2}$
$M_X(t)$	$\left(\frac{\lambda}{\lambda - t}\right)^\alpha$

Chi-square Distribution

Density	$\frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-x/2}, \quad x \geq 0$
$E(X)$	k
var(X)	$2k$
$M_X(t)$	$(1-2t)^{-k/2}$

Fisher-Snedecor F Distribution

Density	$\frac{1}{xB\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \left(\frac{n_1 x}{n_1 x + n_2}\right)^{\frac{n_1}{2}} \left(1 - \frac{n_1 x}{n_1 x + n_2}\right)^{\frac{n_2}{2}}$
$E(X)$	$\frac{n_2}{n_2 - 2}$ for $n_2 > 2$
var(X)	$\frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2-2)^2(n_2-4)}$ for $n > 4$
$M_X(t)$	NA

Student's t Distribution

Density	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}(n/2)(1+x^2/2)^{(n+1)/2}}, \quad x \in \mathbb{R}$
$E(X)$	0
var(X)	$\frac{n}{n-2}$ for $n > 2$
$M_X(t)$	NA

Weibull Distribution

Density	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^{k}\right), \quad x \geq 0$
$E(X)$	$\lambda\Gamma(1 + 1/k)$
var(X)	$\lambda^2\Gamma(1+2/k) - \lambda^2\Gamma^2(1+1/k)$
$M_X(t)$	NA

Beta Distribution

Density	$\frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1$
$E(X)$	$\frac{\alpha}{\alpha+\beta}$
var(X)	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
$M_X(t)$	$1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k} \frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$

Erlang Distribution

Density	$\frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}, \quad x \geq 0$
$E(X)$	$\frac{k}{\lambda}$
var(X)	$\frac{k}{\lambda^2}$
$M_X(t)$	$\left(\frac{\lambda}{\lambda - t}\right)^k$ for $t < \lambda$

Index