Stochastic Modeling & Simulation

April 28, 2021

1 Pseudorandom Number Generation

Summer 2019 - Instructor: Josh Reed

Teaching Assistant: Haotian Song

Generating an arbitrary random variable is dependent upon first being able to simulate a random variable uniformly on the interval $[0, 1]$. This is accomplished by a *pseudorandom number generator* which we discuss now.

1.1 Pseudorandom Number Generators

Most "random numbers" generated on a computer are (somewhat surprisingly) derived from a deterministic sequence of numbers. The function which is used to produce this sequence of non-random numbers is referred to as a *pseudorandom number generator*. There are several ways to construct a pseudorandom number generator which we discuss below but all such generators have the same goal. They attempt to produce a sequence of numbers $u_1, u_2, u_3...$, which mimic a sequence of Uniform$[0, 1]$ random numbers that are independent of one another.

One example of a pseudorandom number generator is the *linear congruential generator*. In a linear congruential generator, the sequence $u_1, u_2, u_3...$, is generated by setting

$$\begin{aligned} x_{i+1} &= (ax_i + c) \bmod m, \\ u_{i+1} &= x_{i+1}/m, \end{aligned}$$

for $i = 0, 1, 2, ...$ The initial value x_0 is referred to as the *seed value* and it is chosen by the user to be an integer between 0 and $m - 1$. The value a is referred to as the *multiplier*, c is the *increment* and m is the *modulus* of the generator. All of these numbers are assumed to be integers. The mod operator takes the remainder after long division. For example $7 \bmod 2 = 1$ and $9 \bmod 3 = 0$. This implies that $x_{i+1} = (c + ax_i) \bmod m$ will always be an integer between 0 and $m - 1$, and so $u_{i+1} = x_{i+1}/m$ will always be a number between between 0 and 1 as desired.

Linear congruential generators are one of the oldest known pseudorandom number generators. They are also one of the best known and widely used pseudorandom number generators. This is because they are relatively easy to understand and may be efficiently implemented on a computer. If $c = 0$ in the above, then we may refer to the generator more specifically as a *multiplicative congruential generator* and if $c \neq 0$ we may refer to it as a *mixed congruential generator*.

In order to better understand how linear congruential generators work, let us set the multiplier a to 7, the increment c to 0 and the modulus m to 11. Starting with an initial seed value of 4, the sequence of pseudorandom numbers generated is given by

$$\begin{aligned}
x_0 &= 4 \\
x_1 &= 28 \bmod 11 = 6 \\
x_2 &= 42 \bmod 11 = 9 \\
x_3 &= 63 \bmod 11 = 8 \\
x_4 &= 56 \bmod 11 = 1 \\
x_5 &= 7 \bmod 11 = 7 \\
x_6 &= 49 \bmod 11 = 5 \\
x_7 &= 35 \bmod 11 = 2 \\
x_8 &= 14 \bmod 11 = 3 \\
x_9 &= 21 \bmod 11 = 10 \\
x_{10} &= 70 \bmod 11 = 4 \\
x_{11} &= 28 \bmod 11 = 6 \\
\vdots &= \quad \vdots \quad = \vdots
\end{aligned}$$

Notice that in the original example above since the modulus m has been set to 11, the x_i's can only take on the integer values 1 through 10 (we rule out 0 in this case because since $c = 0$ we have that if $x_i = 0$, then $x_j = 0$ for all $j > i$). Moreover, all 10 of these values appear before the sequence returns to its initial seed value of 4. This occurs at x_{10} at which the point the sequence starts to repeat itself. In this case, we say that the linear congruential generator has a *full period*. Not all generators have a full period. For instance, consider the case in which the multiplier a is 4, the increment c is 0 and the modulus m is 11. Then, the seed value of 1 generates the sequence $1, 4, 5, 9, 3, 1, 4, ...$, whereas the seed value of 2 generates the sequence $2, 8, 10, 7, 6, 2, 8, ...$

The longer the period of a generator, the more distinct numbers it will produce before repeating itself. This is desirable because it implies that that the generator can more closely mimic a uniform distribution. In fact, if the modulus m is a prime number and the increment c is 0, then the following two conditions on the multiplier a ensure that the generator has a full period.

1. $a^{m-1} - 1$ is a multiple of m.

2. a^j is not a multiple of m for $j = 1, ..., m - 2$.

The following table taken from L'Ecuyer (1998) provides the multipliers and moduli for some commonly recommended multiplicative congruential generators. All of these generators have full periods.

Multiplier	Modulus
39373	$2^{31} - 1$
16807	$2^{31} - 1$
40692	2147483399
40014	2147483563

Multiplier	Modulus
42024	2147482801

One common feature of all linear congruential generators is that all pairs of consecutive u_i's will lie on a common set of parallel lines in the unit square. This is referred to as the lattice structure of linear congruential generators and it may or may not be a problem depending on how far apart or close together the parallel lines are spaced. The spectral test for linear congruential generators was developed by Coveyou and Macpherson (1967) in order to address this issue and measures how strong the lattice effect is. It works by taking the maximum value of the distance between all sets of parallel lines that the consecutive pairs (u_i, u_{i+1}) lie on. Smaller resulting values imply more random appearing patterns. The spectral test can also be extended to higher dimensions.

2 Session 2: Probability Overview

Summer 2019 - Instructor: Josh Reed

Teaching Assistant: Haotian Song

In this session we provide an overview of the fundamental ideas of probability. These notes are a summary of Chapter 1 of Ross' *Introduction to Probability Models*.

3 Samples Spaces

One of the main motivations for studying probability is the desire to assign weights or *probabilities* to events which might occur in the future. To this end, a key concept which underlies all of probability (although it is not always up front and visible) is the concept of a sample space, which is usually denoted by $\mathcal{S}$. The sample space is the set of all possible outcomes which might occur. Subsets of the sample space are referred to as events, which are usually denoted by E.

Example. Suppose that a coin will be tossed and will land either on heads or on tails. Then, the sample space is $\mathcal{S} = \{H, T\}$ which are the two possible outcomes of tossing the coin. There are three events which may be formed. They are $\{H\}, \{T\}$ and $\{H, T\}$. The first 2 of these events represent the coin landing on heads and the coin landing on tails, respectively. The third event represents the coin landing on either heads *or* tails.

Example. Suppose that a 6-sided will be rolled. The sample space of possible outcomes is $S = \{1, 2, 3, 4, 5, 6\}$. There are many events which may be formed in this case. For instance $\{1\}$ is the event that the number 1 is rolled. On the other hand, $\{2, 4, 6\}$ is the event that an even number is rolled and $\{4, 5, 6\}$ is the event that a number 4 or larger is rolled.

Example. Suppose that a coin will be tossed two times in a row and that each time the coin will land either on heads or on tails. Then, the sample space of possible outcomes is $\mathcal{S} = \{HH, HT, TH, TT\}$ representing the possible outcomes of tossing the coin twice. There are several events which may be formed. For instanace, $\{HT, TH\}$ represents one heads and one tails being tossed, while $\{HH, TT\}$ represents either both heads or both tails.

We can also create combinations of events through the union and intersection of events. In particular, $A \cup B$ is the *union* of events A and B and represents the set of all outcomes that are either in event A or event B or both. Next, $A \cap B$ is the *intersection* of events A and B and represents the set of all outcomes that are both in event A and event B. Finally, A^C is the *complement* of event A and is the set of all outcomes that are not in event A.

Example. Suppose that again a coin will be tossed and will land either on heads or on tails. Then, the events $\{H\}$ and $\{T\}$ represent the coin landing on heads or tails, respectively. The union of these two events $\{H\} \cup \{T\} = \{H, T\}$ is a coin landing on either heads or tails. The intersection $\{H\} \cap \{T\} = \emptyset$ is the emptyset which is also paradoxically considered to be an event. The complement $(\{H\}^C = \{T\})$.

Example. Suppose that a 6-sided will be rolled. In this case, $\{2\} \cup \{4\} \cup \{6\} \cup = \{2, 4, 6\}$ is the event that an even number is rolled, and $\{2, 4, 6\}^C = \{1, 2, 3\}$ is the event that an odd number is rolled. Also, $\{1, 2, 3, 4\} \cap \{3, 4, 5, 6\} = \{3, 4\}$ is the event that a number greater than 2 but less 5 is rolled.

Example. Suppose that a coin will be tossed two times in a row and that each time the coin will land either on heads or on tails. Then, $\{HH\} \cup \{HT\} \cup \{TH\} \cup = \{HH, HT, TH\}$ is the event that the coin lands on heads at least once, and $\{HH, HT, TH\}^C = \{TT\}$ is the event that the coin never lands on heads. Also, $\{HH, HT, TH\} \cap \{HT, TH, TT\} = \{HT, TH\}$ is the event that the coin lands on heads at least once and tails at least once.

4 Assigning Probabilities to Events

One of the advantages of speaking in terms of sample spaces and events is that we can assign probabilities to each event E, denote by $P(E)$. There are 3 rules for assigning probabilities to events.

1. Every event has a probability between 0 and 1. That is, $0 \leq P(E) \leq 1$ for every event.
2. The probability that some event occurs is 1. That is, $P(S) = 1$.
3. If $E_1, E_2, E_3, ...,$ are disjoint events, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

In the above, 2 events A and B are said to be disjoint if $A \cap B = \emptyset.\backslash$

Example. Consider the example above of tossing a coin and seeing if it lands on heads or tails. If the coin is fair then $P(H) = P(T) = 1$. On the other hand, if the coin is not fair then we may that $P(H) = 4/5$. In this case, since $P(S) = 1$, it must be the cas that $P(T) = 1 - P(H) = 1/5$.

Example. Consider the example above of rolling a die. Suppose we wish to calculate the probability that the die lands on 2,3,4,5 or 6. In order to this we can use the fact that for an event E we have that $P(E) = 1 - P(E^C)$. Hence, $P(2, 3, 4, 5, 6) = 1 - P(1) = 5/6$.

Example. Consider the example above of rolling a die. Suppose we wish to calculate the probability that the die lands on 2,3,4,5 or 6. In order to this we can use the fact that for an event E we have that $P(E) = 1 - P(E^C)$. Hence, $P(2, 3, 4, 5, 6) = 1 - P(1) = 5/6.\backslash$

Example. Consider the example above of flipping a coin two times in a row. Suppose we wish to calculate the probability that at least 1 of the 2 coins lands on heads. In this case let the event E be defined by $E = \{(H, H), (H, T)\}$ and let the event F be defined by $F = \{(H, H), (T, H)\}$. In words E is the event that the first toss of the coin lands on heads and F is the event that the second toss of the coin lands on heads. We want to calculate the probability of $E \cup F$. In order to do this we can make use of the relationship

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

In particular both E and F have a $1/2$ probability since the probability of either of the coins landing on heads is $1/2$. Next, $E \cap F = \{(H, H)\}$ which has a $1/4$ probability of occurring. Hence, substituting into the above we obtain that $P(E \cup F) = 3/4$.

5 Conditional Probabilities

Consider the die rolling example above and suppose that we are told that an even number has been rolled. Now we wish to determine the probability that a 2 was rolled. How may we proceed? Since an even number was rolled this limits the outcomes of the roll of the die to to the set $\{2, 4, 6\}$. Moreover, since originally the die was equally likely to land on 1 through 6, given that we now know either a 2,4 or 6 was rolled each of these must be equally likely to have occurred too. Hence, the probability that a 2 was rolled given an even number was rolled is $1/3$.

The above example illustrates what is known as a conditional probability. In this case we are calculating the probability that a 2 is rolled conditional on an even number number being rolled. If we let $F = \{2, 4, 6\}$ be the event that an even number is rolled and $E = \{2\}$ be the event that a 2 is rolled, then the conditional probability is denoted by $P(E|F)$. There exists a general formula for calculating conditonal probability and it is given by

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Example. Suppose that cards numbered 1 through 20 are randomly shuffled and the top card is then flipped over. Given that the number of the card is at least an 8, what is probability that it is a 12? In order to answer this question, we can use conditional probabilities. Let F denote the event that the top card is at least an 8 and let E denote the event that the top card is a 12. Then, $E \cap F$ is the event that at least an 8 is drawn and that a 12 is drawn, which is equivalent to a 12 being drawn. Hence, using the conditional probability formula the desired probability is

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{1}{20}}{\frac{13}{20}} = \frac{13}{20}.$$

Example. Consider a family that has 2 children and that at least one of them is a girl. What is the probability that both children are girls? In order to answer this question, we can again use conditional probabilities. Let F denote the event that at least one of the children is a girl and let E denote the event that both children are girls. Then, $E \cap F$ is the event that both children are girls. Assuming that boys and girls are equally likely to be born, this event has probability $1/4$.

Enumerating the possible outcomes, the probability that at least one child is a girl is $3/4$. Using the conditional probability formula, the desired probability is given by

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

Example. John is deciding to whether to take a course in either machine learning or data mining. If John takes a machine learning course, he will receive an A with probability $1/3$. If John takes a data mining course, he will receive an A with probability $1/2$. In order to decide which course to take, John has decided to flip a fair coin. What is the probability that in the end John receives an A in machine learning? In order to answer this question, we can use the conditional probability formula. Let E be the event that John receives an A, regardless of the course he takes, and let F be the event that John takes the machine learning course. Then, $E \cap F$ is the event that John receives an A in the machine learning course, and rearranging the conditional probability formula we have

$$P(E \cap F) = P(F)P(E|F) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}.$$

6 Independent Events

Independence is a concept which will be important in this course. Two events are said to be independent of one another if

$$P(E \cap F) = P(E)P(F).$$

Note that E and F are independent of each other. Meaning that if E is independent of F, then F is also independent of E. Using the conditional probability formula, we also have that if E and F are independent, then

$$P(E|F) = P(E).$$

Example. Suppose that a fair coin is tossed 2 times in a row. What is the probability that it lands on heads each time? To answer this question let E_1 be the probability that it lands on heads on the first toss, and let E_2 be the probability that it lands on heads on the second toss. Then, since the 2 tosses are independent of each other, we have that

$$P(E_1 \cap E_2) = P(E_1)P(E_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

Example. Suppose that a 6 sided die is rolled two times in a row. Is the event that the first die rolls a 4 independent of the event that the sum of the two rolls is a 7? The answer to this question is not obvious so we will resort to calculating the relevant probabilities and seeing if the condition for independence holds. Let E_1 be the event that the first roll of the die is a 4 and let E_2 be the event that the sum of the two dice is 7. Now note that in order for $E_1 \cap E_2$ to occur the first roll must be a 4 and the second roll a 3. Since there are a total of $6 \times 6 = 36$ possible ordered outcomes of the

two dice rolls, we have that $P(E_1 \cap E_2) = 1/36$. On the other hand, since the die has six sides it follows that

$$P(E_1) \times P(E_2) = \frac{1}{6} \times \frac{1}{6} = 1/36$$

and so the two events are independent.

The events E, F and G are said to be independent if E and F are independent, F and G are independent, G and E are independent, and

$$P(E \cap F \cap G) = P(E)P(F)P(G).$$

In other words, if all subsets of events are independent of one another. A similar definition holds for 4 or more events.

Example. Suppose that an urn contains 4 balls labeled 1,2,3 and 4 and that each of the balls is equally likely to be drawn from the urn. Let the events $E = \{1,2\}, F = \{1,3\}$ and $G = \{1,4\}$ where the numbers denote the ball drawn from the urn. In this case one can verify that

$$\begin{aligned} P(E \cap F) &= P(E \cap F) = \frac{1}{4}, \\ P(F \cap G) &= P(F \cap G) = \frac{1}{4}, \\ P(E \cap G) &= P(E \cap G) = \frac{1}{4}. \end{aligned}$$

This implies that E, F and G are *pairwise* independent. However,

$$\frac{1}{4} = P(E \cap F \cap G) \neq P(E)(F)(G) = \frac{1}{8},$$

and so the events E, F and G are not independent.

7 Bayes' Formula

Sometimes when calculating a conditional probability $P(E|F)$, it is easier to calculate the opposite conditional probability $P(F|E)$. Luckily, it turns out that $P(E|F)$ and $P(F|E)$ may be related to one another via

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}.$$

This relationship is referred to as Baye's law.

Example. Suppose that we have two urns. The first urn contains 3 white balls and 4 black balls. The second urn contains 5 white balls and 6 black balls. Suppose that we flip a fair coin and then take a ball from the first or second urn if the coin lands on heads or tails, respectively. What is

the probability that the coin landed on heads given that a white ball was picked? To answer this question we can use Baye's rule. Let H be the event that the coin lands on heads and let W be the event that a while ball was selected. Then,

$$P(H|W) = \frac{P(W|H)P(H)}{P(W)}.$$

Now, $P(W|H) = 3/7$ and $P(H) = 1/2$. To calculate $P(W)$ we use the fact that

$$P(W) = P(W|H)P(H) + P(W|H^C)P(H^C) = \frac{34}{77},$$

from which it follows that $P(H|W) = 33/68$.

Example. A student is taking a multiple choice test and for each question she either knows the correct answer with probability p or guesses with probability $1 - p$. Each question has m choices and so a guess results in the correct answer with probability $1/m$. Given that a student answered a questions correctly, what is the probability that she guessed? Let C be the event that the student answers a question correctly and G the probability that she guessed. Then, by Baye's rule

$$P(G|C) = \frac{P(C|G)P(G)}{P(C)} = \frac{P(C|G)P(G)}{P(C|G)P(G) + P(C|G^C)P(G^C)}.$$

Now $P(C|G) = 1/m$ and $P(C|G^C) = 1$. Also $P(G) = 1 - p$ and $P(G^C) = p$. Substituting these quantities into the above we obtain that

$$P(G|C) = \frac{1-p}{1+(m-1)p}.$$

Example. A test is 99% accurate in detecting if a person is sick. However, 1% of healthy people also yield a positive result of being sick. If 0.5% of the population actually has the disease, what is the probability that a person with a positive test is sick? We can use Baye's rule as follows. Let D be the event that a person is sick and E the event that their test is positive. Then,

$$P(D|E) = \frac{P(E|D)P(D)}{P(E)} = \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|D^C)P(D^C)}.$$

Now substituting the appropriate probabilities into the above we obtain that $P(D|E)$ is approximately 83.2%.

8 Random Variables

Often times we will assign a numeric value to each outcome in the sample space. For instance if we flip a coin we might assign a 0 if it lands on tails and a 1 if it lands on heads. If we roll two fair die in a row we might sum up the two numbers that are rolled. Any function which assigns a numeric value to each outcome is the sample space is referred to as a *random variable*.

Example. Suppose that two fair die are rolled in a row and let Z be the random variable that is the sum of the two numbers that are rolled. We can then calculate the probability that Z takes on a certain value by summing up the probabilities of the events associated with the value. For instance,

$$\begin{aligned} P(Z=2) &= P((1,1)) = \frac{1}{36}, \\ P(Z=5) &= P(\{(1,4),(2,3),(3,2),(4,1)\}) = \frac{1}{9}, \end{aligned}$$

and

$$P(Z=8) = P(\{(2,6),(3,5),(4,4),(5,3),(6,2)\}) = \frac{5}{36}.$$

Note also that since Z must take on a value between 2 and 12, it follows that

$$\sum_{z=2}^{12} P(Z=z) = 1.$$

This can be verified by going through all of the calculations above from $z=2$ through $z=12$.

Example. Suppose that two fair coins are tossed in a row and let Z be the random variable that is the number of heads that are tossed. Then, we have that

$$\begin{aligned} P(Z=0) &= P((T,T)) = \frac{1}{4}, \\ P(Z=1) &= P(\{(H,T),(T,H)\}) = \frac{1}{2}, \end{aligned}$$

and

$$P(Z=2) = P((H,H)) = \frac{1}{4}.$$

Note that in this case $P(Z=0)+P(Z=1)+P(Z=2)=1$.

Example. Suppose that we are interested in whether a machine will break down within a year from now and that we define the random variable I by

$$I = \begin{cases} 0, \text{ if the machine is still working a year from now,} \\ 1, \text{ if the machine breaks down within a year from now.} \end{cases}$$

The random variables I is referred to as an *indicator function* for whether the machine breaks down or not. This type of random variable will be useful in the course.

If Z is a random variable, then its cumulative distribution function F is defined by

$$F(z) = P(Z \leq z), \ \ z \in \mathbb{R}.$$

The cumulative distribution function encodes all of the information about the distribution of Z. For instance, if one wants to know the probability that Z is greater than some number z, then this is given by $1 - F(z)$. If one wants to know the probability that Z lies between two numbers a and b, then this is given by

$$P(a < Z \leq b) = F(b) - F(a).$$

As we proceed throughout the course, the cumulative distribution function will show up many times.

9 Discrete Random Variables

A random variable Z is said to be *discrete* if it can only take on a finite number of values or at most as many values as there are integers (this is referred to as a *countable* number of values). In this case, we can write down the values that Z can take as $z_1, z_2, ...$ The probability mass function (or pmf for short) of Z is a function p defined on $z_1, z_2, ...$, such that

$$p(z_i) = P(Z = z_i), \ i = 1, 2, ...$$

It must always be the case that

$$\sum_{i=1}^{\infty} p(z_i) = 1.$$

We can also express the CDF of a discrete random variable in terms of its pmf by writing

$$F(z) = \sum_{i: z_i \leq z} p(z_i), \ \ z \in \mathbb{R}.$$

Example. Let Z be the random variable representing the outcome of rolling a 6 sided die. Then, Z is a discrete random variable and the values it can take are 1,2,3,4,5 and 6. The pmf of Z is

$$p(i) = 1/6, \ \ i = 1, 2, 3, 4, 5, 6.$$

The CDF of Z is given by

$$F(z) = \begin{cases} 0, & \text{if } z < 1, \\ 1/6, & \text{if } z < 2, \\ 1/3, & \text{if } z < 3, \\ 1/2, & \text{if } z < 4, \\ 2/3, & \text{if } z < 5, \\ 5/6, & \text{if } z < 6, \\ 1, & \text{if } z \geq 6. \end{cases}$$

Example. Consider a random variable Z with probability mass function

$$p(1) = 1/3, \quad p(2) = 5/12, \quad p(3) = 1/4.$$

The CDF of Z is given by

$$F(z) = \begin{cases} 0, & \text{if } z < 1, \\ 1/3, & \text{if } z < 2, \\ 3/4, & \text{if } z < 3, \\ 1, & \text{if } z \geq 3. \end{cases}$$

10 Continuous Random Variables

A random variable Z is said to be *continuous* if it can take on an infinite number of values that cannot be counted with integers. For instance, a random variable taking all possible values in the interval $[0, 1]$ would be continuous or, more generally, a random variables taking all values in $(-\infty, \infty)$ would be continuous. Because a continuous random variable can take so many values, the probability of it taking any particular one is zero and so there is no probability mass function. Instead, for a continuous random variable we have a probability density function (or pdf for short) which is a function f such that if F is the CDF of Z, then we may write

$$F(z) = \int_{-\infty}^{z} f(u)du, \ \ z \in \mathbb{R}.$$

Another way to think of the pdf f is as the derivative of the CDF F. In other words,

$$f(z) = \frac{dF(z)}{dz}, \ \ z \in \mathbb{R}.$$

Because Z must take on some value with probability 1, it will always be the case

$$\int_{-\infty}^{\infty} f(u)du = 1.$$

To find the probability that Z is *greater* than some number z, we have

$$P(Z > z) = 1 - F(z) = \int_z^\infty f(u)du = 1,$$

and the probability that Z lies between two numbers a and b is

$$P(a < Z < b) = F(b) - F(a) = \int_a^b f(u)du.$$

Example. One example of a continuous random variable is a random variable Z which is uniformly distributed over the interval $[0, 1]$. This random variable will be play a prominent role in our lectures on simulation. The pdf of a Uniform$[0, 1]$ random variable is given by

$$f(z) = \begin{cases} 1, & \text{if } 0 < z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

This implies as expected that a Uniform$[0, 1]$ random variable is equally likely to take any value between 0 and 1. Suppose now that we would like to find the probability Z is less than $1/3$ and the probability that Z is between $1/4$ and $3/4$. Integrating the pdf, we find that the CDF of a Uniform$[0, 1]$ random variable is

$$F(z) = \begin{cases} 0, & \text{if } z < 0, \\ z, & \text{if } 0 \leq z < 1, \\ 1, & \text{if } z \geq 1. \end{cases}$$

The probability that Z is less than $1/3$ is now given by $F(1/3) = 1/3$ and the probability that Z lies between $1/4$ and $3/4$ is given by $F(3/4) - F(1/4) = 1/2$. \

Example. Consider a random variable Z which takes values between 0 and 1 and whose pdf is given by

$$f(z) = \begin{cases} 4z, & \text{if } \quad 0 < z < 1/2, \\ 4 - 4z, & \text{if } 1/2 \leq z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

This is an example of a symmetric triangular distribution (due to the fact that shape of the pdf is a triangle) and it is also the distribution of the average of two independent Uniform$[0, 1]$ random variables. That is, if U_1 and U_2 are independent Uniform$[0, 1]$ random variables, then Z has the same distribution as $(U_1 + U_2)/2$. Suppose now as above that we would like to find the probability Z is less than $1/3$ and the probability that Z is between $1/4$ and $3/4$. Integrating the pdf, we find that the CDF of Z is

$$F(z) = \begin{cases} 0, & \text{if } z < 0, \\ 2z^2, & \text{if } 0 \leq z < 1/2, \\ 2z^2 - (2z - 1)^2, & \text{if } 1/2 \leq z < 1, \\ 1, & \text{if } z \geq 1. \end{cases}$$

The probability that Z is less than $1/3$ is now given by $F(1/3) = 2/9$ and the probability that Z lies between $1/4$ and $3/4$ is given by $F(3/4) - F(1/4) = 3/4$.

Example. Suppose that Z is a random variable with pdf

$$f(z) = \begin{cases} cz^2, & \text{if } -3 < z < 3, \\ 0, & \text{otherwise.} \end{cases}$$

We now wish to find the value of c. In order to do this we can use the fact that

$$\int_{-3}^{3} f(z)dz = 1.$$

Hence, we obtain that

$$c = \left(\int_{-3}^{3} z^2 dz \right)^{-1} = 1/18.$$

11 Expectation of a Random Variable

The expectation of a discrete random variable Z with probability mass function p is given by

$$E[Z] = \sum_{i=1}^{\infty} z_i p(z_i),$$

where $z_1, z_2, ...$, are the set of values that the random variable Z can take. One can think of the expecation as being the average value of Z.\

Example. Let Z be the random variable representing the outcome of rolling a 6 sided die. Then, all 6 values of the die are equally likely and the pmf of Z is given in the exercise above. The expected value of Z is

$$E[Z] = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{7}{2}.$$

Example. Consider a random variable Z with probability mass function

$$p(1) = 1/3, \quad p(2) = 5/12, \quad p(3) = 1/4.$$

The expected value of Z is

$$E[Z] = 1\left(\frac{1}{3}\right) + 2\left(\frac{5}{12}\right) + 3\left(\frac{1}{4}\right) = \frac{23}{12}.$$

Example. Let I by a random variable defined by

$$I = \begin{cases} 0, \text{ if a machine is still working a year from now,} \\ 1, \text{ if a machine breaks down within a year from now.} \end{cases}$$

Recall from the exercise above that I is referred to as an *indicator function* for whether the machine breaks down or not. The expectation of I is given by

$$0 \times P(\text{machine is still working a year from now}) + 1 \times P(\text{machine breaks down within a year from now})$$

which is equal to $P(\text{machine breaks down within a year from now})$. Hence, from this example we see that the expectation of an indicator function is equal to the probability of the event that it is meant to indicate.\

The expectation of a continuous random variable is defined in a similar way to the expectation of a discrete random variable. If Z is a continuous random variable with probability density function f, then its expectation is given by

$$E[Z] = \int_{-\infty}^{\infty} zf(z)dz.$$

Example. Consider a random variable Z which is uniformly distributed over the interval $[0, 1]$. Recall that the pdf of this random variable is given by

$$f(z) = \begin{cases} 1, & \text{if } 0 < z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

The expectation of Z is then

$$E[Z] = \int_{-\infty}^{\infty} zf(z)dz = \int_0^1 zdz = \frac{z^2}{2}\Big|_0^1 = \frac{1}{2}.$$

Example. Consider a random variable Z which takes values between 0 and 1 and whose pdf is given by

$$f(z) = \begin{cases} 4z, & \text{if } \quad 0 < z < 1/2, \\ 4 - 4z, & \text{if } 1/2 \leq z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

This is an example of a symmetric triangular distribution. The expected value of Z is given by

$$E[Z] = \int_{-\infty}^{\infty} zf(x)dz = \int_0^{1/2} 4zdz + \int_{1/2}^1 (4 - 4z)dz = \frac{1}{2}.$$

It turns out that there is an easier way to find out that the expectation of Z is $1/2$. If X_1 and X_2 are two random variables, then the expectation of their sum is equal to the sum of their expectations. In other words,

$$E[X_1 + X_2] = E[X_1] + E[X_2].$$

Also, if c is a constant and X is a random variable, then

$$E[cX] = cE[X].$$

From the exercise above, we know that Z is equal in distribution to $(U_1 + U_2)/2$ where U_1 and U_2 are Uniform$[0, 1]$ random variables. Also, from the exercise above we have that $E[U_1] = E[U_2] = 1/2$. Therefore,

$$E[Z] = E[(U_1 + U_2)/2] = \frac{1}{2}(E[U_1] + E[U_2]) = \frac{1}{2}.$$

Example. Suppose that Z is a random variable with pdf

$$f(z) = \begin{cases} (1/18)z^2, & \text{if } -3 < z < 3, \\ 0, & \text{otherwise.} \end{cases}$$

The expected value of Z is then given by

$$E[Z] = \int_{-\infty}^{\infty} zf(x)dz = \int_{-3}^{3} (1/18)z^3 dz = 0.$$

It should not be surprising that the expected value of Z is 0 because its pdf is symmetric about the origin.

Many times in this course we will be interested in learning the expectation of a function of a random variable. In, particular if g is a function and Z a random variable, we would like to compute $E[g(Z)]$. If Z is a discrete random variable with pmf p, this can be accomplished by setting

$$E[g(Z)] = \sum_{i=1}^{\infty} g(z_i)p(z_i).$$

Example. Let $g(z) = z^2$ and let Z be the random variable representing the outcome of rolling a 6 sided die. Then, all 6 values of the die are equally likely and the pmf of Z is given in the exercise above. The expected value of $g(Z)$ is

$$E[g(Z)] = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) = \frac{91}{6}.$$

Example. Let g be the function defined by

$$g(1) = 4, \quad g(2) = 2, \quad g(3) = 9.$$

and consider a random variable Z with probability mass function

$$p(1) = 1/3, \quad p(2) = 5/12, \quad p(3) = 1/4.$$

The expected value of $g(Z)$ is

$$E[g(Z)] = 4\left(\frac{1}{3}\right) + 2\left(\frac{5}{12}\right) + 9\left(\frac{1}{4}\right) = \frac{53}{12}.$$

If Z is a continuous random variable with pdf f, then $E[g(Z)]$ is given by

$$E[g(Z)] = \int_{-\infty}^{\infty} g(z)f(z)dz.$$

Example. Consider a random variable Z which is uniformly distributed over the interval $[0, 1]$. The pdf of this random variable is given by

$$f(z) = \begin{cases} 1, & \text{if } 0 < z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

If g is a function, the expectation of $g(Z)$ is

$$E[g(Z)] = \int_{-\infty}^{\infty} g(z)f(z)dz = \int_{0}^{1} g(z)dz.$$

Example. Consider a random variable Z which takes values between 0 and 1 and whose pdf is given by

$$f(z) = \begin{cases} 4z, & \text{if } \quad 0 < z < 1/2, \\ 4 - 4z, & \text{if } 1/2 \leq z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

This is an example of a symmetric triangular distribution. Now suppose that $g(z) = z^2$. The expected value of $g(Z)$ is then given by

$$E[g(Z)] = \int_{-\infty}^{\infty} g(z)f(z)dz = \int_{0}^{1/2} 4z^3 dz + \int_{1/2}^{1} z^2(4 - 4z)dz = \frac{7}{24}.$$

Example. Suppose that Z is a random variable with pdf

$$f(z) = \begin{cases} (1/18)z^2, & \text{if } -3 < z < 3, \\ 0, & \text{otherwise,} \end{cases}$$

and let $g(z) = 2 - z$. The expected value of $g(Z)$ is then given by

$$E[g(Z)] = \int_{-\infty}^{\infty} g(z)f(z)dz = \int_{-3}^{3} 2(1/18)z^3dz - \int_{-3}^{3}(1/18)z^3dz = 2\int_{-3}^{3}(1/18)z^3dz = 2.$$

12 Variance of a Random Variable

If Z is a random variable, then its variance is defined by

$$\text{Var}(Z) = E[(Z - E[Z])^2].$$

Another way to express variance which follows from this definition is

$$\text{Var}(Z) = E[Z^2] - E^2[Z].$$

Depending on the situation it may be easier to compute one or the other of these two expressions. The importance of knowing the variance of a random variable is that it provides an indication of how spread out the random variable is around its mean. Because the variance is expressed entirely in terms of some function g of the random variable Z, we can use the results from the previous section to calculate it.

Example. Let Z be the random variable representing the outcome of rolling a 6 sided die. Then, from the examples above know that $E[Z] = 7/2$ and $E[Z^2] = 91/6$. The variance of Z is therefore

$$\text{Var}(Z) = E[Z^2] - E^2[Z] = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$

Example. Let Z be the random variable with probability mass function

$$p(1) = 1/3, \quad p(2) = 5/12, \quad p(3) = 1/4.$$

From the above exercise, we have that the expected value of Z is $23/12$. Therefore, the variance of Z is

$$\text{Var}(Z) = E[(Z - (23/12))^2] = \left(\frac{11}{12}\right)^2\frac{1}{3} + \left(\frac{1}{12}\right)^2\frac{5}{12} + \left(\frac{13}{12}\right)^2\frac{1}{4} = \frac{83}{144}.$$

Example. Let I by a random variable defined by

$$I = \begin{cases} 0, \text{ if a machine is still working a year from now,} \\ 1, \text{ if a machine breaks down within a year from now.} \end{cases}$$

In this case, since $0^2 = 0$ and $1^2 = 1$, the variance of I is equal to

$$E[I] - E^2[I] = E[I](1 - E[I])$$

which, using the result above is equal to

$$P(\text{machine breaks down within a year from now}) \times (1 - P(\text{machine breaks down within a year from now})).$$

Example. Consider a random variable Z which is uniformly distributed over the interval $[0, 1]$. The pdf of this random variable is given by

$$f(z) = \begin{cases} 1, & \text{if } 0 < z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

We know from the above that the expected value of Z is $1/2$. Hence, the variance of Z is

$$\text{Var}(Z) = E[(Z - (1/2))^2] = \int_0^1 (z - (1/2))^2 dz = \frac{1}{12}.$$

Example. Consider a random variable Z which takes values between 0 and 1 and whose pdf is given by

$$f(z) = \begin{cases} 4z, & \text{if } \quad 0 < z < 1/2, \\ 4 - 4z, & \text{if } 1/2 \leq z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Using the examples above, we have that $E[Z] = 1/2$ and $E[Z^2] = 7/24$. Hence,

$$\text{Var}(Z) = E[Z^2] - E^2[Z] = \frac{7}{24} - \left(\frac{1}{2}\right)^2 \; = \; \frac{1}{24}.$$

Example. Suppose that Z is a random variable with pdf

$$f(z) = \begin{cases} (1/18)z^2, & \text{if } -3 < z < 3, \\ 0, & \text{otherwise.} \end{cases}$$

From the exercise above, we have that $E[Z] = 0$. Hence,

$$\text{Var}(Z) = E[Z^2] = \frac{1}{18} \int_{-3}^{3} z^4 \; = \; \frac{27}{5}.$$

13 Joint Distributions

So far we have discussed random variables in isolation. We now extend our results to the case of two or more random variables, also known as a random vector. If X and Y are two random variables, then their joint distribution function is given by

$$F(x,y) = P(X \leq x, Y \leq y), \quad x,y \in \mathbb{R}.$$

The joint distribution function encodes all of the information about the random vector (X,Y). For instance, from the joint distribution function we can obtain the distribution function of X,

$$F_X(x) = P(X \leq x) = F(x,\infty),\ x \in \mathbb{R},$$

or the distribution function of Y,

$$F_Y(y) = P(Y \leq y) = F(\infty,y),\ y \in \mathbb{R}.$$

These distribution functions are commonly referred to as the marginal distributions of the random vector (X,Y).

In the specific case where X and Y are both discrete random variables, we can define their joint probability mass function by

$$p(x,y) = P(X = x, Y = y),\ x,y \in \mathbb{R}.$$

The probability mass function of X is then given by

$$p_X(x) = \sum_{y:p(x,y)>0} p(x,y),\ x \in \mathbb{R},$$

and the probability mass function of Y is given by

$$p_Y(y) = \sum_{x:p(x,y)>0} p(x,y),\ y \in \mathbb{R}.$$

If X and Y are continuous random variable, then they are said to be jointly continuous if there exists a function $f(x,y)$ such that we may write

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(x,y)dxdy,\ x,y \in \mathbb{R}.$$

The function $f(x,y)$ is referred to as the joint probability density function of X and Y. The cumulative distribution function of X is then given by

$$F_X(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f(x,y)dxdy,\ x \in \mathbb{R},$$

and the cumulative distribution function of Y is given by

$$F_X(x) = \int_{-\infty}^{y}\int_{-\infty}^{\infty} f(x,y)dxdy,\ y \in \mathbb{R}.$$

Also, the joint density function f may be recovered from the joint distribution function F via the relationship

$$f(x,y) = \frac{d^2}{dxdy}F(x,y), \;\; x,y \in \mathbb{R}.$$

14 Independence

The random variables X and Y are said to be independent if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y), \; x,y \in \mathbb{R}.$$

This can be expressed in terms of the joint distribution function of X and Y by writing $F(x,y) = F(x)F(y)$. If X and Y are independent discrete random variable, then in terms of probability mass functions, we have that

$$p(x,y) = p_X(x)p_Y(y) \; x,y, \in \mathbb{R},$$

whereas if X and Y are jointly continuous random variables, then in terms of probability density functions, we have that

$$f(x,y) = f_X(x)f_Y(y) \; x,y, \in \mathbb{R}.$$

One important fact regarding independent random variables is that if X and Y are independent, then $E[XY] = E[X]E[Y]$. More generally, if g and h are functions, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Example. Suppose that X is uniformly distributed on $[0,1]$ and that Y is is a random variable with a mean of 3. Also suppose that X and Y are independent of each other. Calculate $E[XY]$. In order to calculate $E[XY]$ we use the fact that the mean of a Uniform$[0,1]$ random variable is $1/2$ together with the fact that by independence, $E[XY] = E[X]E[Y]$, in order to find that $E[XY] = 3/2$.

15 Covariance

If X and Y are random variables with joint distribution function F, then their covariance is defined to be

$$\text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])].$$

From this definition, it may also be shown that

$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y],$$

which sometimes makes $\text{Cov}(X, Y)$ easier to compute. If X and Y are independent, then $\text{Cov}(X, Y) = 0$. Also, if $X = Y$, then $\text{Cov}(X, Y)$ reduces to $\text{Var}(X)$ or $\text{Var}(Y)$.

Example. Suppose that X is the indicator function of the event A and that Y is the indicator function of the event B. Recall that this means that

$$X = \begin{cases} 0, \text{ if event A does not occur,} \\ 1, \text{ if event A does occur,} \end{cases}$$

and

$$Y = \begin{cases} 0, \text{ if event B does not occur,} \\ 1, \text{ if event B does occur.} \end{cases}$$

Then, we may calculate that

$$\text{Cov}(X, Y) = P(X = 1, Y = 1) - P(X = 1)P(Y - 1).$$

In this case it may be shown that $\text{Cov}(X, Y) > 0$ if and only if

$$P(Y = 1|X = 1) > P(Y = 1).$$

In other words, the covariance between X and Y is positive if and only if Y is more likely to be 1 given that X equal to 1. More generally, if the two random variables X and Y have a positive covariance, then X tends to be large when Y is large, whereas if X and Y have a negative covariance, then X tends to be small when Y is large.

Another measure of dependence between two random variables X and Y which will be used in the course is their correlation, denoted by $\rho_{X,Y}$. The correlation between X and Y is defined by

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

[ ]:

16 Session 3: Random Number Generation

Summer 2019 - Instructor: Josh Reed

Teaching Assistant: Haotian Song

16.1 The Inverse Transform Method

Let X be a random variable with cumulative distribution function F. Recall that this means that $F(x) = P(X \leq x)$ for each $x \in \mathbb{R}$. The quantile function associated with F is defined by

$$F^{-1}(u) \quad = \quad \min\{x \in \mathbb{R} : u \leq F(x)\}, 0 < u < 1.$$

The inverse transform method of generating a random variable X with CDF F is the following. First generate a Uniform$[0,1]$ random variable U, and then set

$$X = F^{-1}(U).$$

To verify that X has the proper distribution, note that

$$P(X \leq x) \quad = \quad P(F^{-1}(U) \leq x) \; = \; P(F(F^{-1}(U)) \leq F(x)) \; = \; P(U \leq F(x)) \; = \; F(x).$$

The following is pseudocode for the inverse transform method.

1. Generate a random variable U which has a Uniform$[0,1]$ distribution.

2. Return $X = F^{-1}(U)$.

Example. Consider a random variable X which takes the value 0 with probability $1 - p$, and the value 1 with probability p. This is referred to as the *Bernoulli distribution* and it is an example of a *discrete distribution*. The CDF of X is given by

$$F(x) \quad = \quad \begin{cases} 0 \text{ if } x < 0, \\ 1 - p \text{ if } 0 \leq x < 1, \\ 1 \text{ if } x \geq 1. \end{cases}$$

Its quantile function is therefore

$$F^{-1}(u) \quad = \quad \begin{cases} 0 \text{ if } 0 < u \leq 1 - p, \\ 1 \text{ if } 1 - p < u \leq 1. \end{cases}$$

Hence, according to the inverse transform method, if we generate a Uniform$[0,1]$ random variable, we then set $X = 0$ if $0 < U \leq 1 - p$ or $X = 1$ if $1 - p < U \leq 1$.

Example. The *Poisson distribution* is a second example of a discrete distribution. It is commonly used to count the number of events occuring over some interval of time. The Poisson distribution is paramterized by its mean λ. Specifically, if N is a Poisson random variable with mean λ, then

$P(N = n) = e^{-\lambda}\lambda^n/n!$ for $n = 0, 1, 2, ...$ This implies that the CDF of N is given by $F(x) = 0$ for $x < 0$ and

$$F(x) = \sum_{k=0}^{\lfloor x \rfloor} e^{-\lambda}\frac{\lambda^k}{k!}, \ x \geq 0,$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. It is not possible to provide a simple expression for the sum appearing above and so we cannot write the quantile function of F in a nice way. Nevertheless, we can still apply the inverse transform method as follows. First, let U be a uniform random variable on the interval $[0, 1]$. Then, set $X = n$ where n is such that $F(n-1) < U < F(n)$. There is no formula for finding the n satisfying the preceding condition but one may iteratively search through n starting from $n = 0$ until the proper n is found.

Example. The *geometric distribution* is an example of a discrete distribution which represents the number of repeated, independent trials until success. For example, consider the number of times a coin must be flipped before it lands on heads. The probability of success on any given trial is denoted by $0 \leq p \leq 1$. If X is a geometric random variable with probability of success p, then the probability of n trials until success is given by $P(X = n) = (1-p)^{n-1}p$ for $n = 1, 2, ...$ A geometric random variable may be simulated using the inverse transform method by first generating a Uniform$[0, 1]$ random variable U and then setting $X = 1 + \lfloor \ln(U)/\ln(1-p) \rfloor$.

Example. For a final example of the inverse transform method, consider the exponential distribution. This is an example of a continuous distribution. Like the Poisson distribution, the exponential distribution is parameterized by its mean, which is given by $1/\lambda$. The parameter λ is referred to as the *rate* of the distribution. The CDF of an exponential distribution with rate λ is given by

$$F(x) = 1 - e^{-\lambda x}, \ x \geq 0.$$

Inverting this CDF, one obtain its quantile function $F^{-1}(q) = -\ln(1-u)/\lambda$ for $0 < u < 1$. Note also that if U is a Uniform$[0, 1]$ random variable, then $1 - U$ is a Uniform$[0, 1]$ random variable too. Hence, setting $X = -\ln(U)/\lambda$ one obtains a exponential random variable with mean $1/\lambda$.

16.1.1 Conditional Distributions

Sometimes it is desirable to generate a random variable X conditioned on it taking values in some interval $[a, b]$. The inverse transform method can be used in this situation. Suppose that X has CDF F. Then, the CDF of F *conditioned* on X being in the interval $[a, b]$ is given by

$$F_{a,b}(x) = \frac{F(x) - F(a)}{F(b) - F(a)}, \ a \leq x \leq b.$$

So, letting U be uniformly distributed between $F(a)$ and $F(b)$ and set $X = F^{-1}(U)$, one can verify that X has the CDF given by $F_{a,b}$ above.

In-Class Exercise. Use the inverse transform method to generate a random variable X having the cumulative distribution function

$$F(x) = 1 - e^{-\sqrt{x}}, \;\; x > 0.$$

Create a histrogram of this distribution function based off of 100,000 simulation replicatons.

Solution. This is an example of a Weibull distribution with scale parameter 1 and shape parameter $1/2$. Inverting the cumulative distribution function F, we find that

$$F^{-1}(q) = (\ln(1-q))^2, \;\; 0 < q < 1.$$

To use the inverse transform method to generate a random variable X having the cumulative distribution function F we may therefore set $X = (\ln(U))^2$ where U is a Uniform$[0,1]$ random variable. Here we have used the fact that $1 - U$ has the same distribution at U.

17 The Acceptance-Rejection Method

The acceptance-rejection method for generating random variables first generates a random variable according to some alternative distribution G and then decides whether to accept or reject the sample. The distribution of the accepted samples turns out to have the desired distribution F. The origin of this method dates back to the work of the mathematician John von Neumann. Its specifics are as follows.

17.1 Continuous Distributions

A random variable X with a CDF F is said to be a *continuous distribution* if we may write

$$F(x) = \int_{-\infty}^{x} f(u)du, \;\; x \in \mathbb{R}.$$

In this case f is referred to as the *probability density function* of f (or *pdf* for short) and it is also the derivative of F. Suppose now that we wish to generate a random variable X from a distribution with CDF F and density f. To implement the acceptrance-rejection method we first pick an alternative continuous distribution which has CDF G and density g and also select a $c > 0$ such that $f(y) \leq cg(y)$ for all $y \in \mathbb{R}$. We then generate a random variable Y according to the distribution G and set $X = Y$ with probability $f(y)/(cg(y))$. This is the acceptance-rejection portion of the algorithm. If X is not set equal to Y, we generate a new sample Y from the distribution G and perform the acceptance-rejection test again. We continue on in this manner until eventually we have accepted some Y and set $X = Y$.

The following is pseudocode for the continuous form of the acceptance-rejection method.

1. Generate a random variable Y according to the distribution G.

2. Generate a random variable U which has a Uniform$[0,1]$ distribution.

3. If $U \leq f(Y)/(cg(Y))$, return Y. Otherwise, return to Step 1.

The number of samples from the alternative distribution G that the acceptance-rejection method must generate before an acceptance occurs is a geometric random variable with probability of success $1/c$. Hence, on average the algorithm will require c samples from the alternative distribution G. This motivates one to choose an alternative distribution such that c may be set close to 1.

Example. Suppose that the positive random variable X has a pdf given by $f(x) = xe^{-x}$ for $x \geq 0$. In this case, the alternative pdf g can be chosen to be the exponential distribution with rate $\lambda = 1/2$. In other words, $g(x) = (1/2)e^{-x/2}$. The constant c may be chosen to be $4/e$ which is about 1.472.

Example. Suppose that X is an arbitrary random variable which takes values in the unit interval $[0, 1]$ and which has a pdf f. In this case, one alternative distribution G which we may pick is the Uniform$[0, 1]$ distribution. Since the pdf g of a Uniform$[0, 1]$ is constant and equal to 1, in order to implement the acceptance-rejection method we must select a value of c which is less than $f(x)$ for all $x \in [0, 1]$. One option is to set c equal to the maximum of f over $[0, 1]$. The specifics of the acceptance-rejection method are then as follows. Generate two Uniform$[0, 1]$ random variables U_1 and U_2. If $U\_2 \leq f(U\_1)/c$, return U_1, otherwise repeat.

17.2 Discrete Distributions

There is also a version of the acceptance-rejection method for discrete random variables. Suppose that X is a random variable following a *discrete distribution*. In other words, there exists a set of x_j for $j = 1, 2, ...$, such that $P(X = x_j) = p_j$ and $p_1 + p_2 + ... = 1$. In this case, when performing the acceptance-rejection method we find an alternative discrete random variable Y with some distribution G such that $P(Y = x_j) = q_j$ and $q_1 + q_2 + ... = 1$. Moreover, it must be the case that there exists a $c \geq 0$ satisfying $p_j/q_j \leq c$ for each $j = 1, 2, ...$ We then proceed similar to the case of the acceptance-rejection method for continuous distributions. The only difference is that rather than work with probability density functions, we work with *probability mass functions* instead.

Pseudocode for the discrete form of the acceptance-rejection method is given below.

1. Generate a discrete random variable Y according to the distribution G. Call the value that this random variable takes y_j.

2. Generate a random variable U which has a Uniform$[0, 1]$ distribution.

3. If $U \leq P(X = y\_j)/(cP(Y = y\_j))$, return y_j. Otherwise, return to Step 1.

Example. Suppose that X is a discrete random variable which takes the integer values 1 through 10 with the following probabilities.

xj	1	2	3	4	5	6	7	8	9	10
P(X=xj)	0.05	0.08	0.15	0.28	0.17	0.09	0.07	0.04	0.04	0.03

Next, suppose that Y is the unfiform distribution on the integers 1 through 10, meaning that $P(Y = i) = 1/10$ for $i = 1, ..., 10$. In this case, the maximum value of p_j/q_j is given by $0.28/0.10 = 2.8$ and so we may set $c = 2.8$.

Example. Suppose that X is a Poisson random variable with mean 2. In this case, $P(X = j) = e^{-2}2^j/j!$ for $j = 0, 1, 2, ...$ The first 7 values of this formula are as follows.

j	0	1	2	3	4	5	6
P(X=j)	0.14	0.27	0.27	0.18	0.09	0.04	0.01

A natural candidate for the distribution of Y is the geometric distribution with a matching mean of 2. This corresponds to a probability of success of $p = 0.5$. The first 7 values of the probability mass function of Y are as follows.

j	0	1	2	3	4	5	6
P(Y=j)	0.50	0.25	0.125	0.06	0.035	0.02	0.01

A graph of the ratio $P(X = j)/P(Y = j)$ is given in the chart below.The maximum value of this ratio is approximately 2.89, which c may be set equal to.

In-Class Exercise. Use the acceptance-rejection method for continuous distributions to simulate a random variable X having probability density function $f(x) = 6x(1 - x)$ on the interval $[0, 1]$. This is a so-called Beta(2,2) distribution. Create a histrogram of the distribution of X based off of 100,000 simulation replicatons.

Solution. In this case since X is defined on the interval $[0, 1]$, we will choose the alternative distribution G to be a Uniform$[0, 1]$ random variable. The maximum value of $f(x)$ on $[0, 1]$ is $3/2$ and the probability density function of a Uniform$[0, 1]$ is $g(y) = 1$. We may therefore set $c = 3/2$.

18 Simulating Normal Random Variables

The family of normal random variables is paramterized by a mean μ and variance σ^2. A normal random variable with a mean of 0 and a variance of 1 is referred to as a *standard normal random variable*. A standard normal random variable has a probability density function given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2}, x \in \mathbb{R}.$$

If X is a standard normal random variable, then

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

is a $N(\mu, \sigma^2)$ random variable.

The inverse transform method is difficult to implement for normal random variables because the CDF of a normal random variable cannot be expressed in a simple form and so the quantile function cannot be written explicitly. The acceptance-rejection method is a possibility if one chooses an appropriate alternative density g. In this section we present two alternative methods specifically designed to simulate normal random variables.

18.1 The Box-Muller method

The *Box-Muller method* generates two independent $N(0,1)$ random variables. One of these can then be used via the relationship above to generate a $N(\mu, \sigma^2)$ random variable. The method works as follows. First, generate two independent Uniform$[0,1]$ random variables, labeled U_1 and U_2. Then, set $R = -2\ln(U_1)$ and $V = 2\pi U_2$. Finally, the two independent $N(0,1)$ random variables are given by $Z_1 = \sqrt{R}\cos(V)$ and $Z_2 = \sqrt{R}\sin(V)$. Some pseudocode for the Box-Muller method is given below.

1. Generate 2 independent random variables U_1 and U_2 both of which have a Uniform$[0,1]$ distribution.

2. Set $R = -2\ln(U_1)$ and $V = 2\pi U_2$.

3. Return $Z_1 = \sqrt{R}\cos(V)$ and $Z_2 = \sqrt{R}\sin(V)$.

18.2 The Polar Form of the Box-Muller method

A variant of the Box-Muller method is the *polar form* of the Box-Muller method. Both methods generate a pair of $N(0,1)$ random variables, but the polar method is slightly faster since it avoids having to calculate sines and cosines. The method works as follows. First generate a pair U_1 and U_2 of independent Uniform$[-1,1]$ random variables. Next, set $s = U_1^2 + U_2^2$. If $s > 1$, discard U_1 and U_2 and repeat by generating a new pair of Uniform$[-1,1]$ random variables. Otherwise, return

$$Z_1 = (U_1/\sqrt{s})\sqrt{-2\ln(s)} \quad \text{and} \quad Z_2 = (U_2/\sqrt{s})\sqrt{-2\ln(s)}.$$

The pair Z_1 and Z_2 will be independent $N(0,1)$ random variables. Some pseudocode for the polar form of the Box-Muller method is given below.

1. Generate 2 independent random variables U_1 and U_2 both of which have a Uniform$[0,1]$ distribution.

2. Set $s = U_1^2 + U_2^2$.

3. If $s < 1$, return $Z_1 = (U_1/\sqrt{s})\sqrt{-2\ln(s)}$ and $Z_2 = (U_1/\sqrt{s})\sqrt{-2\ln(s)}$. Otherwise, return to Step 1.

18.3 Normal Random Vectors

A d-dimensional random vector $X = (X_1, ..., X_d)$ is said to be a *standard normal random vector* if each of the X_1, X_2 through X_d are independent standard normal random variables. A d-dimensional random vector $X = (X_1, ..., X_d)$ is said to be a *normal random vector* if it may be expressed as $X = \mu + AZ$, where $\mu = (\mu_1, ..., \mu_d)$ is a $d-$dimensional vector, A is a $d \times d$ matrix and Z is a d-dimensional standard normal random vector. Recall that in general the covariance between X_i and X_j for $1 \leq i, j \leq d$ is given by

$$\text{Cov}(X_i, X_j) \quad = \quad E[(X_i - \mu_i)(X_j - \mu_j)].$$

The $d \times d$ matrix $\Sigma = AA^T$ is referred to as the *covariance matrix* of X. This is because the (i,j)-th entry of Σ is given by $\text{Cov}(X_i, X_j)$. The vector $\mu = (\mu_1, ..., \mu_d)$ is referred to as the *mean vector* of X.

In order to simulate a d-dimensional standard normal random vector Z, one can first simulate d independent standard normal random variables using either the Box-Muller method or its polar form given above, and then put them together into a vector. Thus, to simulate a normal random vector one can use the relationship $X = \mu + AZ$, where Z is a standard normal vector. In many cases a normal random vector is parameterized by its mean vector and covariance matrix, denoted by $N(\mu, \Sigma)$. Hence, in order to simulate using the relationship $X = \mu + AZ$, a matrix A satisfying $AA^T = \Sigma$ must be found. We present two ways of doing this.

18.3.1 Cholesky Factorization

The *Cholesky factorization* of Σ is a lower triangular matrix A such that $AA^T = \Sigma$. For simulation purposes this is useful because it reduces the number of computations required in order to compute AZ. Most programming languages contain a Cholseky factorization function. The following pseudocode simulates a normal random vector $N(\mu, \Sigma)$ using the Cholesky factorization of its covariance matrix.

1. Compute the Cholesky factorzation $AA^T = \Sigma$ of the covariance matrix Σ.

2. Generate a standard normal random vector Z.

3. Set $X = \mu + AZ$.

Example. Suppose that X is a normal random vector with mean vector $\mu = 0$ and covariance matrix

$$\Sigma = \begin{pmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{pmatrix}.$$

Then, the Cholseky factorization of Σ is given by $\Sigma = AA^T$, where

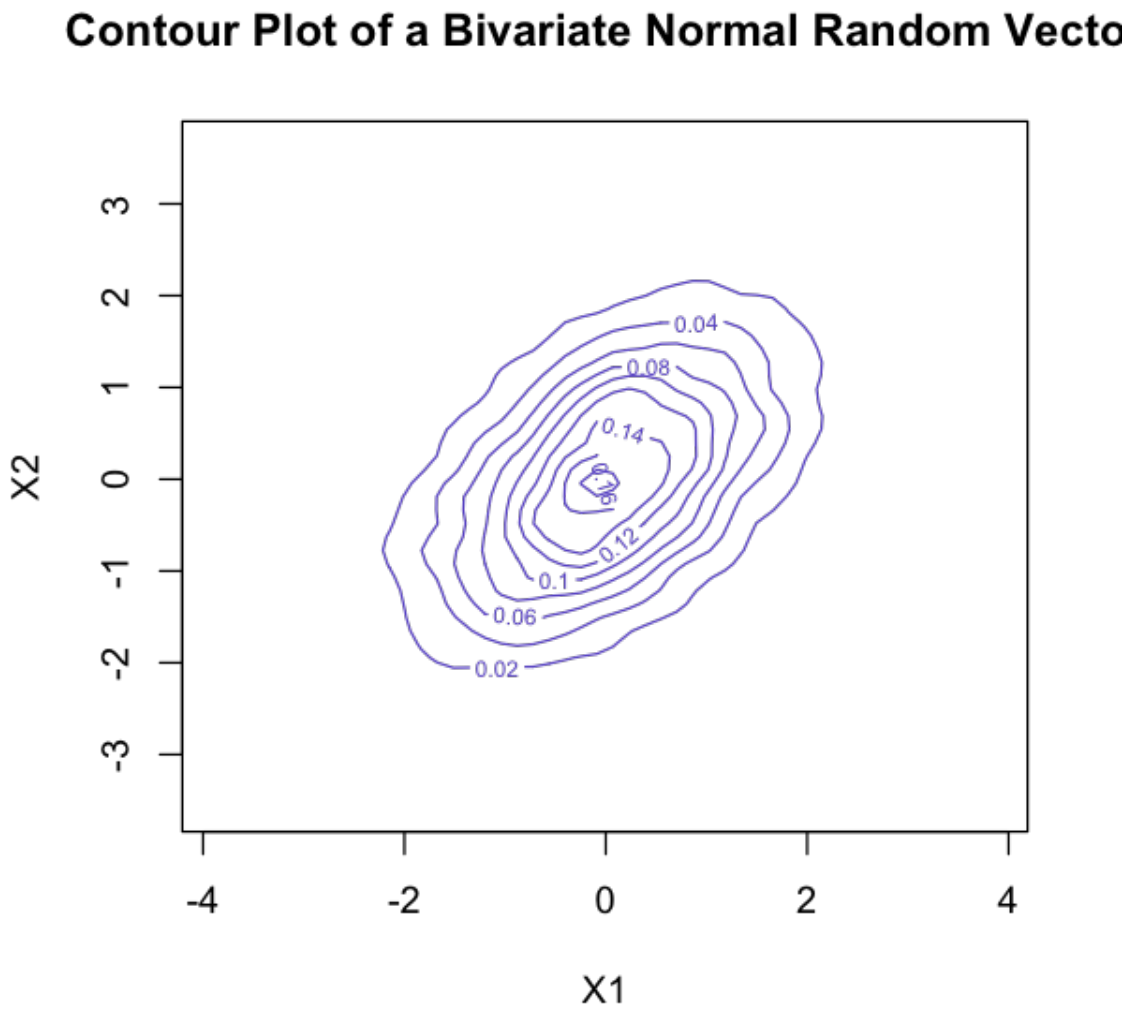
$$A = \begin{pmatrix} 5 & 0 & 0 \\ 3 & 3 & 0 \\ -1 & 1 & 3 \end{pmatrix}.$$

In order to simulate X we write $X = AZ$, where Z is 3-dimensional standard normal vector. Writing this out in long form, one obtains that

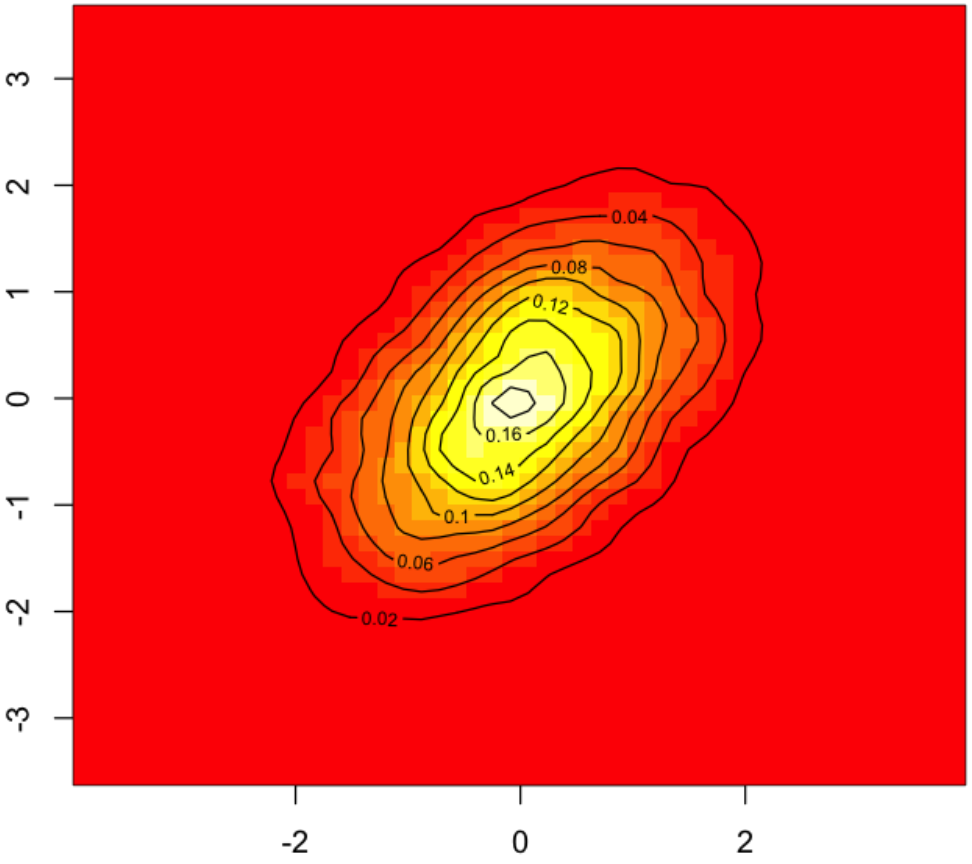
$$\begin{aligned} X_1 &= 5Z_1 \\ X_2 &= 3Z_1 + 3Z_2 \\ X_3 &= -Z_1 + Z_2 + 3Z_3. \end{aligned}$$

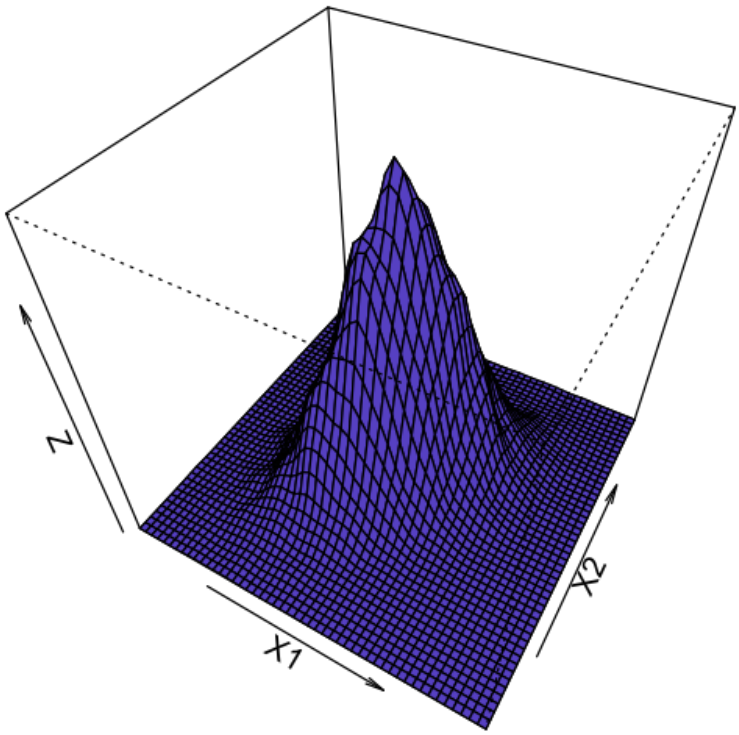
Example. Consider a bivariate normal random vector with a mean vector μ equal to 0 and a covariance matrix given by

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$



The figure below is a fancier version of this plot.


3
2
1
0
-1
-2
-3
-2
0
2
0.04
0.08
0.12
0.16
0.14
0.1
0.06
0.02



Example. Recall that if X_1 and X_2 are random variables, then their *correlation* is given by

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}},$$

where σ_{X_1} and σ_{X_2} are the standard deviations of X_1 and X_2, respectively. The Cholesky decomposition of the covariance matrix of $X = (X_1, X_2)$ is given by $\Sigma = AA^T$, where

$$A = \begin{pmatrix} \sigma_{X_1} & 0 \\ \rho\sigma_{X_2} & \sqrt{1-\rho^2}\sigma_{X_2} \end{pmatrix}.$$

In-Class Exercise. Consider a bivariate normal random vector with a mean vector of $\mu = (1, 1)$ and a covariance matrix

$$\Sigma = \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}.$$

Using the Cholesky factorization approach, generate 100,000 simulations of this random vector and graph the output as a scatter plot.

18.3.2 The Principle Components Method

The principle components method decomposes a normal random vector X into its different sources of variation. This is helpful for instance if the vector $X = (X_1, ..., X_d)$ represents the random stock prices of d different companies. Each of the companies may be exposed to certain macroeconomic factors representing the economy at large (such as interest rates, unemployment rates, etc.) and these factors may represent the largest variations in X. Next, there may be industry specific factors such as the prices of certain commodities which affect some but not all of the companies, and finally there is most likely some idiosyncratic noise that exists at the firm level. Principal component analysis identifies and ranks these sources of variation from largest to smallest.

Implementing principal components analysis requires a technique referred to as *diagonalizing* the covariance matrix Σ. When diagonalizing Σ, one first creates a matrix V whose columns are the unit eigenvectors of Σ and a diagonal matrix Λ whose diagonal elements are the associated eigenvalues of Σ. It then follows that $\Sigma = V\Lambda V^T$ and so one may take $A = V(\Lambda)^{1/2}$. Most programming languages provide functions to diagonalize a matrix Σ. The following pseudocode simulates a normal random vector $N(\mu, \Sigma)$ by diagonalizing its covariance matrix.

1. Diagonalize the covariance matrix Σ by solving $V\Lambda V^T = \Sigma$.

2. Generate a standard normal random vector Z.

3. Set $X = \mu + AZ$ where $A = V(\Lambda)^{1/2}$.

Constructing A by diagonalizing Σ does not reduce the amount of time required to simulate X. However, it can be used to construct optimal (in a sense described below) approximations to X. The idea is as follows. Suppose that X will be simulated by setting $X = \mu + AZ$, where Z is a standard normal random variable and A is obtained diagonalizing Σ as above. That is, $A = V(\Lambda)^{1/2}$. Also assume that the diagonal elements of Λ are ordered from largest to smallest. Now suppose that due to processing constraints the entire matrix computation AZ cannot be performed. Then, one option is to use the *first* $1 \leq k < d$ components of Z when calculating AZ. In other words, approximate X by $a_1Z_1 + ... + a_kZ_k$ where a_j is the jth column of A. This approximation provides the optimal k dimensional approximation to X which minimizes the mean-squared error.

Example. Suppose again that X is a normal random vector with mean vector $\mu = 0$ and covariance matrix

$$\Sigma = \begin{pmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{pmatrix}.$$

Then, the diagonalization of Σ (up to two decimals) is given by $\Sigma = V\Lambda V^T$, where

$$V = \begin{pmatrix} 0.78 & -0.18 & 0.60 \\ 0.60 & 0.45 & -0.66 \\ -0.15 & 0.88 & 0.46 \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} 37.50 & 0 & 0 \\ 0 & 12.02 & 0 \\ 0 & 0 & 4.50 \end{pmatrix}.$$

Hence, the matrix $A = V(\Lambda)^{1/2}$ is given (up to two decimals) by

$$A = \begin{pmatrix} 4.80 & -0.62 & 1.26 \\ 3.69 & 1.55 & -1.40 \\ -0.91 & 3.04 & 0.97 \end{pmatrix}.$$

18.4 The Composition Method

The composition method works as follows. Suppose that the distribution function of a random variable X is given by

$$F(x) = p_1F_1(x) + p_2F_2(x) + ... + p_KF_K(x),\ x \in \mathbb{R},$$

where $p_1 + p_2 + ... + p_K = 1$ and each F_k is the distribution function of some random variable X_k. This for instance would be the distribution of a random variable generated by first rolling a k-sided die to determine the distribution of the random variable and then generating a random variable from that distribution.

The composition method follows from the description of X. First, generate a random variable Y which is equal to k with probability p_k for $k = 1, 2, ..., K$. Next, suppose that $Y = y$. Then, generate a random variable X_y which has distribution F_y. In order for the method to be useful, it should be easy to generate a random variable from any of the distributions F_1 through F_K.

If each F_k in the above is an exponential distribution with rate λ_k, then the distribution of F is referred to as a *hyperexponential distribution*. In this case, since an exponential distribution is easy to simulate, the composition approach is straightforward to implement.

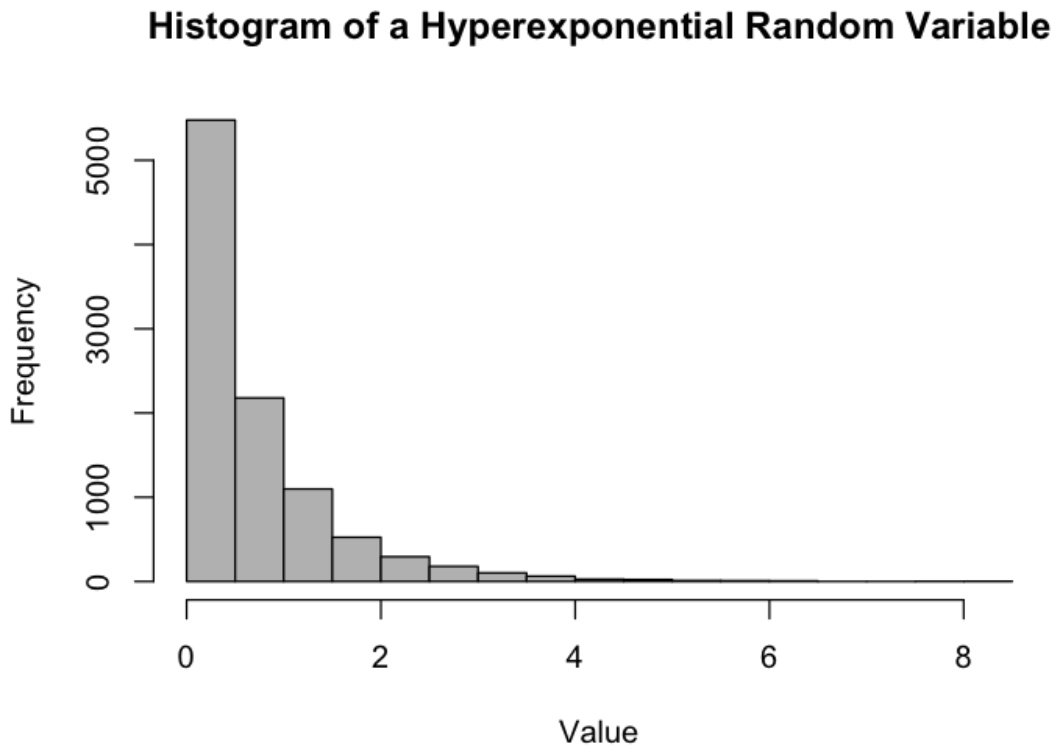
In-Class Exercise. Consider a hyperexonential distribution with

$$p_1 = 1/2,\ \ p_2 = 1/4 \text{ and } \ p_3 = 1/4,$$

and

$$\lambda_1 = 1,\ \ \lambda_2 = 2 \text{ and } \ \lambda_3 = 3,$$

Using the composition approach, generate 100,000 simulations of this random variable and graph the output as a histogram.



18.5 The Convolution Approach

Some random variables are constructed as the summation of two simpler random variables. For instance, if X_1 represents the time at which customer 1 arrives to a bank, and X_2 represents the amount of time *between* when customers 1 and 2 arrive to the bank, then $T = X_1 + X_2$ is the time at which customer 2 arrives to the bank. The distribution of T is referred to as the *convolution* of the distributions of X_1 and X_2. In general, the distribution of the convolution is hard to compute, which makes it difficult to apply the techniques we have learned so far in order to simulate it. However, if one can simulate directly from the distributions of X_1 and X_2, then T may be simulated by adding X_1 and X_2 together. This method of simulation is referred to as the *convolution approach*.

Example. Suppose that X_1 and X_2 in the bank example above are exponential random variables with rate λ. Then, $T = X_1 + X_2$ has an *Erlang distribution* with *shape parameter* 2 and *scale parameter* λ. More generally, an *Erlang distribution with shape parameter n and scale parameter* λ may be written as the sum of n independent and identically distributed exponential random variables with rate parameter λ.

Example. A *lognormal* random variable X may be written as $X = \exp(Y)$, where Y is a normally distributed random variable with mean μ and variance σ^2. We use the notation $LN(\mu, \sigma^2)$ to denote this random variable. Lognormal random variables are commonly used to model stock prices. Using the convolution approach one can simulate a lognormal random variable by fist simulating the normal random variable Y and then exponentiating to obtain X.

Example. A *chi-square distribution with 1 degree of freedom* may be written as $X = Z^2$, where Z is a standard normal random variable. Hence, in order to simulate a chi-square random variable with 1 degree of freedom one can first simulate a standard normal random variable and then square it. A *chi-square distribution with n degrees of freedom* is the sum of n independent chi-square random variables each with 1 degree of freedom. This may also be simulated using the convolution approach.

Example. Student's* *t-distribution with ν degrees of freedom* may be written as the random variable $X = Z/\sqrt{V/\nu}$, where Z is a standard normal random variable and V is an independent chi-square distribution with ν degrees of freedom.

19 Session 4: Monte Carlo Simulation

Summer 2019 - Instructor: Josh Reed

Teaching Assistant: Haotian Song

In this session, Monte Carlo simulation is introduced. Also discussed are output analysis and run legnth control.

20 Monte Carlo Overview

Monte Carlo simulation is a statistical technique which provides estimates of some unknown quantity of importance. Although it can seemingly take many different forms the underyling idea of Monte Carlo is nearly always the same. There exists some unknown quantity μ which we would like to estimate. In order to do so a sequence of independent and identically distributed random variables $Z_1, Z_2, ..., Z_N$ with mean equal to μ is simulated on a computer. Their sample mean

$$\bar{Z} = \frac{Z_1 + Z_2 + ... + Z_N}{N}$$

is then taken as our *estimate of* μ.

The quantity μ to be estimated can take many different forms. It can be the price of a stock option, the probability that a company goes bankrupt or even the value of some complicated integral. However, the underlying Monte Carlo framework will not change. This is a useful point to remember as we move forward.

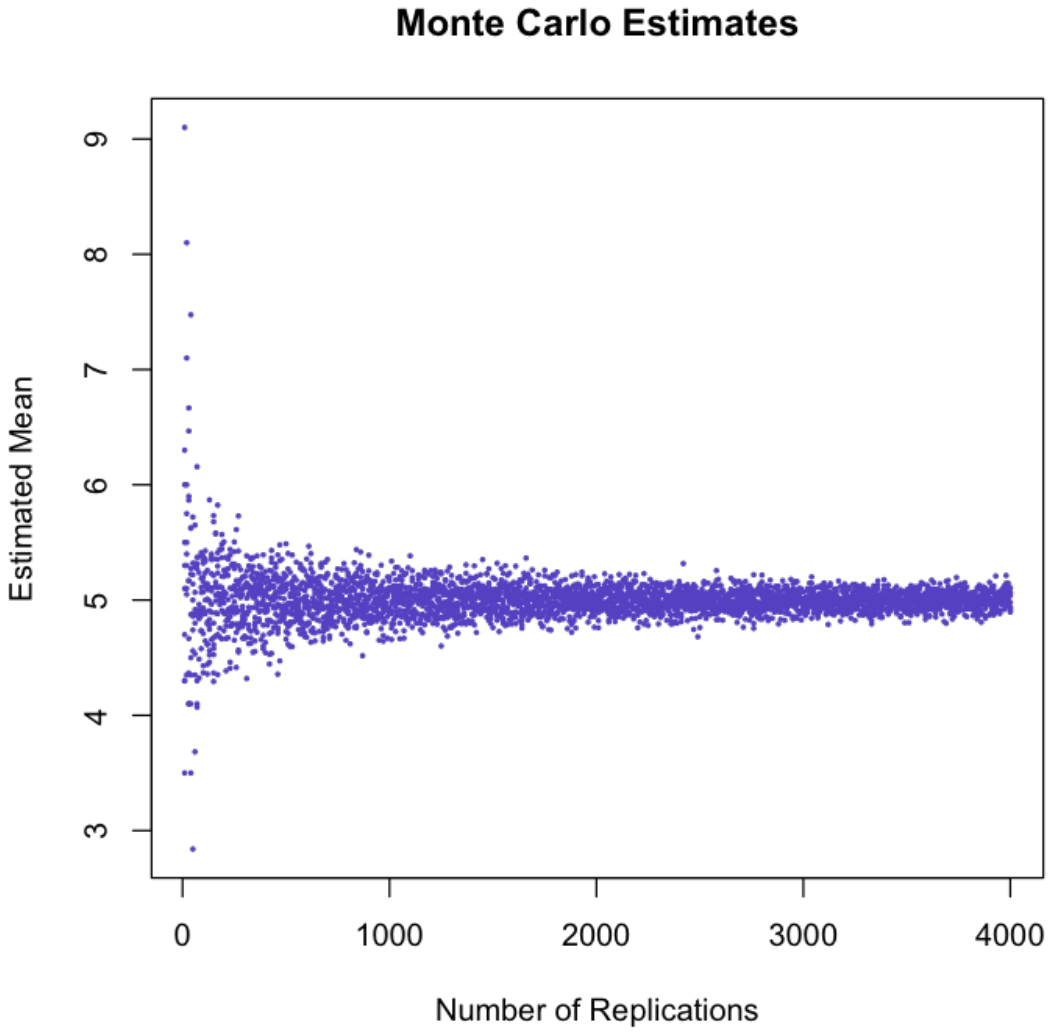
21 The Law of Large Numbers

The theoretical underpinning for Monte Carlo simulation is the *law of large numbers*. This important theorem from probability states that as the size of a sample of independent and identically distributed random variables grows to ∞, the sample mean approaches the population mean. As a reminder from probability, the law of large numbers is stated here.

The Law of Large Numbers. Let $Z_1, Z_2, ...,$ be independent and identically distributed random variables with mean μ. Then, with probability 1,

$$\frac{Z_1 + Z_2 + ... + Z_N}{N} \quad \to \quad \mu \text{ as } N \to \infty.$$

Example. Consider the problem of estimating the mean of a geoemtric random variable Z with probability of success $1/p$. Although we know ahead of time that $\mu = 1/p$, it is still instructive to see Monte Carlo simulation in action. The following R code simulates N copies of Z and returns their sample mean as an estimate of μ. It then runs the function `geometric_estimate_mu` with $p = 1/5$ for various values of N and graphs the output as a scatter plot. As the number of replications grows large the estimates become closer to the true mean of 5.



Example. Suppose that X is an exponential random variable with rate U^3 where U is uniformly distributed on the interval $[1,2]$. Calculate $P(X > 1)$. By conditioning on the value of U this probability may expressed as

$$P(X > 1) = \int_1^2 e^{-u^3} du,$$

which cannot be simplified. Monte Carlo simulation may be used to estimate the probability instead. First generate a sequence $U_1, U_2, ..., U_N$ of Uniform$[1,2]$ random variables and their corresponding exponential random variables $X_1, X_2, ...X_N$. Next, for each $n = 1, ..., N$, set

$$Z_n = \begin{cases} 1 & \text{if } X_n > 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Each Z_n is now a Bernoulli random variable with a mean of $P(X > 1)$.

The following R code implements the procedure above sequentially from $N = 1$ to $1,000$. For each value of N the previous $N - 1$ estimate is used together with the value of Z_N. As the number of replications grows the estimate converges to $P(X > 1)$ as shown in the graph.

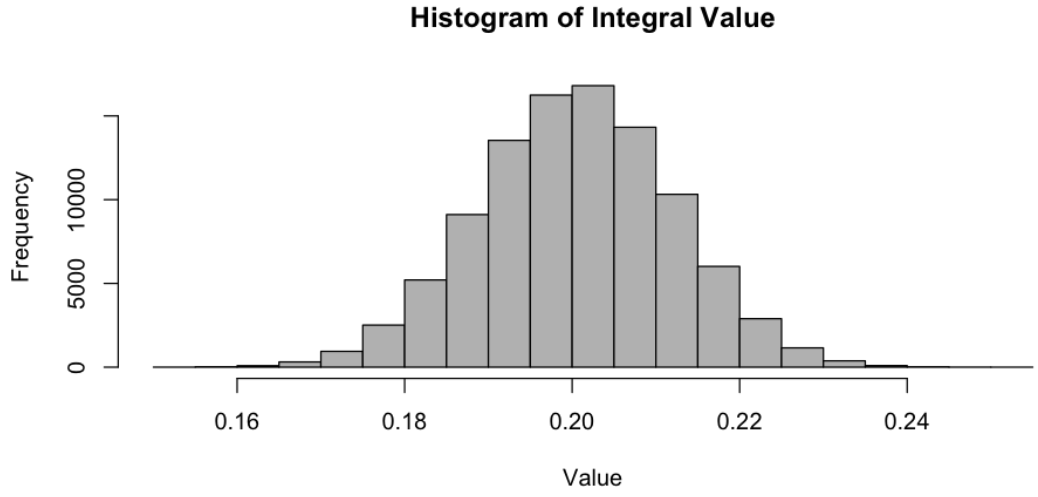
Example. One surprising application of Monte Carlo simulation is computing difficult integrals. Consider the integral

$$\int_0^1 x^{3/2} e^{-x} dx.$$

This integral does not have a nice solution but it can be computed using Monte Carlo simulation. Let $Z = U^{3/2} e^{-U}$ where U is a Uniform$[0,1]$ random variable. Then,

$$E[Z] = E[U^{3/2} e^{-U}] = \int_0^1 x^{3/2} e^{-x} dx.$$

Thus by simulating independent and identically distribution copies of Z Monte Carlo can be used to estimate the value of the integral. The following is a histogram of 100,000 Monte Carlo estimated values of the integral where each simulation was run for 100 replications.



Example. A well known application of Monte Carlo simulation is to estimate the area or volume of an object. This can be illustrated by trying to estimate the area of a circle with a radius of 1. The answer is of course π but it it useul to see how Monte Carlo can accomplish this. Plus it shows a way to use Monte Carlo to estimate π!

Suppose that a circle with radius 1 is centered inside the square $[-1,1] \times [-1,1]$. Let $U = (U_1, U_2)$ be a random point uniformly distributed inside the square and set

$$X = \begin{cases} 1 & \text{if } U_1^2 + U_2^2 \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Then, X equals 1 if U lands in the cirlce and is 0 otherwise. The mean of X is the probability that U lands in the circle which is equal to the area of the circle divided by 4, the area of the square. Setting $Z = 4X$ we now obtain a random variable whose mean is the area of the circle, or π!

In-Class Exercise. Suppose that a room contains S individuals. What is the probability at least 2 people share a common birthday? Assume that each person's birthday is uniformly distributed over all 365 days in a year (leap years are not counted here). The answer is surprisingly larger than one might expect. For intance it turns out that only 23 people are needed to achieve at least a 50% probability. Write a Monte-Carlo simulation in R that estimates the probability for an arbitrary S.

In-Class Exercise. Suppose that n guests are invited to a party and that each guest upon arrival hands their coat to an attendant. At the end of the night, when each guest leaves the party, the attendant returns to them a random coat. Write a Monte Carlo simulation to estimate the average number of guests who receive the correct coat for values of n ranging from 2 to 100.

22 An Application to Option Pricing

A European call option is an agreement between a buyer and seller where the seller provides the buyer with the *option* to buy a share of stock in a specific company at some future date $T \geq 0$ referred to as the *expiration date*. The price at which the seller agrees to sell the stock is the *strike price* of the option and is denoted by $K \geq 0$. If the market price of the stock on the expiration date is greater than the strike price of the option, then the buyer should *exercise* the option to buy the stock at K and sell it on the market at $S(T)$ yielding a profit of $S(T) - K$. If the price of the stock on the expiration date is less than the strike price, then the option expires worthless. We say that the *payoff* of the option on the expiration date is $\max(S(T) - K, 0)$.

Under certain assumptions on the behavior of the stock price the Black-Scholes formula may be used to price the option at time 0. The formula requires 3 additional quantities besides the strike price and expiration date. They are as follows. The risk-free interest rate which is denoted by r. This can be thought of as the return on a short-dated government bond. The volatility of the underlying stock which is denoted by σ. Loosely speaking this is related to the swings in price of the stock. The initial price of the stock at time 0, denoted by $S(0)$.

The Black-Scholes price of the option at time 0 is now given by

$$\Phi(d_1)S(0) - \Phi(d_2)Ke^{-rT},$$

where

$$d_1 = \frac{1}{\sigma\sqrt{T}}\left(\ln\left(\frac{S(0)}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)T\right),$$

$d_2 = d_1 - \sigma\sqrt{T}$, and Φ is the CDF of a standard normal random variable.

Example. Consider a European call option with a strike price of $K = \$45$ and an expiration date of $T = 1/2$ years from now. Suppose that the underlying stock is currently trading at $S(0) = \$40$ a share, has a volatility of $\sigma = 30\%$ and the risk-free interest rate is 2% per year. In order to price the option using the Black-Scholes formula, we first calculate

$$d_1 = \frac{1}{0.3\sqrt{1/2}}\left(\ln\left(\frac{40}{45}\right) + \left(0.02 + \frac{(0.3)^2}{2}\right)(1/2)\right) = 0.34$$

and $d_2 = 0.34 - 0.3 \times \sqrt{1/2} = 0.27$. The price of option is then given by

$$\Phi(0.34)40 - \Phi(0.27)45e^{-0.01} = \$1.74.$$

In the Black-Scholes framework the price of the underlying stock at time T is given by the lognormal random variable

$$S(T) \quad = \quad S(0)\exp((r-(1/2)\sigma^2)T+\sigma N(0,T)).$$

The value of the option at the initial time $T=0$ is given by its expected discounted payoff at the expiration date. In other words, it is given by

$$E[e^{-rT}\max(S(T)-K,0)].$$

This suggests the following Monte Carlo approach to call option pricing.

1. For each $j=1,...,N$,

```
Generate $S_j$ and

Set $Z_j = e^{-rT}max(S_j-K,0)$
```

2. Return $(Z_1+...+Z_N)/N$

The following R code implements the function `mc_call` which uses the Monte Carlo pseudocode above to price a European call option. For a given number of simulation replications, the function returns a sequence of estimates starting with the estimate formed just from Z_1 and ending with the estimate formed by Z_1 through Z_N.

Exampe. Consider again a European call option with a strike price of $K=\$45$ and an expiration date of $T=1/2$ years from now, where the underlying stock is currently trading at $S(0)=\$40$ a share, has a volatility of $\sigma=30\%$ and the risk-free interest rate is 2% per year.

In-Class Exercise. A European *put* option is an agreement between a buyer and seller where the seller provides the buyer with the *option* to sell a share of stock in a specific company at some future date $T\geq 0$ referred to as the *expiration date*. The price at which the seller agrees to buy the stock is the *strike price* of the option and is denoted by $K\geq 0$. If the market price of the stock on the expiration date is less than the strike price of the option, then the buyer should *exercise* the option to buy the stock on the market at $S(T)$ and sell the stock at K thus yielding a profit of $K-S(T)$. If the price of the stock on the expiration date is greater than the strike price, then the option expires worthless. We say that the *payoff* of the option on the expiration date is $\max(K-S(T),0)$.

Just as in the case of a European call option, in the Black-Scholes framework, assuming a risk-free interest rate of r, a volatility of the underlying stock denoted by σ, and an initial price of the stock at time 0, denoted by $S(0)$, the price of the underlying stock at time T is given by the lognormal random variable

$$S(T) \quad = \quad S(0)\exp((r-(1/2)\sigma^2)T+\sigma N(0,T)).$$

The value of the put option at the initial time $T=0$ is given by its expected discounted payoff at the expiration date. In other words, it is given by

$$E[e^{-rT}\max(K-S(T),0)].$$

Now consider a European *put* option with a strike price of $K = \$45$ and an expiration date of $T = 1/2$ years from now, where the underlying stock is currently trading at $S(0) = \$40$ a share, has a volatility of $\sigma = 30\%$ and the risk-free interest rate is 2% per year. Write a Monte Carlo simulation to estimate the price of this option based off of 10,000 simulations.

23 Confidence Intervals and The Central Limit Theorem

The sample mean estimate of μ reported from a Monte Carlo simulation is close but not equal to the true value of μ. One common way to deal with this is to use the statistical concept of a *confidence interval*. A confidence interval provides a range of possible values for μ and the true value of μ is said to lie in this interval with a certain *level of confidence* which is usually expressed as a perecentage. The confidence level can roughly be interpreted as the probability that μ lies within the confidence interval, given the samples generated.

The theoretical result used for constructing a confidence interval is the central limit theorem which we state below.

Theorem Let $Z_1, Z_2, ...,$ be independent and identically distributed random variables with finite expectation μ and variance σ^2. Then, for each $z \in \mathbb{R}$,

$$P\left(\frac{(Z_1 + Z_2 + ... + Z_N) - N\mu}{\sigma\sqrt{N}} \leq z\right) \quad \rightarrow \quad P(N(0,1) \leq z) \text{ as } n \rightarrow \infty.$$

In words, the central limit theorem states that a sum of independent and identically distributed random variables is approximately a normal random variable.

Example. Recall that an Erlang distribution with shape parameter n and scale parameter λ may be written as the sum of n independent and identically distributed exponential random variables with rate parameter λ. The central limit theorem implies that if n is large then the Erlang distribution is approximately normally distributed. The last 3 graphs below iilustrate the density function of an Erlang distribution with scale parameter 1. As n increases from 1 to 2 to 3 the density function of the Erlang distribution resembles the normal density function more and more.

Example. Recall that an Erlang distribution with shape parameter n and scale parameter λ may be written as the sum of n independent and identically distributed exponential random variables with rate parameter λ. The central limit theorem implies that if n is large then the Erlang distribution is approximately normally distributed. The last 3 graphs below iilustrate the density function of an Erlang distribution with scale parameter 1. As n increases from 1 to 2 to 3 the density function of the Erlang distribution resembles the normal density function more and more.

The central limit theorem can be used to construct confidence intervals for Monte Carlo simulation as follows. First, simulate a sequence $Z_1, Z_2, ..., Z_N$ of independent and identically distributed random variables with mean μ (the unknown quantity) and standard deviation σ and then compute their sample mean

$$\bar{Z} \quad = \quad \frac{Z_1 + Z_2 + ... + Z_N}{N}.$$

Next, let $z_{\alpha/2}$ be such that $P(N(0,1) > z_{\alpha/2}) = \alpha/2$ and set

$$L \;=\; \bar{Z} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \quad \text{and} \quad U \;=\; \bar{Z} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}.$$

It then follows by the central limit theorem that approximately

$$P(L \leq \mu \leq U) \;=\; 1 - \alpha.$$

Thus, thus confidence interval

$$\left[\bar{Z} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{Z} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

yields a confidence level of $100(1-\alpha)\%$ for the unknown quantity μ.

The pseudocode below summarizes the procedure for finding a $100(1-\alpha)\%$ confidence interval.

1. For j=1,. . . ,N,

`Generate $Z_j$`

2. Set $\bar{Z} = (Z_1 + ... + Z_N)/N$ and

$$L \;=\; \bar{Z} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad U \;=\; \bar{Z} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

3. Return [L,U]

The standard deviation σ of each Z_j is necessary to construct the confidence interval above but in most cases it is not known. Replacing σ by the *sample standard deviation* s is one way to deal with this. Recall that if $Z_1, Z_2, ..., Z_N$ is a sequence of independent and identically distributed random variables, then their sample standard deviation is

$$s \;=\; \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(Z_j - \bar{Z})^2},$$

where $\bar{Z}$ as usual is their sample mean. In this case,

$$\left[\bar{Z} - z_{\alpha/2}\frac{s}{\sqrt{N}}, \bar{Z} + z_{\alpha/2}\frac{s}{\sqrt{N}}\right]$$

may be used as a $100(1-\alpha)\%$ confidence interval for μ.

Sometimes in order to account for the fact that $\bar{Z}$ is only approximately normal, $z_{\alpha/2}$ is replaced with the equivalent quantile from Student's t-distribution. Assuming that N random variables from the distribution of Z have been simulated, Student's t-distribution with $N-1$ degrees of freedom should be used. Denoting the corresponding quantile by $t_{\alpha/2,N-1}$, the $100(1-\alpha)\%$ confidence interval is then given by

$$\left[\bar{Z} - t_{\alpha/2,N-1}\frac{s}{\sqrt{N}}, \bar{Z} + t_{\alpha/2,N-1}\frac{s}{\sqrt{N}}\right].$$

The quantile $t_{\alpha/2,N-1}$ from Student's t-distribution is larger than the corresponding quantile $z_{\alpha/2}$ from the normal distribution and so using it produces wider and more conservative confidence intervals.

24 Confidence Intervals for Option Pricing

Consider again a European call option with a strike price of $K = \$45$ and an expiration date of $T = 1/2$ years from now. Also suppose that the underlying stock is currently trading at $S(0) = \$40$ a share, has a volatility of $\sigma = 30\%$ and the risk-free interest rate is 2% per year.

The R code below calls the function `mc_call_ci` 200 times, each time calculating a 95% confidence interval based on 10,0000 simulations. The output is then presented as a graph. Most of the confidence intervals contain the true price of the call option but their upper and lower limits varry.

25 Run Length Control for Confidence Intervals

The process of determining the minimum number of simulation replications required in order to ensure that the Monte Carlo estimation error of μ is less than or equal to some pre-specified amount is referred to as *run length control*.

There are several methods of run length control, two of which are discussed below. Both of these approaches are based on the following analysis. Suppose that due to either internal or external constraints, some unknown quantity μ must be estimated by a confidence interval with an error of no more than $\kappa > 0$ at a confidence level of $100(1-\alpha)\%$. Next recall that if N replications of the unbiased random variable Z are generated, the resulting $100(1-\alpha)\%$ confidence interval is

$$\left[\bar{Z} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{Z} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right].$$

Here, σ is the standard deviation of Z and is assumed to be known. In this case, with confidence level $100(1-\alpha)\%$ the maximum estimation error made when using this confidence interval is $z_{\alpha/2}\sigma/\sqrt{N}$. So, the minimum number of replications of Z that must be generated is given by

$$N \geq \frac{z_{\alpha/2}^2\sigma^2}{\kappa^2}.$$

The one problem in implementing this analysis is that in most cases the value of σ, the standard deviation of Z, is not known ahead of time. The following two methods address this issue.

25.1 The Two Stage Approach

In the two stage approach, two independent batches of replications of the random variable Z are created. In the first batch, $N_1 \geq 1$ replications of Z are generated and from these replications the sample standard deviation of Z, denoted by s, is calculated. The standard deviation σ is then replaced by the sample standard deviation s in the inequality above and it is determined if the number N of replications already run satisfies the inequality or not. If it turns out that enough simulations have already been run, then the approach proceeds directly to calculating the sample mean estimate of μ and the corresponding confidence interval from the N replications already run. On the other hand, if the inequality is not satisfied, then the approach continues to run enough replications until it is. Generally speaking, the value of N_1 should be set large enough to ensure a reasonably accurate estimate of the sample standard deviation but not too large that it substantially reduces the efficiency of the approach.

Pseudocode for the two stage approach is given below.

1. For j=1,. . . ,N_1, Generate Z_j.

2. Set $\bar{Z} = (Z_1 + ... + Z_{N_1})/N_1$ and

$$s = \sqrt{\frac{1}{N_1 - 1} \sum_{j=1}^{N_1} (Z_j - \bar{Z})^2}.$$

3. Set $N^\star = \lceil z_{\alpha/2}^2 s^2 / \kappa^2 \rceil$.

4. If $N_1 \geq N^\star$, return $\bar{Z}$ and its associated confidence interval

$$\left[\bar{Z} - z_{\alpha/2} \frac{s}{\sqrt{N_1}}, \bar{Z} + z_{\alpha/2} \frac{s}{\sqrt{N_1}} \right].$$

5. If $N_1 < N^\star$, generate Z_j for $j = N_1 + 1, ..., N^\star$.

6. Set $\bar{Z} = (Z_1 + ... + Z_{N^\star})/N^\star$ and

$$s = \sqrt{\frac{1}{N^\star - 1} \sum_{j=1}^{N^\star} (Z_j - \bar{Z})^2}.$$

7. Return $\bar{Z}$ and its associated confidence interval

$$\left[\bar{Z} - z_{\alpha/2} \frac{s}{N^\star}, \bar{Z} + z_{\alpha/2} \frac{s}{\sqrt{N^\star}} \right].$$

25.2 The Sequential Approach

The sequential approach updates its estimate of the sample mean and sample standard deviation of Z after each new replication of Z_j . It then checks if the inequality above is satisfied by the updated sample mean, sample standard deviation and the increased number of replications. The moment the inequality is satisfied, the approach returns the most recent sample mean along with the desired confidence interval.

Pseudocode for the sequential approach is given below.

1. Set $N = 0$

2. While TRUE

Set $N = N + 1$.

Generate Z_N.

Set $\bar{Z} = (Z_1 + ... + Z_N)/N$ and

$$s = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(Z_j - \bar{Z})^2}.$$

Set $N^\star = \lceil z_{\alpha/2}^2 s^2/\kappa^2 \rceil$.

If $N \geq N^\star$, end while.

4. Return $\bar{Z}$ and its associated confidence interval

$$\left[\bar{Z} - z_{\alpha/2}\frac{s}{\sqrt{N}}, \bar{Z} + z_{\alpha/2}\frac{s}{\sqrt{N}}\right].$$

25.3 Example

Consider again a European call option with a strike price of $K = \$45$ and an expiration date of $T = 1/2$ years from now. Also suppose that the underlying stock is currently trading at $S(0) = \$40$ a share, has a volatility of $\sigma = 30\%$ and the risk-free interest rate is 2% per year.

The R code below implements the sequential approach which guarantees an error of no more than \$0.10 when using Monte Carlo simulation to price this option. The blue curve on the output graph is the approach's estimate of the minimum number of replications needed in order to achieve an error of no more than \$0.10. The orange line is a 45 degree line. The point at which the two lines cross is the minimum number of replications required in order to achieve the desired level of accuracy.

26 Absolute vs. Relative Errors

There are at least 2 ways to compare an estimate $\bar{Z}$ to its true value μ. The first is referred to as the *absolute error* and is given by $|\bar{Z} - \mu|$. Both the two stage and the sequential approach control the

simulation run length in order to bring the absolute error below some predefined threshold. One potential pitfall of using absolute error is that if the true value itself is tiny, then even seemingly small absolute errors may be large relative to the true value itself. Oppositely, first glance big absolute errors may actually be relatively small if the true value itself is large. This leads to the concept of *relative error* which is defined by

$$\frac{|\bar{Z} - \mu|}{|\mu|}.$$

Notice that the numerator in the definition of relative error is just the absolute error and the $|\mu|$ in the denominator is a normalizing constant. Both of the run length control algorithms discussed above can be modified in order to reduce the relative error below some predefined threshold.

27 Processing Time Allocation

An estimator $\bar{Z}$ is said to be *unbiased* if its expectation is equal to the true value μ of the quantity being estimated. Often times one may have available two unbiased estimators of the same underlying quantity μ, and a decision must be made on which of the two estimators to use. One way to make this decision is to choose the estimator with the lower variance. This make sense since all else being equal, a lower variance estimator will result in tighter confidence intervals.

In many settings there may only be a limited amount of computer processing time available for running a simulation algorithm and creating an estimate. In this setting variance is still an important factor in selecting an estimator but it is not the only consideration either. For instance, consider again two unbiased sample mean estimators $\bar{Z}$ and $\bar{W}$ which are based off of the replications $Z_1, Z_2, ...$, and $W_1, W_2, ...$, respectively. Suppose that the variance of each Z_j is smaller than the variance of each W_j but the amount of processing time required to generate a Z_j is greater than the amount of processing time required to generate a W_j. In this case, even though each Z_j may have a smaller individual variance than a corresponding W_j, one cannot simulate as many Z_j as W_j due to the high-level processing time constraint on the total amount of time that can be devoted to the simulation.

In order to balance the tradeoff between the variance and the run time of an estimator, it is helpful to introduce some notation. Suppose that $\bar{Z}$ is a sample mean estimator which is based off of generating a sequence of random variables $Z_1, Z_2, ...$, where each Z_j has a variance of σ^2 and takes τ units of processing time to generate. Also suppose that due to external constraints there are only T total units of processing time available to devote to the simulation. In this case, because each replication requires τ units of processing time, and only a total of T units of time are available, it follows that at most $\lfloor T/\tau \rfloor$ replications of Z_j can be generated before terminating the simulation. Hence, the estimator is given by

$$\bar{Z} \quad = \quad \frac{Z_1 + Z_2 + ... + Z_{\lfloor T/\tau \rfloor}}{\lfloor T/\tau \rfloor}.$$

Now recall from the central limit theorem that $\bar{Z}$ may be approximated by a normal random variable with a mean of μ and a variance of $\sigma^2/\lfloor T/\tau \rfloor$, which is roughly equal to $\sigma^2\tau/T$. Thus, if

several sample mean estimators are available to choose from, each with different variances and processing times per replication, in order to reduce the variance of $\bar{Z}$, one should select the estimator whose replications which minimizes $\sigma^2\tau$.

28 Biased Estimators

An estimator $\bar{Z}$ is said to be *biased* if $E[\bar{Z}]$ is not equal to μ, the underlying quantity which is to be estimated. It might seem strange at first to construct an estimate which is biased, however in many cases it may be very difficult or overly complicated to construct an estimate which is unbiased.

Example. Consider the problem of estimating the ratio of the expectations of two random variables X and Y. That is, the problem of estimating $E[X]/E[Y]$. In this case a natural approach is to generate say n samples $X_1, ..., X_n$ from the distribution X and n samples $Y_1, ..., Y_n$ from the distribution Y, and then to take the ratio of their sample means. That is to compute $\bar{X}/\bar{Y}$. However, this estimate is biased as

$$E\left[\frac{\bar{X}}{\bar{Y}}\right] \quad \neq \quad \frac{E[X]}{E[Y]}.$$

Even though the estimate $\bar{X}/\bar{Y}$ in the example above is biased, the bias becomes smaller as the number of samples drawn from the distributions of X and Y increases. This is a common feature of many unbiased estimators. One can reduce their bias at the expense of taking up more processing time. Although not the case in the example above, an increase in processing time is often required to lower the bias of each individual replication X_j. In a situation with limited processing time, one has to make a decision between simulating a small number of replications each with a low bias or a larger number of replications but each with a higher bias. The central limit theorem implies that the benefit of simulating a larger number of replications is that it reduces the variance of the sample mean estimate $\bar{X}$.

There exists an error measurement which captures both the bias and the variance in an estimate. The *mean squared error* of an estimate $\bar{X}$ relative to its true value μ is given by

$$\text{MSE} \quad = \quad E[(\bar{X} - \mu)^2].$$

The expression on the righthand side above can be decomposed as

$$E[(\bar{X} - \mu)^2] \quad = \quad (E[\bar{X}] - \mu)^2 + E[(\bar{X} - E[\bar{X}])^2].$$

Note that the first term on the righthand side above is the bias squared and the second term is the variance of the estimator. Hence, by selecting an estimator that minimizes the mean squared error one is able to minimize the sum of these two important quantities.

29 Session 5: Variance Reduction

Summer 2019 - Instructor: Josh Reed

Teaching Assistant: Haotian Song

Recall that in Monte Carlo simulation if $\bar{Z}$ is the sample mean estimator of some unkown quantity μ, then the $100(1-\alpha)\%$ confidence interval for μ is given by

$$\left[\bar{Z} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{Z} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

where σ^2 is the variance of the Z_j's used to form $\bar{Z}$. In this session we discuss two methods which change the distribution of Z in order to reduce its variance. The benefit of these *variance reduction techniques* is that they result in tighther confidence intervals and more accurate estimates of μ.

30 The Method of Control Variates

Suppose that $(X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N)$ is a sequence of independent and identically distributed random pairs where the mean of each X_j is equal to μ, the unknown quantity that we would like to estimate. Next, let μ_Y be the mean of the Y_j's and b be a real number (to be chosen further below) and set

$$Z_j \quad = \quad X_j - b(Y_j - \mu_Y), \;\; j = 1, 2, ..., N.$$

It then follows that $\bar{Z}$, the sample mean of the Z_j's is an unbiased estimator of μ.

The idea behind control variates is to select a good value of b which will reduce the variance of the Z_j's realtive to the variance of the X_j's. This will result in $\bar{Z}$ having a lower variance than $\bar{X}$ and so tighter confidence intervals too. It turns out that the optimal b to use is given by

$$b^\star = \text{Cov}(X, Y)/\sigma_X^2$$

and that using this value of b the ratio between the variances of $\bar{Z}$ and $\bar{X}$ is $1 - \rho_{X,Y}^2$, where the correlation between X and Y is

$$\rho_{X,Y} \quad = \quad \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

This implies that the control variates approach works best when there is a high correlation between X and the control variate Y.

One problem when implementing the control variate approach is that $\text{Cov}(X, Y)$ and σ_X^2 may not be known ahead of time in order to determine $b^\star$. In this case $b^\star$ may be approximated by

$$b^N = \frac{\sum_{j=1}^N (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^N (X_j - \bar{X})^2}.$$

If N is large enough this will be close to the true value of $b^\star$.

Example. Consider the problem of estimating $E[e^U]$ where U is a Uniform$[0,1]$ random variable. In this case ordinary Monte Carlo results in an estimator with a variance of $\text{Var}(e^U) = 0.2420$. On the other hand using U as a control variate results in an estimator with a variance of 0.0039. The following R code plots both the ordinary Monte Carlo estimate (red dots) and the optimal control variate estimate (blue dots)

Example. Let X be a random variable and consider the problem of estimating $P(X \leq x)$ for some value x. As discussed previously this can be accomplished using the Bernoulli random variable

$$Y = \begin{cases} 1 & \text{if } X \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

A good control variate to pick in this case is the value of X itself. This is because X is usually highly negatively correlated with Y. The optimal value of b is given by

$$b^\star = \frac{E[X|X \leq x] - E[X]}{P(X > x)}$$

and the relative reduction in variances is

$$1 - \rho_{XY}^2 = 1 - \frac{1}{\sigma_X^2} \left(\frac{F(x)}{1 - F(x)} \right) (E[X|X < x] - E[X])^2.$$

Example. Consider our usual example of a European call option with a strike price of $K = \$45$ and an expiration date of $T = 1/2$ years from now. The underlying stock has an initial price at time $T = 0$ of $S(0) = \$40$ a share, a volatility of $\sigma = 30\%$ and we assume that the risk-free interest rate is 2% per year.

We now price this option by Monte Carlo simulation using the method of control variates. It is important that any control variate we pick possess two properties. The first is that its expectation be known ahead of time. The second is that it possess a high degree of correlation with the original estimate. In the case of a European call option, one natural choice which satisfies both of these criteria is the terminal price $S(T)$ of the underlying asset. In particular, recalling that $S(T)$ may be written as

$$S(T) = S(0) \exp\left(\left(r - \frac{\sigma^2}{2} \right) T + \sigma N(0,T) \right),$$

it follows that $E[S(T)] = \$40.40$. **In-Class Exercise.** Consider a European *put* option with a strike price of $K = \$45$ and an expiration date of $T = 1/2$ years from now. Also suppose that the

underlying stock is currently trading at $S(0) = \$40$ a share, has a volatility of $\sigma = 30\%$ and the risk-free interest rate is 2% per year.

Use the method of control variates to price this option based on 10,0000 simulation replications where the terminal price of the underlying stock is used as the control variate.

31 Multiple Control Variates

The control variates approach can be extended to *multiple* control variates. Suppose that

$$(X_1, Y_{1,1}, Y_{1,2}, ..., Y_{1,r}), (X_2, Y_{2,1}, Y_{2,2}, ..., Y_{2,r}), ..., (X_N, Y_{N,1}, Y_{N,2}, ..., Y_{N,r})$$

is a sequence of independent and identically distributed random vectors where the mean of each X_n is equal to μ, the unknown quantity that we would like to estimate. Next, let μ_p be the mean of the $Y_{n,p}$'s and $(b_1, b_2, ..., b_r)$ a vector of coefficients to be discussed below. Then, for each simulation replication $n = 1, ..., N$, set

$$Z_n = X_n - b_1(Y_{n,1} - \mu_1) - b_2(Y_{n,2} - \mu_2) - .... - b_r(Y_{n,r} - \mu_r).$$

It follows that the variance of each Z_n is given by

$$\sigma_Z^2 = \sigma_X^2 - 2\sum_{p=1}^{r} b_p \text{Cov}(X_1, Y_{1,p}) + \sum_{p=1}^{r}\sum_{q=1}^{r} b_p b_q \text{Cov}(Y_{1,p}, Y_{1,q}).$$

and that $\bar{Z}$, the sample mean of the Z_n's, is an unbiased estimator of μ.

The optimal coefficients $b_1^\star, ..., b_r^\star$ to minimize σ_Z^2 can be difficult to find since in most situations the covariances terms are not known. An alternative is to estimate each covariance term individually and then minimize the variance expression above resulting in a coefficient vector $b^N = (b_1^N, ..., b_r^N)$. A second characterization of b^N is the solution to the least squares linear regression

$$X_n = a + b_1 Y_{n,1} + b_2 Y_{n,2} + ... + b_r Y_{n,r} + \varepsilon_n,$$

where ε_n is a random noise term for $n = 1, ..., N$.

31.1 The Asian Call Option

The payoff of an *Asian call option* depends on the average price of the underlying stock over a set of *monitoring dates*. Specifically, suppose that an Asian call option has monitoring dates $t_1, t_2, ..., t_M$ and a strike price of K. Then, at its expiration date $T \geq 0$ the option has a payoff of $\max(\bar{S} - K, 0)$, where

$$\bar{S} = \frac{1}{M}(S(t_1) + S(t_2) + ... + S(t_M)).$$

The price of the option at the initial time $T = 0$ is equal to its expected discounted payoff $e^{-rT}E[\max(\bar{S} - K, 0)]$.

When pricing an Asian call option in the Black-Scholes framework it is assumed that the distribution of the random vector $(S(t_1), S(t_2), ..., S(t_M))$ is a multivariate log normal distribution with parameters μ and Σ, where

$$\mu_m = \left(r - \frac{\sigma^2}{2}\right) t_m$$

for $m = 1, ..., M$, and

$$\Sigma_{m,p} = \sigma^2 \min(t_m, t_p)$$

for $m = 1, ..., M$, and $p = 1, ..., M$. As usual r is the risk-free interest rate and σ is the volatility of the underlying stock. This is equivalent to saying that for each monitoring date t_m,

$$S(t_m) = \exp(X_m),$$

where $(X_1, ..., X_M)$ is a multivariate normal random vector with mean vector μ and variance-covariance matrix Σ.

Example. Consider an Asian call option with a strike price of $K = \$45$ and an expiration date of $T = 1/2$ years from now. Also assume that the underlying stock has an initial price at time $T = 0$ of $S(0) = \$40$ a share, a volatility of $\sigma = 30\%$ and that the risk-free interest rate r is 2% per year. Suppose also that there are 4 monitoring dates of the option given by $1/8, 1/4, 3/8$ and $1/2$ of a year from now.

One natural multiple control variate to use in this case is the vector of stock prices at each of the monitoring dates. That is, to let

$$Y_n = (S_n(1/8), S_n(1/4), S_n(3/8), S_n(1/2))$$

for $n = 1, ..., N$. The expected value of the stock price on each of the monitoring dates is given by

$$E[S(1/8)] = \$40.10, \quad E[S(1/4)] = \$40.20, \quad E[S(3/8)] = \$40.30, \quad E[S(1/2)] = \$40.40.$$

32 Antithetic Variables

Antithetic variables is a technique to reduce simulation variance by generating *pairs* of random variables where each random variable in the pair has an expectation equal to the unknown quantity μ but there is some negative correlation between the two components. The idea is as follows.

Suppose that $(X_1, \dot{X}_1), (X_2, \dot{X}_2), ..., (X_N, \dot{X}_N)$ is a sequence of independent and identically distributed pairs of random variables where each X_n and $\dot{X}_n$ have the same distribution whose mean is equal to the unknown quantity μ. Ordinarily there will be some correlation between X_n and $\dot{X}_n$ in order for the method to be effective. Now setting

$$Z_n = \frac{X_n + \dot{X}_n}{2},$$

each Z_n has a mean equal to μ and a variance which may be written as

$$\sigma^2 = 2\text{Var}(X) + \text{Cov}(X_1, \dot{X}_1).$$

The antithetic estimator of μ is given by $\bar{Z}$, the sample means of the Z_n's. Recalling that each Z_n requires 2 replications (one for X_n and one for $\dot{X}_n$) we now see that the antithetic variable estimator $\bar{Z}$ has a smaller variance than the ordinary Monte Carlo estimator $\bar{X}$ (with $2n$ samples) if

$$\text{Cov}(X_1, \dot{X}_1) < 0.$$

Example. Suppose that $X_1, X_2, ..., X_N$ is a sequence of independent and identically distributed random variables whose mean is equal to the unknown quantity μ. Moreover, suppose that each X_n is generated by the inverse transform method. That is,

$$X_n = F^{-1}(U_n),$$

where F is the CDF of the X_n's and U_n is a Uniform$[0, 1]$ random variable. Ordinary Monte Carlo will estimate μ by the sample mean $\bar{X}$. In the antithetic approach a second sequence $\dot{X}_1, \dot{X}_2, ..., \dot{X}_N$ of independent and identically distributed random variables can be generated by setting

$$\dot{X}_n = F^{-1}(1 - U_n),$$

where U_n is *the same* Uniform$[0, 1]$ random variable used above to generate X_n. Because $1 - U_n$ is also a Uniform$[0, 1]$ random variable, X_n and $\dot{X}_n$ have the same distribution but since the same U_n is used to generate both of them, there is some correlation between the two. In fact, it can be shown that $\text{Cov}(X_n, \dot{X}_n) < 0$ so it is always the case that this approach results in a variance reduction.

Example. Suppose that X_n is an exponential random variable with rate 1 that is generated by using the inverse transform method. This means that

$$X_n = -\ln(1 - U_n)$$

where U_n is a Uniform$[0, 1]$ random variable. Its corresponding antithetic random variable is given by

$$\dot{X}_n = -\ln(U_n)$$

.

Example. If both X_n and $\dot{X}_n$ are functions of some underlying variable Y_n, where X_n is an increasing function of Y_n, and $\dot{X}_n$ is a decreasing function of Z_n, then it is always the case that $\text{Cov}(X_n, \dot{X}_n) < 0$ and the antithetic approach results in a variance reduction.

Example. Consider again the problem of estimating $E[e^U]$ where U is a Uniform$[0,1]$ random variable. In this case the ordinary Monte Carlo estimator results in a variance of $\text{Var}(e^U) = 0.2420$. On the other hand, setting

$$Z = \frac{e^U + e^{1-U}}{2},$$

results $E[Z] = E[e^U]$ and $\text{Var}(Z) = 0.0039$.

32.1 An Application to Option Pricing

Recall that in the Black-Scholes framework the price of a European call option with a strike price of K expiring at time T is given by $E[X]$ where

$$X = e^{-rT} \max(S(T) - K, 0)$$

is the discounted payoff of the option at expiration. We now apply the antithetic variable technique by finding a random variable $\dot{X}$ which has the same distribution as X but is negatively correlated with it. In order to do so recall that

$$S(T) = S(0) \exp\left(\left(r - \frac{\sigma^2}{2}\right) T + \sigma N(0,T)\right),$$

where $N(0,T)$ is a normal random variable with a mean of 0 and a variance of T. Since the distribution of $N(0,T)$ is symmetric about the origin it has the same distribution of $-N(0,T)$ and so we can set

$$\dot{X} = e^{-rT} \max(\dot{S}(T) - K, 0),$$

where

$$\dot{S}(T) = S(0) \exp\left(\left(r - \frac{\sigma^2}{2}\right) T - \sigma N(0,T)\right).$$

Both X and $\dot{X}$ will have the same distribution as one another. Moreover, X is increasing in $N(0,T)$ and $\dot{X}$ is decreasing in $N(0,T)$ and so $(X, \dot{X})$ is a candidate antithetic pair.

The following R code implements the function `mc_european_call_antithetic` which prices a European call option using Monte Carlo simulation with the antithetic variables described above. For a given number of simulation replications, the function returns a sequence of running estimates.

Example. Consider as usual a European call option with a strike price of $K = \$45$ and an expiration date of $T = 1/2$ years from now. The underlying stock has an initial price at time $T = 0$ of $S(0) = \$40$ a share, a volatility of $\sigma = 30\%$ and we assume that the risk-free interest rate is 2% per year.

In-Class Exercise. Consider a European *put* option with a strike price of $K = \$45$ and an expiration date of $T = 1/2$ years from now. Also suppose that the underlying stock is currently trading at $S(0) = \$40$ a share, has a volatility of $\sigma = 30\%$ and the risk-free interest rate is 2% per year.

Use the method of antithetic variables to price this option based on 10,0000 simulation replications.

33 Session 6: Variance Reduction Continued

Summer 2019 - Instructor: Josh Reed

Teaching Assistant: Haotian Song

In this session, stratified sampling and importance sampling variance reduction techniques are introduced.

34 Stratified Sampling

Suppose that Z is a random variable whose mean is equal to some unknown quantity μ which we would like to estimate. Stratified sampling is a variance reduction technique where the set of values that Z can take is first partitioned into different subsets or "strata" and then Z is conditionally generated from each stratum a specified number of times. The sequence of random variables generated from this procedure are not identically distributed. Their sample mean is however an unbiased estimator of μ which is sufficient for most Monte Carlo purposes.

The details of stratified sampling are as follows. The set of values that Z can take is first partitioned into a family of disjoint subsets $A_1, A_2, ..., A_K$. Next, let

$$p_k \quad = \quad P(Z \in A_k),\ k = 1, ..., K,$$

be the probability that Z lies in the k^{th} stratum. Moreover, suppose that it has been determined ahead of time that a total of N replications of Z will be simulated. Then, the algorithm generates $N_k = Np_k$ replications of Z conditional on Z lying in the kth stratum. That is, corresponding to the k^{th} stratum, a sequence of independent random variables $Z_{k,1}, Z_{k,2}, ..., Z_{k,N_k}$ is generated who distribution is given by

$$P(Z_{k,n} \leq x) \quad = \quad P(Z \leq x \mid Z \in A_k).$$

We assume for convenience that each N_k is an integer. The estimator of μ is given by

$$\bar{Z} \quad = \quad \frac{1}{n} \sum_{k=1}^{K} \sum_{n=1}^{N_k} Z_{k,n}.$$

The fact that $\bar{Z}$ is unbiased follows from the identity

$$E[Z] \quad = \quad p_1 E[Z \mid Z \in A_1] + p_2 E[Z \mid Z \in A_2] + ... + p_K E[Z \mid Z \in A_K].$$

The variance of $\bar{Z}$ is given by

$$\mathrm{Var}(\bar{Z}) \quad = \quad \frac{1}{N^2} \sum_{k=1}^{K} N_k \sigma_k^2 \; = \; \frac{1}{N} \sum_{k=1}^{K} p_k \sigma_k^2,$$

where σ_k^2 is the variance of the $Z_{k,n}$'s. It may be shown that $\mathrm{Var}(\bar{Z})$ is *always* smaller than the variance of the estimator formed by averaging N independent and identically distributed samples from the distribution of Z.

The quantities σ_k^2 in the expression above for the variance of the stratified sampling estimator are typically unknown, which means they cannot be used to construct confidence intervals for $\bar{Z}$. However, they may be estimated from the simulation output. Specifically, for each $k = 1, ..., K$, let

$$\sigma_k \quad = \quad \sqrt{\frac{1}{N_k - 1} \sum_{n=1}^{N_k} (Z_{k,n} - \bar{Z}_k)^2}$$

be the sample standard deviation of $Z_{k,1}, Z_{k,2}, ..., Z_{k,N_k}$, where $\bar{Z}_k$ is the sample mean of $Z_{k,1}, Z_{k,2}, ..., Z_{k,N_k}$. Then, setting

$$s^2 \quad = \quad \sum_{k=1}^{K} p_k \sigma_k^2,$$

it follows that a $100(1-\alpha)\%$ confidence interval for $\bar{Z}$ is given by

$$\left[\bar{Z} - z_{\alpha/2} \frac{s}{\sqrt{N}}, \bar{Z} + z_{\alpha/2} \frac{s}{\sqrt{N}} \right].$$

Example. Consider the problem of sampling N random variables which are uniformly distributed in the square $[0,1] \times [0,1]$. This can be accomplished using ordinary Monte Carlo by generating

two independent Uniform$[0,1]$ random variables U_1 and U_2 and then forming the pair (U_1, U_2). Stratified sampling may also be applied. For instance the square $[0,1] \times [0,1]$ may be partitioned into 4 quadrants with $N/4$ random variables uniformly sampled from each quadrant. The R code below illustrates this stratified sampling for $N = 10,000$ with the samples from each quadrant labeled different colors.

Example. Suppose that

$$f(u) = \begin{cases} -u & \text{if } -1 \leq u \leq 0, \\ u^2 & \text{if } \quad 0 \leq u \leq 1. \end{cases} \tag{3}$$

Then, the integral

$$\int_{-1}^{1} f(u)du$$

is equal to $E[f(U)]$ where U is a Uniform$[-1,1]$ random variable. This expectation may be estimated using stratified sampling by first setting $Z_{1,n} = f(U_{1,n})$ where $U_{1,1}, U_{1,2}, ..., U_{1,N}$ are Uniform $[-1,0]$ random variables, next setting $Z_{2,n} = f(U_{2,n})$ where $U_{2,1}, U_{2,2}, ..., U_{2,N}$ are Uniform $[0,1]$ random variables and finally setting

$$\bar{Z} \quad = \quad \frac{1}{2N}\left(\sum_{n=1}^{N} Z_{1,n} + \sum_{n=1}^{N} Z_{2,n}\right).$$

Example. Consider a game show where contestants select 1 of 4 doors and receive whatever prize is behind it. The distribution of the value of the prize behind door i has CDF F_i for $i = 1,2,3,4$. Let Z be a generic random variable represnting the prize that the contestant receives assume that there is a $1/4$ chance of selecting each door. Rather than using the composition method to simulate Z and estimate its expected value, stratified sampling may be used. In this case, exactly $1/4$ of the samples come from the distribution F_i of the value of the prize behind door i for $i = 1,2,3,4$.

35 An Application to Option Pricing

Recall that in the Black-Scholes framework the price of a European call option with a strike price of K expiring at time T is given by $E[Z]$ where

$$Z = e^{-rT}\max(S(T) - K, 0)$$

is the discounted payoff of the option at expiration. We now apply stratified sampling to estimate $E[Z]$. In order to do so recall that

$$S(T) = S(0)\exp\left(\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}N(0,1)\right),$$

where $N(0,1)$ is a standard normal random variable. In this example we choose to stratify the standard normal random variable $N(0,1)$. In order to do so we select 4 strata corresponding to

the 1st, 2nd, 3rd and 4th quantiles of the distribution of $N(0,1)$. The corresponding intervals are given by

$$(-\infty, -0.67], \quad (-0.67, 0], \quad (0, 0.67], \quad (0.67, +\infty).$$

Since each strata corresponds to a quantile, it follows that $p_k = 1/4$ for $k = 1, 2, 3, 4$.

Example. Consider a European call option with a strike price of $K = \$45$ and a expiration date of $T = 1/2$ a year from now. Assume that the underlying stock has an initial price at time $T = 0$ of $S(0) = \$40$ a share, a volatility of $\sigma = 30\%$ and the risk-free interest rate is 2% per year. The R code below uses the function `mc_call_strat` to graph 100 instances of the running estimate of the option price for up to 1,600 simulation replications.

35.1 Optimal Sampling Proportions for Stratified Sampling

In the above stratified sampling procedure the number of samples drawn from each strata is in direct proportion to the probability that the random variable Z lies in that strata. A more general sampling procedure is as follows.

Suppose as above that the stratum $A_1, A_2, ..., A_K$ have been chosen already and that N samples of Z will be generated with N_k samples coming from stratum k for $k = 1, ..., K$, where

$$N_1 + N_2 + ... + N_k = N.$$

We do not assume that $N\_k = Np\_k$ where $p_k = P(Z \in A_k)$ but do require that the sum of the N_k's be equal to N. Now let $Z_{k,1}, Z_{k,2}, ..., Z_{k,N_k}$ be the sequence of random variables corresponding to the kth strata. Then, an unbiased estimator of μ is given by

$$\bar{Z} \quad = \quad \frac{1}{N} \sum_{k=1}^{K} \frac{p_k}{q_k} \sum_{n=1}^{N_k} Z_{k,n},$$

where $q_k = N_k/N$ is the fraction of samples that correspond to the kth strata. If $q_k = p_k$, this estimator corresponds to the original stratified sampling estimator above.

It turns out that the variance of the estimator is given by

$$\mathrm{Var}(\bar{Z}) \quad = \quad \frac{1}{N} \sum_{k=1}^{K} \frac{p_k}{q_k} \sigma_k^2,$$

where σ_k^2 is the variance of the $Z_{k,n}$'s, and the optimal q_k's in order to minimze the variance are given by

$$q_k^\star \quad = \quad \frac{p_k \sigma_k}{\sum_{k=1}^{K} p_k \sigma_k},$$

for $k = 1, ..., K$. Substituting this expression into the formula for the variance of $\bar{Z}$ given above, the minimum variance for the estimator is

$$\frac{1}{n}\left(\sum_{k=1}^{K} p_k \sigma_k^2\right).$$

When the standard deviations σ_k for $k = 1, ..., K$, are not known ahead of time a test run can be performed for each strata where the conditional random variables are simulated and the sample standard deviation of the strata is calculated from the output. The quantities $q_k^\star$ are next estimated using the formula above with the true standard deviation replaced by the sample standard deviation. One may then proceed as usual with the stratified sampling algorithm.

Example. Consider again the problem of computing the integral

$$\int_{-1}^{1} f(u)du$$

where

$$f(u) = \begin{cases} -u & \text{if } -1 \leq u \leq 0, \\ u^2 & \text{if } \quad 0 \leq u \leq 1. \end{cases} \tag{4}$$

In this case the integral can be written as $E[f(U)]$ where U is a Uniform$[-1, 1]$ random variable. Defining the two strata $[-1, 0]$ and $[0, 1]$ we have that $\sigma_1^2 = 1/12$ and $\sigma_2^2 = 4/45$ in which case $q_1^\star = 0.49$ and $q_2^\star = 0.51$.

36 Importance Sampling

Importance sampling is a powerful technique for reducing estimator variance. The basic setup is as follows. Suppose that the unknown quantity μ to be estimated may be written as the expectation of a function h of some random variable X. That is, assuming that X has a pdf f, one has

$$\mu \;=\; E[h(X)] \;=\; \int_{-\infty}^{\infty} h(x) f(x) dx.$$

Hence, one can estimate μ by first generating $X_1, X_2, ..., X_N$ according to the distribution of X and then calculating the sample average

$$\frac{1}{N}\sum_{n=1}^{N} h(X_n).$$

The variance of $h(X)$ might be large making implementation of the algorithm above inefficient. Importance sampling attempts to reduce variation by selecting an alternative distribution G and

expressing μ as the expectation of a different function of the random variable Y which has distribution G . Specifically, assuming that G has a pdf g, note that

$$\int_{-\infty}^{\infty} h(x)f(x)dx \quad = \quad \int_{-\infty}^{\infty} h(y)\,\frac{f(y)}{g(y)}g(y)dy.$$

So, it also follows that

$$E\left[h(Y)\,\frac{f(Y)}{g(Y)}\right] \quad = \quad \int_{-\infty}^{\infty} h(y)\,\frac{f(y)}{g(y)}g(y)dy \quad = \quad \mu.$$

This suggests the following alternative Monte Carlo simulation approach for estimating μ. First generate $Y_1, Y_2, ..., Y_N$ according to the alternative distribution G and set

$$Z_n = h(Y_n)\cdot\frac{f(Y_n)}{g(Y_n)}, \text{ for } n = 1, 2, ..., N.$$

The sample average $\bar{Z}$ is then an ubiased estimate of μ. The function f/g is referred to as the *likelihood function*.

In order for importance sampling to work, it must be the case that if $g(y) = 0$, then $h(y)f(y) = 0$, otherwise the integral above will be infinity. Also, in some cases the variance of the importance sampling estimator is larger than the variance of the ordinary Monte Carlo estimaor. One must therefore be careful when selecting an alternative distribution G. This is discussed further below.

Example. Consider the problem of estimating $p = P(N(0,1) > 4)$ where $N(0,1)$ is a standard normal random variable. This can be accomplished using ordinary Monte Carlo by first generating a sequence $X_1, X_2, ..., X_N$ of independent and identically distributed $N(0,1)$ random variables, and then setting $Z_n = h(X_n)$ where

$$h(x) = \begin{cases} 1 & \text{if } x > 4, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

and computing the sample average $\bar{Z}$. The standard deviation of each Z_n in this case is $\sqrt{p(1-p)}$ which is very large compared to p the quantity being estimated.

Now let G be an alternative distribution with pdf

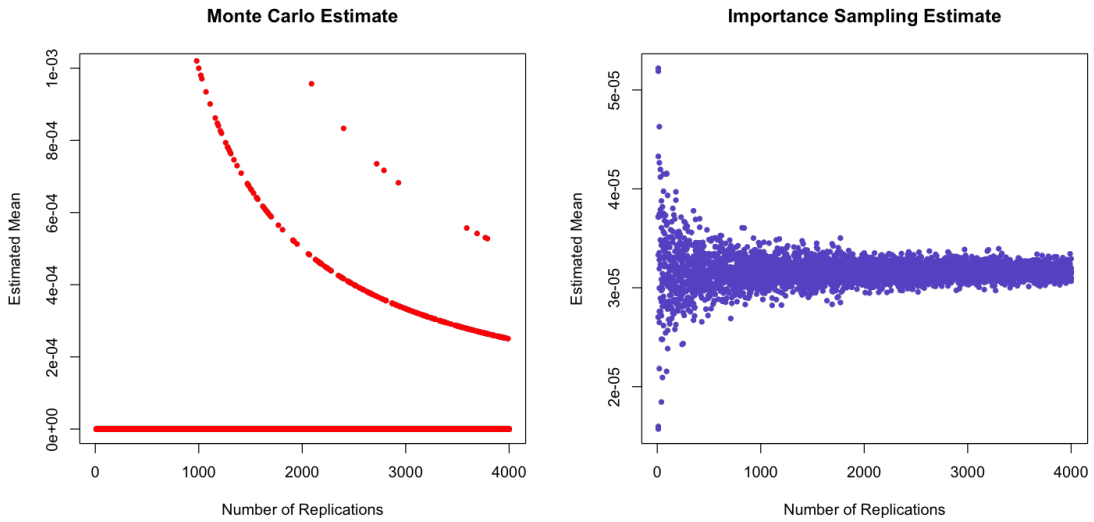
$$g(y) = \begin{cases} e^{-(y-4)} & \text{if } y > 4, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Then, letting $Y_1, Y_2, ..., Y_N$ be a sequence of independent and identically distributed $N(0,1)$ random variables, and setting

$$Z_n = h(Y_n)\cdot\frac{f(Y_n)}{g(Y_n)}$$

where f is the pdf of a standard normal random variable, we have that $\bar{Z}$ is an unbiased estimate of p.

The following R code plots both the ordinary Monte Carlo estimate (red dots) and the importance sampling estimate (blue dots) for $P(X > 4)$. In many cases because $P(X > 4)$ is so small the Monte Carlo estimate is zero.



Example. Consider the problem of computing the integral

$$\int_0^1 x^{-1/2} e^{-x} dx.$$

This can be accomplished using ordinary Monte Carlo by first generating a sequence $X_1, X_2, ..., X_N$ of independent and identically distributed $U[0,1]$ random variables, and then setting $Z_n = h(X_n)$ where $h(x) = x^{-1/2}e^{-x}$ and computing the sample average $\bar{Z}$. Unfortunately, this will result in an estimator with infinite variance.

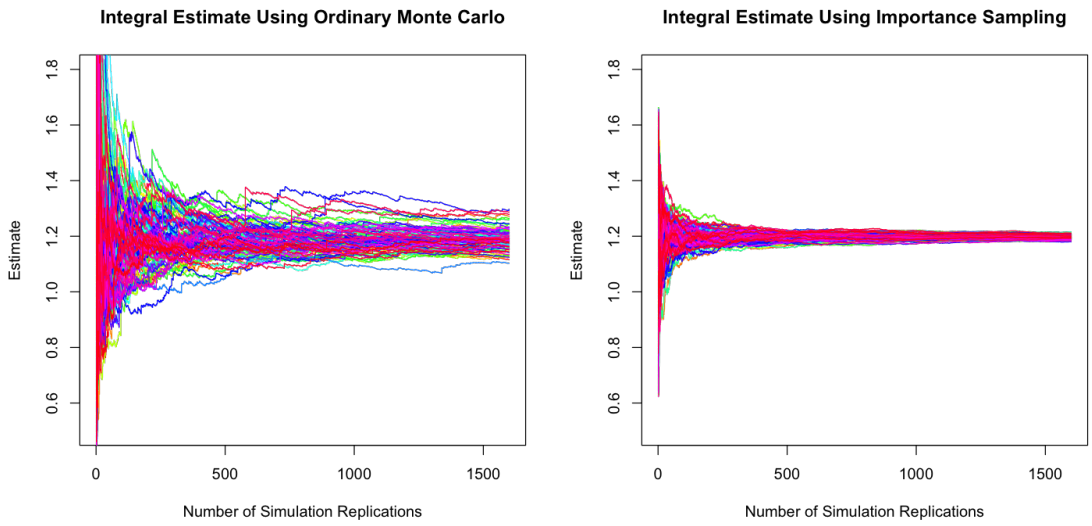
Suppose instead that we use importance sampling with an alternative distribution G which has pdf $g(x) = 2/\sqrt{x}$ on $[0,1]$. In this case

$$h(y)\frac{f(y)}{g(y)} = 2e^{-y}$$

and so generating $Y_1, Y_2, ..., Y_N$ according to the alternative distribution G and then setting

$$Z_n = 2e^{-Y_n}, \quad \text{for } n = 1, ..., N,$$

the sample average $\bar{Z}$ may also be used as an unbiased estimator for the value of the integral.



37 An Application to Option Pricing

Recall that in the Black-Scholes framework the price of a European call option with a strike price of K expiring at time T is given by $E[Z]$ where

$$Z = e^{-rT} \max(S(T) - K, 0)$$

is the discounted payoff of the option at expiration. We now apply importance sampling to estimate $E[Z]$. In order to do so recall that

$$S(T) = S(0) \exp\left(\left(r - \frac{\sigma^2}{2}\right) T + N(0, \sigma^2 T)\right).$$

Then, letting

$$h(x) \quad = \quad e^{-rT} \max\left(S(0) \exp\left(\left(r - \frac{\sigma^2}{2}\right) T + x\right) - K, 0\right),$$

it follows that

$$E[e^{-rT} \max(S(T) - K, 0)] \quad = \quad \int_{-\infty}^{\infty} h(x) f(x) dx,$$

where f is the pdf of a $N(0, \sigma^2 T)$ random variable. This fits into the importance sampling framework described above.

Now replace the distribution f of the $N(0, \sigma^2 T)$ random variable above with an alternative normal random variable with the same variance but a different mean μ. The likelihood ratio is then given by

$$\frac{f(y)}{g(y)} = \frac{e^{-y^2/(2\sigma^2 T)}}{e^{-(y-\mu)^2/(2\sigma^2 T)}}$$

and the price of the option is

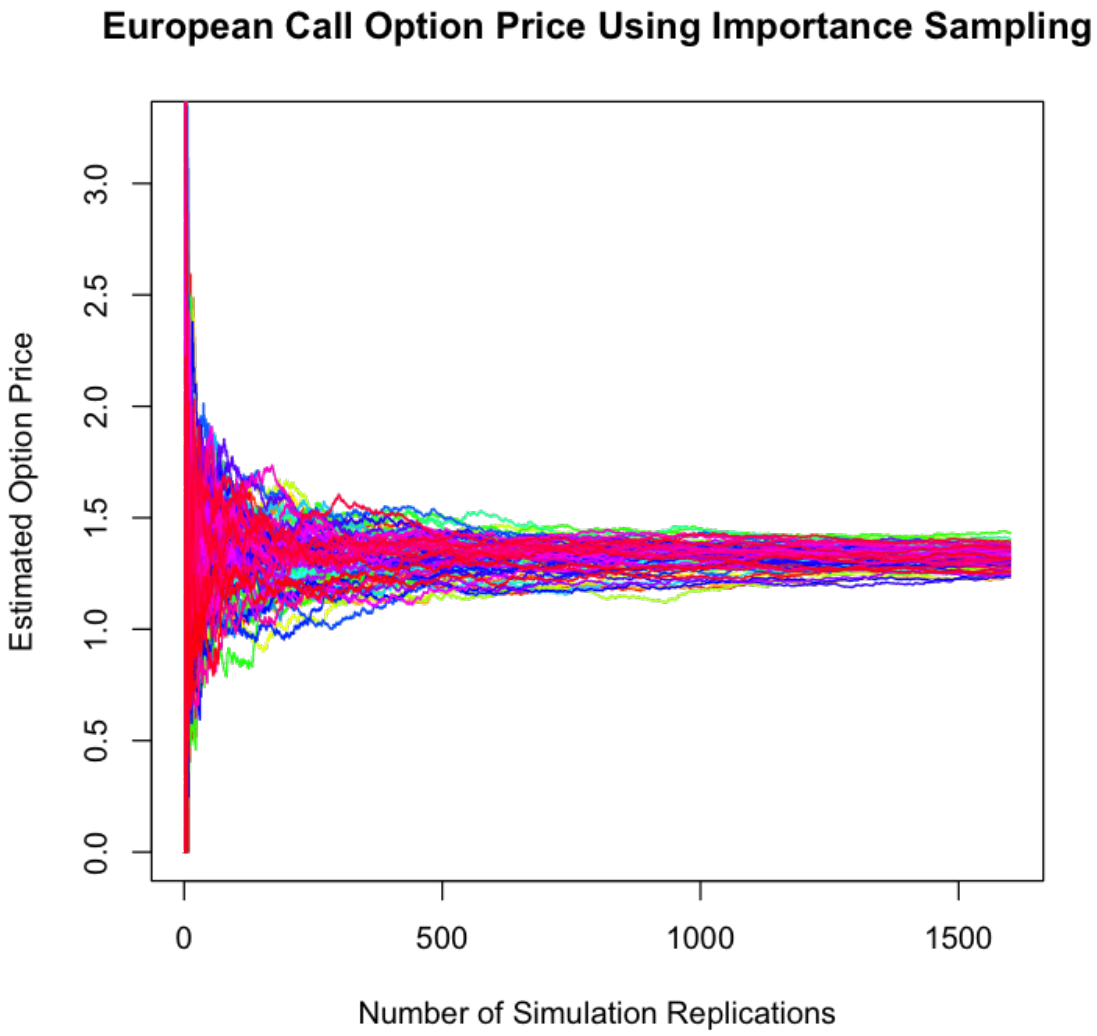
$$E[Z] \quad = \quad E\left[h(Y)\,\frac{f(Y)}{g(Y)}\right].$$

Example. Consider a European call option with a strike price of $K = \$45$ where the initial price of the underlying stock is $S(0) = \$40$. Also assume that the risk free interest rate is $r = 2\%$, the volatility is $\sigma = 30\%$ and the expiration date is $T = 1/2$ year.

Using the formula above for the underlying stock price, the expected value of the underlying stock at the expiration date of $T = 1/2$ years is \$40.40. This is significantly less than the strike price of \$45. The probability that the option is executed at expiration is

$$P(S(1/2) > 45) \quad = \quad 0.27.$$

We can use importance sampling to increase the percentage of the time that the option pays off. Suppose that our goal is to change the distribution of the underlying stock so that its expected value at expiration is equal to the strike price of \$45. In order to do this, we can replace the distribution of the $N(0, \sigma^2 T)$ random variable above with an alternative normal random variable with the same variance but a mean of $\mu = 0.11$.



37.1 Output Analysis for Importance Sampling

Importance sampling does not always reduce variance. Sometimes it even significantly increases variance. This makes properly choosing an alternative distribution g more important. The following discussion shows how to calculate the variance of an importance sampling estimator.

Suppose as above that in order to estimate $\mu = E[h(X)]$ where X has a pdf f, a sequence $Y_1, Y_2, ..., Y_N$ of random variables with alternative distribution G has been generated and we set

$$Z_n = h(Y_n) \cdot \frac{f(Y_n)}{g(Y_n)}, \text{ for } n = 1, 2, ..., N.$$

The estimator is then given by the sample mean $\bar{Z}$. In this case, the variance of each Z_n is

$$\sigma^2 \quad = \quad \int_{-\infty}^{\infty} \frac{h^2(x)f(x)}{g(x)} f(x)dx - \mu^2.$$

Because μ is unknown, σ must be estimated from the simulation output. In particular, we can use the sample standard deviation

$$s \quad = \quad \frac{1}{N-1} \sum_{n=1}^{N} \left(h(Y_n) \, \frac{f(Y_n)}{g(Y_n)} - \bar{\mu} \right)^2.$$

Then, a 100(1-α)% confidence interval for μ is given by

$$\left[\bar{\mu} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{\mu} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right].$$

The variance of $h(X)$ is given by

$$\int_{-\infty}^{\infty} h^2(x)f(x)dx - \mu^2.$$

Comparing the variances of the estimators with and without importance sampling, the condition for importance sampling to result in a reduction of variance is

$$\int_{-\infty}^{\infty} h^2(x) \left(1 - \frac{f(x)}{g(x)} \right) f(x)dx \quad > \quad 0.$$

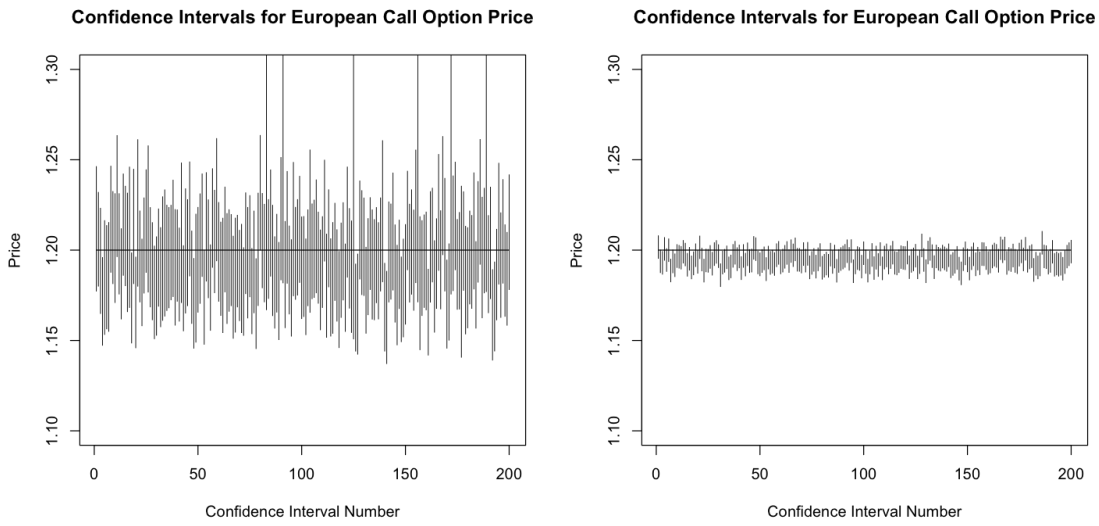
Hence, in order for the alternative sampling distribution $g(x)$ to provide a variance reduction it must roughly speaking be greater than $f(x)$ when $h^2(x)f(x)$ is large, and be less than $f(x)$ when $h^2(x)f(x)$ is small.

One way to choose an alternative distribution g is to note that if $g(x) = h(x)f(x)/\mu$, then the variance of the importance sampling estimator will be zero. Usually the quantity μ is not known and so this cannot be used as an alternative distribution but it still serves as a guide for selecting one. In particular, one rule of thumb is that the alternative distribution $g(x)$ should be proportional or similar to $h(x)f(x)$. One way to achieve this is to select $g(x)$ so that it obtains its maximum at the same x as does $h(x)f(x)$.

Example. Consider again the problem of computing the integral

$$\int_0^1 x^{-1/2} e^{-x} dx.$$

Using the importance sampling technique discussed above, the R code below plots 200 instances of both the ordinary Monte Carlo and the importance sampling 95% condifdence intervals for the integral value. Each instance is constructed from 10,000 simulations.



37.2 Exponential Tilting

The technique used in the application to option pricing to change the mean of a normal random variable is an example of a general method referred to as *exponential tilting*. The method works as follows. Suppose as usual in the importance sampling framework that X is a random variable with a a pdf of f and let

$$\psi(\theta) \;=\; \ln \int_{-\infty}^{\infty} e^{\theta x} f(x) dx, \;\; \theta \in \mathbb{R}.$$

The function ψ is referred to as the *cumulant generating function* of X. It is not always defined for all θ but for convenience lets assume that it is. Now set the pdf of the alternative distribution G to be

$$g(y) \;=\; e^{\theta y - \psi(\theta)} f(y), \;\; y \in \mathbb{R}.$$

Then, the likelihood function is given by

$$\frac{f(y)}{g(y)} \;=\; e^{-\theta y + \psi(\theta)}, \;\; y \in \mathbb{R}.$$

Also, the mean of a random variable Y with the distribution of g is

$$E[Y] \;=\; \int_{-\infty}^{\infty} y g(y) dy \;=\; \int_{-\infty}^{\infty} x \, e^{\theta x - \psi(\theta)} f(x) dx \;=\; \psi^{'}(\theta).$$

So, the expectation of Y is given by the derivative of ψ evaluated at θ.

37.3 Comparison of Variance Reduction Methods

Each of the variance reduction methods discussed have advantages and disadvantages. One way to compare the methods is based upon the simplicity of their implementation versus their effectiveness. The easier the method is to implement the smaller the variance reduction tends to me. Out of the 4 methods explained, antithetic variables is the easiest to use. It requires little knowledge of the context in which the simulation is run but lacks the power of more sophisticated methods. Control variates is the second easiest method to use. Selecting a good control variate requires estimates of correlation but its implementation only requires knowledge the mean of the control variate. Stratified sampling requires the ability to partition the distribution and tends to deliver greater variance reduction. Importance sampling is the most powerful of the 4 methods. It has the potential to reduce variance by levels not achievable by the other 3 methods. However, selecting a good alternative distribution g is not always easy and can even increase variance.

The following graph provides boxplots of the Monte Carlo estimates of the call option price in our running example using ordinary Monte Carlo and each of the 4 variance reduction techniques. Each boxplot is based off of 10,000 simulation replications.

38 Session 7: Poisson Processes

Summer 2019 - Instructor: Josh Reed

Teaching Assistant: Haotian Song

In this session, a brief introduction to stochastic processes is given which is followed by the study of Poisson processes.

39 Introduction to Stochastic Processes

The word "stochastic" comes from the Greek word "stokhastikos" which loosely speaking translates to "able to guess". The English definition for "stochastic" is "random" or "involving chance or probability". Technically speaking, a stochastic process X is a collection of random variables. The collection is usually indexed either according to the integers $n = 0, 1, 2, ...$, in which case the process is referred to as a *discrete time* process, or it can be indexed by time $t \geq 0$, in which case the process is referred to as a *continuous time* process. In most cases, a discrete time process is denoted by

$$X = \{X_n, n = 0, 1, 2, ...\},$$

and a continuous time process is denoted by

$$X = \{X_t, t \geq 0\}.$$

The *state space* of a stochastic process specifies the range of possible values that the random variables of the process can take. The two main classifications for the state space of a stochastic process are a *discrete state space* and a *general state space*. A discrete state space is any finite set of values

or possibly an infinite set of values which may be mapped to the integers. A general state space is much more arbitrary and more or less covers anything else. It is important to point out that these state spaces need not necessarily be described by numbers, they could for instance consist of letters, colors, shapes etc. For the most part however, the stochastic processes encountered in this class will take numeric values.

Some examples of stochastic processes are given below.

Example. The graph below tracks the number of regular season wins of the New York Mets baseball team starting from their inaugural season of 1962 until the 2018 season. This is a discrete time stochastic process with a discrete state space. In particular, its state space is the integers 0 through 162 since there are 162 games in a regular season.

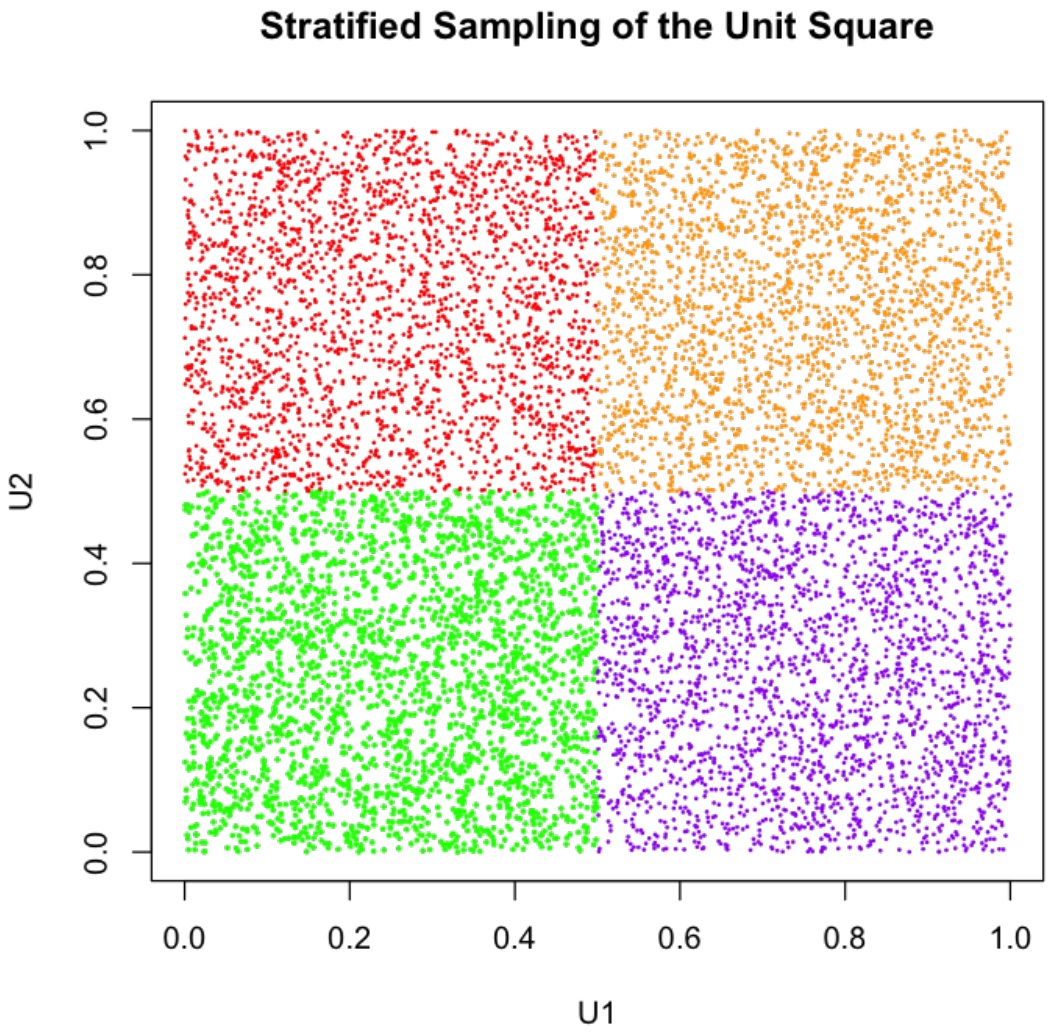
Example. The following is a graph of the high temperature (in degrees Fahrenheit) recorded at Central Park in New York City for each day of January 2019. This is a discrete time stochastic process with a general state space. In particular, its state space is any number greater than -459.67 degrees Fahrenheit (absolute zero).

Example. The following is a graph of 3 stochastic processes. Each is a continuous time stochastic process with a discrete state space. The upper blue line is the asking price for the stock of General Electric trading on New York Stock Exchange over a 1 minute interval on March 21st, 2002. The lower blue line is the bid price for the stock of General Electric over the same interval. The red $+$'s correspond to points in time where transactions occur and their corresponding price. Because activity on the exchange is continuously monitored, this stochastic processes is a continuous time process. Its state space is discrete with the minimum difference between two prices being the tick size.

40 The Bernoulli Process

The Bernoulli process is a simple stochastic process which illustrates several important concepts. It is a discrete time stochastic process $X = \{X_n, n \geq 0\}$ where each X_n is a Bernoulli random variable which is equal to 0 with probability p and equal to 1 with probability $1 - p$. Moreover, the X_n's are assumed to be independent of one another. In order to simulate a Bernoulli process one only needs to simulate n independent and identically distributed Bernoulli random variables.

The following R code simulates a Bernoulli process and graphs its output.

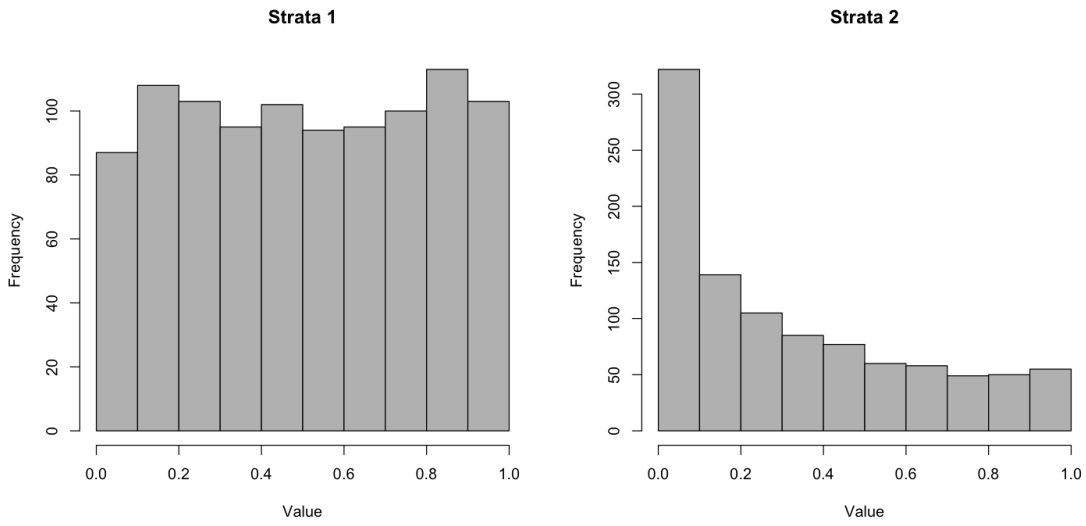


One interpretation of a Bernoulli process is that at each point in time n, the random variable X_n records whether a specific event occurred or not. In particular, the random variable X_n is equal to 1 if the event occurred and 0 if not. As two examples, the event could for be whether a customer arrived to a bank or if a coin which was tossed landed on heads.

In both cases, it natural to also keep track of the total number of events which have occurred up to each point in time. This can be done by noting that for each $n \geq 1$, the quantity

$$S_n = \sum_{k=1}^{n} X_k$$

is the total number of events which have occurred by time n. We assume that $S_0 = 0$. The process $S = \{S_n, n \geq 0\}$ is a new stochastic process. It is zn example of a *counting process* whcih we discuss more below. The R code below graphs S.



Now consider the distribution of the number of events that have occurred by time $n \geq 1$ in the Bernoulli counting process. That is, consider the distribution of the random variable S_n. It turns out that S_n has a binomial distribution with parameters n and p. We denote this type a random variable by Binomial(n, p). In general, a Binomial(n, p) random variable represents the number of times that an event has occurred out of n separate trials, where for each trial the probability of an event occurring is p. A Binomial(n, p) random variable can take the values 0 through n, and the probability of each of these values is given by

$$P(\text{Binomial}(n,p) = k) \quad = \quad \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

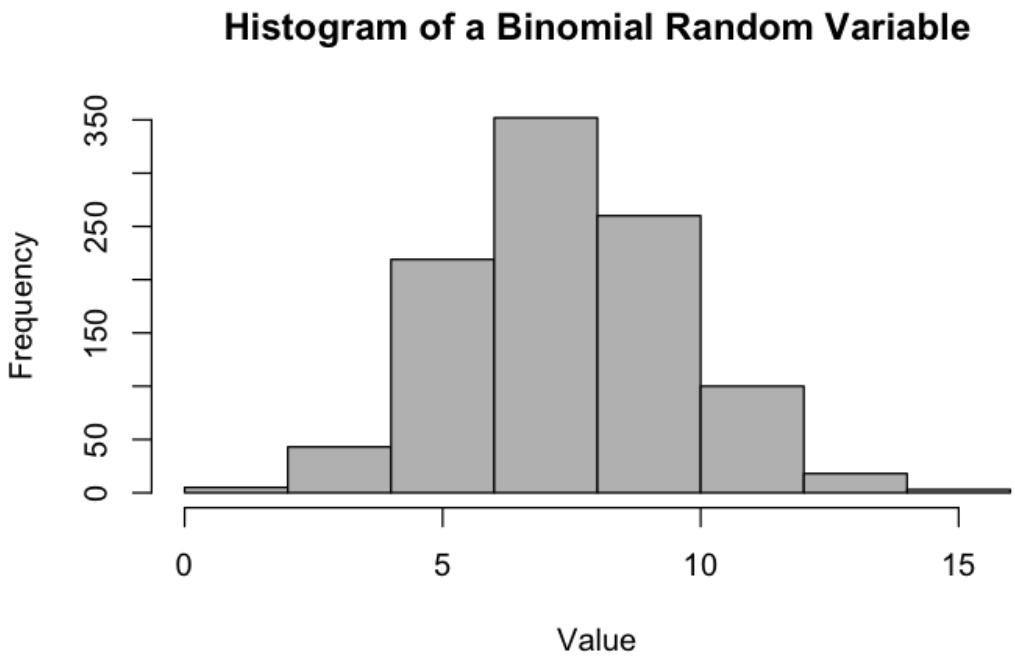
Here $n!$ represents n factorial which is given by

$$n! \quad = \quad n \times (n-1) \times (n-2) \ \times ... \times \ 2 \times 1.$$

Example. Consider the number of times that a six-sided die lands on 1 out of a total of 4 rolls of the die. This is a binomial random variable where the number of trials is $n = 4$ and the probability of an event occurring at each trial is $p = 1/6$. In this case, the probability that 1 out of the 4 rolls of the die is a 1 is given by

$$P(\text{Binomial}(4, 1/6) = 1) \quad = \quad \frac{4!}{3!(1)!}(1/6)(5/6)^3 \ = \ 0.38.$$

One way to simulate a Binomial(n, p) random variable is by summing n Bernoulli random variables, each of which has a probability $1 - p$ of equaling 0 and a probability p of equaling 1.



41 Exponential Random Variables

Before we introduce Poisson processes it is helpful to first discuss exponential random variables. Recall that a non-negative random variable X is said to have an exponential distribution with *rate* $\lambda > 0$ if

$$P(X > x) \;=\; 1 - e^{-\lambda x},\; x \geq 0.$$

The mean of an exponential random variable with rate λ is $1/\lambda$ and the variance is $1/\lambda^2$. It also turns out as we will see shortly that the amount time between when two consecutive events of a Poisson process occur is an exponential random variable.

A random variable X is said to possess the *memoryless property* if

$$P(X > t + s \mid X > t) \;=\; P(X > s) \;=\; e^{-\lambda s},\; s, t \geq 0.$$

In words, the memoryless property states that given that X is greater than t, the probability that X is greater than $t + s$ is equal to the unconditional probability that X is greater than s. It turns out that exponential random variables are the only continuous random variables that possess the memoryless property. Moreover, the memoryless property can be useful in helping to speed up calculations involving exponential random variables.

Example. There are two clerks working at bank and the amount of time that a customer spends speaking with a clerk is an exponential random variable with rate λ. Suppose now that a customer arrives to the bank and finds no line but both clerks are already busy serving customers.

- What is the probability that the newly arrived customer leaves the bank *after* both of the customers already in service leave.

The answer to this problem is $1/2$. The reason is that when the arriving customer enters service, by the memoryless property of the exponential distribution, the remaining service time of the customer still in service is an exponential random variable with rate λ. Hence, since the service time of the customer entering service is also an exponential random variable with rate λ, it follows that the two customers are equally likely to be the next customer to leave.

42 The Poisson Process

The Poisson process is a fundamental continuous time stochastic process. It is named after the French scientist Simeon Poisson who made many contributions to mathematics. The Poisson process is an example of a *counting process*.

Definition. A stochastic process $N = \{N(t), t \geq 0\}$ is said to be a counting proccess if

1. $N(t) \geq 0$ for all $t \geq 0$.
2. $N(t)$ is integer valued for all $t \geq 0$.
3. If $s < t$, then $N(s) \leq N(t)$.
4. For $s < t$, we have that $N(t) - N(s)$ is the number of events occuring in the time interval $(s, t]$.

Counting processes are very useful due to their many potential real life applications such as modeling the number of customers calling a telephone center, the number of buy orders submitted for a particular stock or the number of users arriving to website.

One feature of counting processes that tends to make them easier to analyze is if they have *independent increments*.

Definition. A counting process $N = \{N(t), t \geq 0\}$ is said to possess independent increments if the number of events occuring in disjoint intervals of time are independent of one another.

The definition of a Poisson process is as follows.

Definition. A counting process $N = \{N(t), t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if

1. $N(0) = 0$.
2. The process N has independent increments.
3. For $s < t$, the random variable $N(t) - N(s)$ is Poisson distributed with mean $\lambda(t - s)$.

In words, a Poisson process is a counting process whose increments are independent and Poisson distributed. Note that Condition 3 above implies that for a Poisson process over *any* interval of time $(s, t]$ the average number of events that occur is $\lambda(t - s)$ and

$$P(N(t) - N(s) = k) = e^{-\lambda(t-s)} \frac{(\lambda(t-s))^k}{k!}, \;\; k = 0, 1, 2, ...$$

Because the distribution of the number of events that occur over an interval of time only depends on the *length* of the interval, and not its start and end times, Condition 3 implies that a Poisson process has what is referred to as *stationary increments*.

Example. Suppose that customers arrive to a bank according a Poisson process $N = \{N(t), t \geq 0\}$ with a rate of λ equal to 10 customers per hour.

- What is the averge number of customers that arrive to the bank over a 2 hour period of time? The average number of customers that arrive to the bank over any 2 hour period of time is Poisson distributed with a mean of $2 \times \lambda = 2 \times 10 = 20$ customers.
- What is the probability that 5 customers arrive to the bank over a 1 hour period of time? The number of customers that arrive to the bank over a 1 hour period of time is Poisson distributed with a mean of $\lambda = 10$ customers. The probability that 5 customers arrive over a 1 hour period of time is therefore

$$e^{-10} \frac{10^5}{5!}.$$

- What is the probability that 4 customers arrive to the bank over a 1 hour period of time and that 15 customers arrive to the bank over a separate 2 hour period of time? The number of customers that arrive to the bank over a 1 hour period of time is Poisson distributed with a mean of $\lambda = 10$ customers and the number of customers that arrive to the bank over a 2 hour period of time is Poisson distributed with a mean of $2 \times \lambda = 20$ customers. Moreover, since these two periods of time are assumed separate, the number of customers arriving in each period are independent. The probability is therefore given by

$$e^{-10} \frac{10^4}{4!} e^{-20} \frac{20^{15}}{15!}.$$

The amount of time between when two consecutive events occur is referred to as an *interarrival time*. It turns out that in a Poisson process with rate λ, the interevent times are independent of one another and exponentially distributed with rate λ! This is a surprising and useful fact. By the memoryless property of the exponential distribution, it implies that if we observe a Poisson process at a random point in time, the amount of time until the next event is exponentially distributed with rate λ. This explains why the number of events in disjoint intervals are time are independent of one another.

Example. Suppose that starting at midnight visitors arrive to a webpage according to a Poisson process $N = \{N(t), t \geq 0\}$ with a rate of λ equal to 1,000 customers per hour.

- What is the probability that the first visitor does not arrive to the webpage until at least (1/500)th of an hour after midnight?

The amount of time after midnight until the 1st visitor arrives is an exponential random variable with a rate of $\lambda = \$1,000$ customers per hour. Hence, letting X be the number of hours until the first visitor arrives, we have that

$$P(X > 1/500) \;=\; e^{-(1000\cdot(1/500))} \;=\; e^{-2}.$$

- Suppose that the 1st visitor has not arrived by (3/1,000)th of an hour after midnight. What is the probability that they do not arrive by (1/100)th of an hour after midnight?

This can be answered by calculating the probability that X is greater than (1/100)th of an hour given that it is greater than (3/1000)th of an hour. Applying the memoryless property for the exponential distribution, we obtain that

$$P(X > (1/100) \mid X > (3/1000)) \;=\; P(X > (7/1000)) \;=\; e^{-(1000\cdot(7/1000))} \;=\; e^{-7}.$$

- Suppose that the 1st visitor arrives to the website at (6/1,000)th of an hour after midnight. What is the probability that the 2nd visitor arrives to the website by (11/1000)th of an hour after midnight?

This can be answered by calculating the probability that the amount of time between when the 1st and 2nd visitors arrive is (5/1,000)th of an hour or less. Since the amount of time between when visitors arrives is exponentially ditributed with a rate of 1,000 per hour, the probability is given by

$$= \; 1 - e^{-(1000\cdot(5/1000))} \;=\; 1 - e^{-5}.$$

43 Poisson Process Simulation

There are 2 main ways to simulate a Poisson process. The first is to use the fact that the times between consecutive events in a Poisson process with rate λ are independent and exponentially distributed with rate λ. Using this fact, if X_k is the amount of time between the $(k-1)$st and kth event, then the time at which the nth event occurs is given by

$$S_n \;=\; \sum_{k=1}^{n} X_k, \;\; n = 1, 2, ...$$

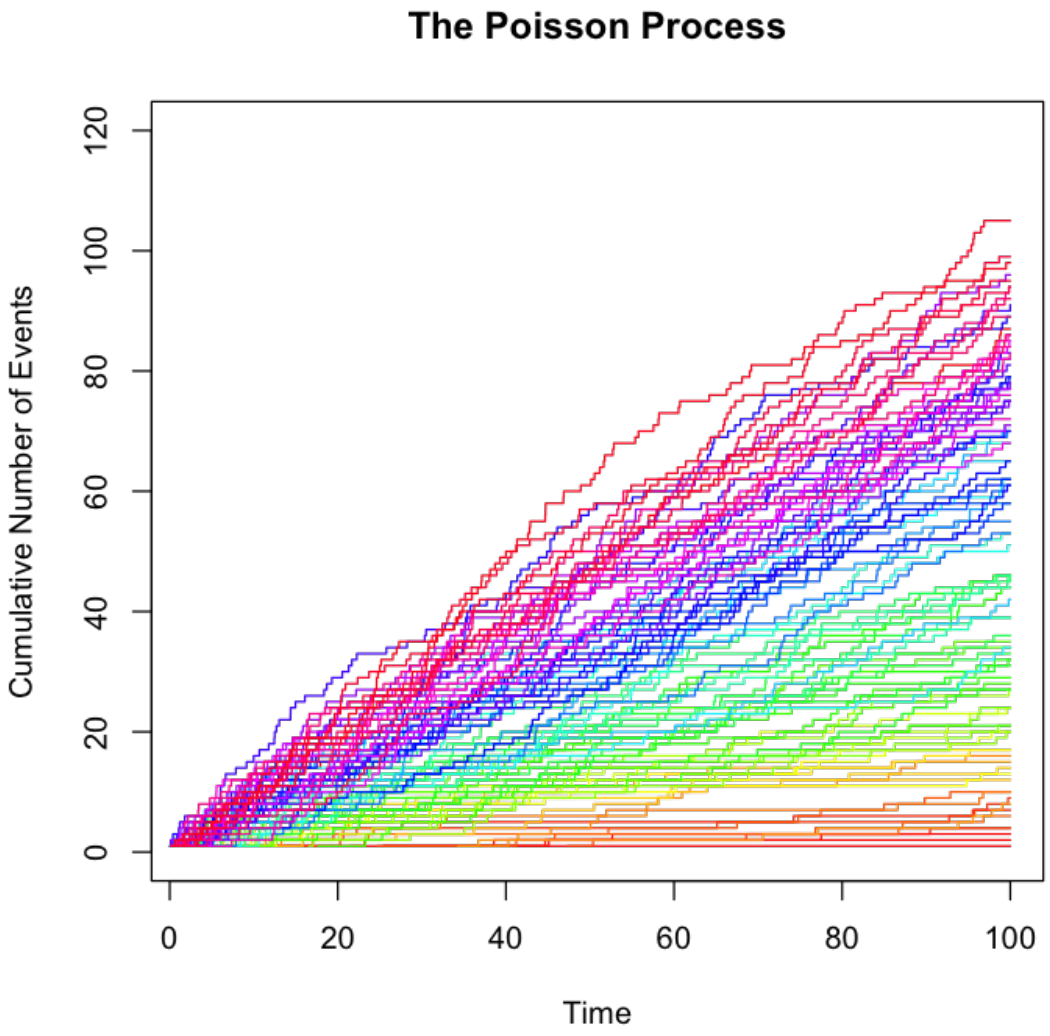
where $X_1, X_2, ...$, are independent and identically distributed with rate λ. Letting $S_0 = 0$, the number of events which have occurred by time $t \geq 0$ is given by

$$N(t) \;\;=\;\; \max\{n \geq 0 : S_n \leq t\}.$$

In practice, one usually only needs to simulate a Poisson process up to some terminal time $T \geq 0$ and so enough X_k's must be generated until an S_n is reached which is greater than T, but nothing more.

The following R code implements a function `PoissonProcessOne`. The function takes as input the values of λ and T and outputs a vector of the ordered times at which the events of a Poisson process with rate λ occur up until the time T using the algorithm alluded to above.

The following R code creates a plot of the cumulative number of events that occur up until time $T = 100$ for 100 independent Poisson processes where the rates of the processes vary from 0.01 to 1 in increments of 0.01.



A second method for simulating a Poisson process up until a terminal time $T \geq 0$ relies on the distributional properties of the times at which events occur up until time T given the total number of events which have occured up until time T. In order to state this distributional result, we first need some notation from statistics.

In order to state the above result in full generality, we first need to introduce the concept of order statistics. Let $\{X_k, 1 \leq k \leq n\}$ be a sequence of random variables. Next, for $k = 1, ..., n$, let $X_{(k)}$ be the kth smallest random variable amongst the sequence $\{X_k, 1 \leq k \leq n\}$. If there are ties, we can pick one of the kth smallest from amongst the $\{X_k, 1 \leq k \leq n\}$ arbitrarily. The random variable $X_{(k)}$ is referred to as the kth-order statistic of $\{X_k, 1 \leq k \leq n\}$. Moreover, the random vector $(X_{(1)}, ..., X_{(n)})$ is referred as the order statistic of the sequence $\{X_k, 1 \leq k \leq n\}$.

Result. Let $N = \{N(t), t \geq 0\}$ be a Poisson process with rate $\lambda > 0$. Given that $N(T) = n$, the distribution of the vector of event times $(S_1, ..., S_n)$ is the same as that of the order statistic of n

independent and identically distributed random variables, each which has a uniform distribution over the interval $[0, T]$.

The above result implies that given that we know that n events have occured by time T in a Poisson process, we can generate the times at which those events occured by simulating n random variables which are uniformly distributed between 0 and T and then arranging them in order.

Example. Suppose that calls arrive to a telephone call center according to a Poisson process with a rate λ of 60 calls an hour.

- Given that 5 calls arrived between 1:00 p.m. and 1:06 p.m., what is the probability that 2 calls arrived betwen 1:00 p.m. and 1:04 p.m.?

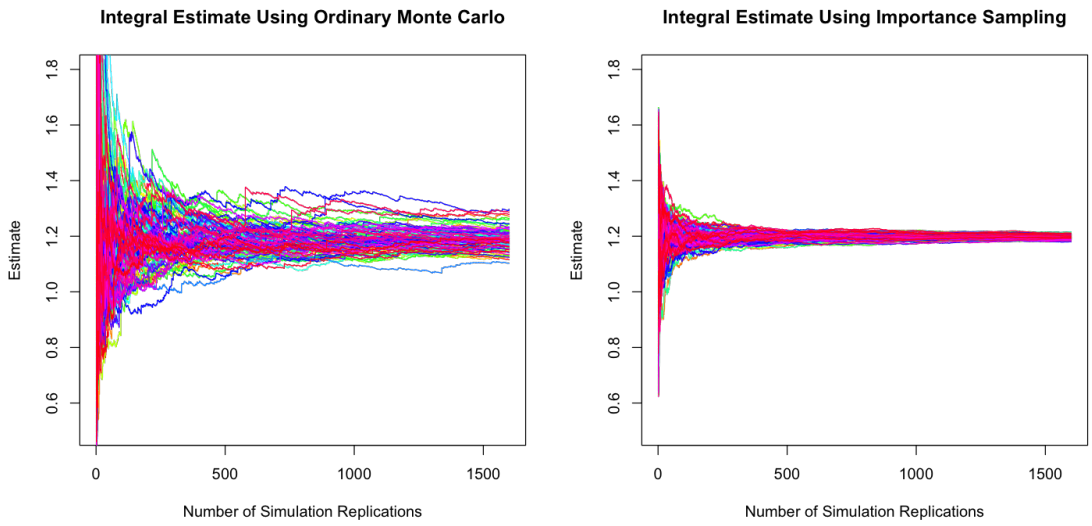
To answer this question we use the result above. Given that 5 calls arrived betwen 1:00 p.m. and 1:06 p.m., the distribution of the ordered times at which the calls arrived is the same as the order statistic of 5 random variables uniformly distrbuted between 1:00 p.m. and 1:06 p.m. Each of these random variables has a $2/3$ probability of being between 1:00 p.m. and 1:04 p.m. Hence, the probability that 2 calls arrived betwen 1:00 p.m. and 1:04 p.m. is the same as the probability of a binomial random variables with 5 trials and probability of success $2/3$ for each trial being equal to 2. That is, the probability is

$$\frac{5!}{2!3!}\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^3.$$

In order to simulate the event times of a Poisson process with rate λ up until a time $T \geq 0$, one can first generate a Poisson random variable with a mean of λT and then, conditional on the outcome of the Poisson random variable, generate the appropriate number of Uniform$[0, T]$ random variables and sort them in order from smallest to largest.

The following R code implements a function `PoissonProcessTwo`. The function takes as input the values of λ and T and outputs a vector of the ordered times at which the events of a Poisson process with rate λ occur up until the time T using the algorithm alluded to above.

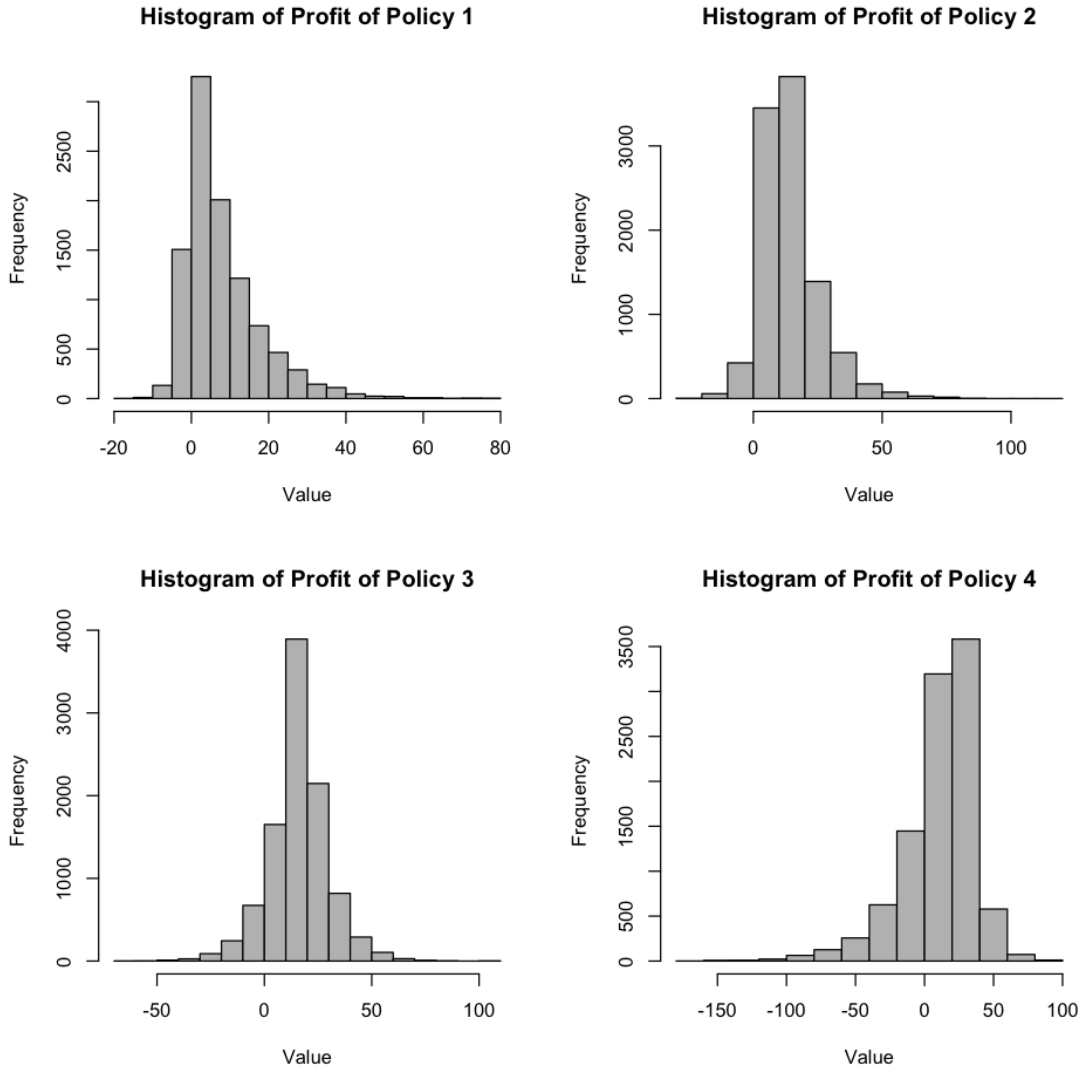
The following R code creates a plot of the cumulative number of events that occur up until time $T = 100$ for 100 independent Poisson processes where the rates of the processes vary from 0.01 to 1 in increments of 0.01.



In-Class Exercise. Suppose that an online store has only 1 unit of inventory of a product and that customers arrive to the store's website according to a Poisson process with a rate λ of 1 customer per minute. Each arriving customer offers a certan amount of money to buy the product and the offers are independent and exponentially distributed with a mean of \$10. For every minute of time that the product is not sold, the store incurs a \$2 holding cost or fractions thereof. The store is considering various policies to sell the product in order to maximize their expected profit. The policies are as follows.

- Accept the offer of the first customer that arrives.
- Accept the first offer that is at least \$9.
- Accept the first offer that is at least \$16.
- Accept the first offer that is at least \$25.

Write a simulation to determine the expected profit for each of the 4 policies listed above. Which one is the best?



The average profits of each of the 4 policies are about \$7.94, \$14.10, \$16.30 and \$10.66, respectively. So, it turns out that the 3rd policy is the best out of the 4.

44 Spatial Poisson Processes

A Poisson process in more than 1 dimension is referred to as a *spatial Poisson process*. The reason for moving to higher dimensions is that not only are the times at which events occur part of the process but their location is as well. This can be useful in applications related to transportation, weather, telecommuncations and more.

We focus here on homogeneous 2-dimensional spatial Poisson processes and assume that the location of each event lies somewhere in the unit square $[0, 1] \times [0, 1]$. The process counting the total number of events that occur in the square is a Poisson process with rate λ. The quantity λ is also

referred to as the *intensity* of the proces. For any subset of the square, the process counting the nunmber of events occuring in that subset is also a Poisson process but with a rate of λ multiplied by the area of the subset. If two subsets do not overlap, then their Poisson processes are independent of each other.

Example. Suppose that the location of crimes in a city occurs according to a spatial Poisson processes with an intensity of 8 crimes per day and that the city is 1 mile long by 1 mile wide.

- Calculate the probability that over 2 days, 5 crimes occur in a neighborhood of the city with an area of $1/4$ square mile.

The counting process of crimes that occur in a neighborhood of the city with an area of $1/4$ square mile is a Poisson processes with a rate of $8(1/4) = 2$ crimes per day. Therefore number of crimes that occur over 2 days is then Poisson distributed with a mean of $2 \times 2 = 4$ crimes and so the probability is

$$e^{-4}\frac{4^5}{5!}.$$

- Calculate the probability that over 2 days 5 crimes occur in each of two separate neighborhoods of the city both with an area of $1/4$ square mile.

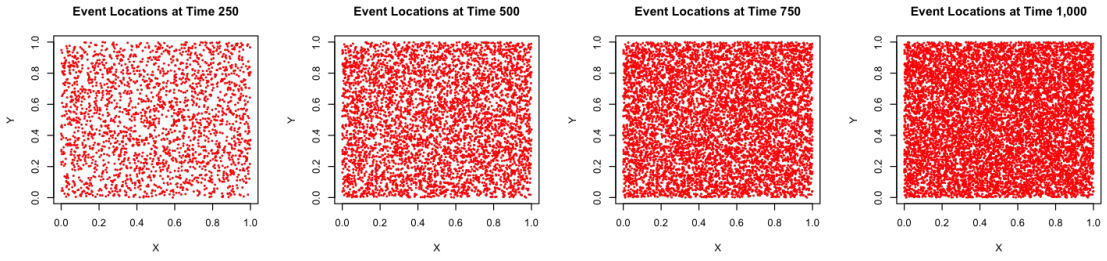
Since the 2 neighborhoods are asssumed to be separate, using the answer above we have that the probability is

$$\left(e^{-4}\frac{4^5}{5!}\right)^2.$$

The following result is useful for simulating spatial Poisson processes.

Theorem. If N is a homogeneous Poisson process with intesnity λ, then the event locations are idependent and identically distribution uniform random variables in the unit square.

The theorem above provides an algorithm for simulating a nonhomogeneous Poisson process with intensity λ. First simulate a Poisson process with rate λ and then assign to each event a location which is uniformly distributed on the unit square. The R code below implements this algorithm. The function `SpatialPoissonUnitSquare` simulates a spatial Poisson process in the unit square. The function takes as input the values of λ and T and outputs a list of 2 vectors. The first vector provides the event times and the second vector the location of each event.



Given a point in the unit square, one interesting quantity to consider is the distance to the nearest event of a spatial Poisson process that has been run up until some terminal time T. This quantity is referred to as the *contact distance*. It turns out that if we ignore boundary effects, the distribution of the contact distance can be computed exactly. Suppose that u is a point in the unit square. Then, there are no events within a distance of r from it if the circle of radius r centered at u has no events in it. But the number of events that occur within a circle of radius r is Poisson distributed with mean $(\lambda T)\pi r^2$. So, we have that

$$P(\text{contact distance is greater than } r) = \exp\left(-\lambda T \pi r^2\right), \;\; r > 0.$$

In-Class Exercise. In an area of a city 1 mile long by 1 mile wide, the locations at which passengers request to be picked up by cars in an online ride hailing service follows an independent spatial Poisson process with a rate of 9 customers per minute. At the same time, the locations at which the online ride hailing service's cars become available folllows a spatial Poisson process with a rate of 10 cars per minute, independent of the passenger arrivals.

Every 30 seconds, the service matches passengers with cars in the following way. The passenger who arrived earliest is matched with the car that is closest to them. Next, the passenger who arrived second earliest is matched out of the remaining cars with the car closest to them, and so on and so forth. Any cars or passengers leftover at the end of this process drop out of the system and the process starts afresh for the next 30 second time interval.

Use simulation to estimate the following for a random 30 second interval.

- The precentage of passengers who are not matched with a car.
- For those passengers who are matched with a car, the average euclidian distance between their location and the location of the car they are matched with.

```
[1] "The average distance between a match in a 30 second interval is: "
```

0.135779320744179

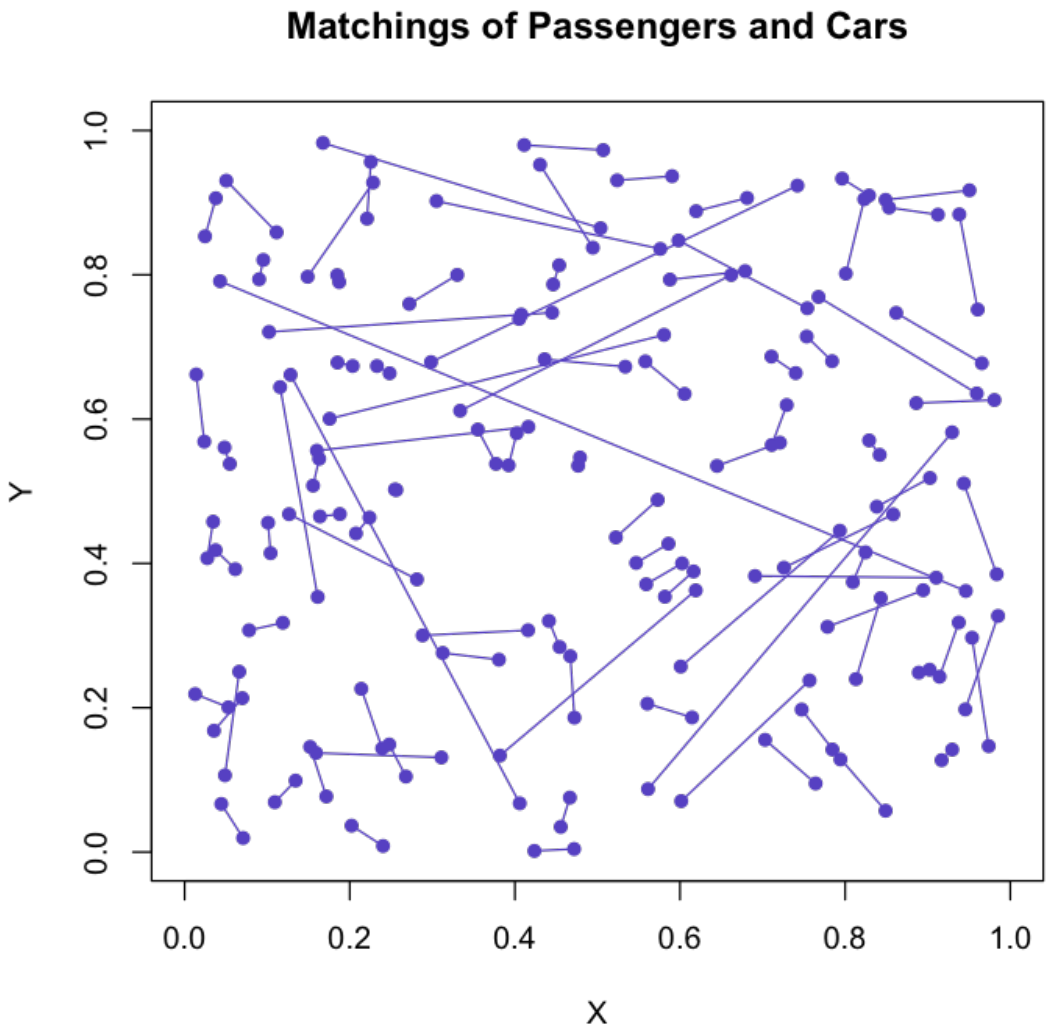
```
[1] "The lower limit for 95% confidence interval for the average distance
between a match in a 30 second interval is: "
```

0.131325031327807

```
[1] "The upper limit for 95% confidence interval for the average distance
between a match in a 30 second interval is: "
```

0.140233610160551

```
[1] "A graph of the matchings from the most recent simulation run is: "
```



45 Session 8: Extensions of Poisson Processes

Summer 2019 - Instructor: Josh Reed

Teaching Assistant: Haotian Song

Recall from the previous session that the definition of a Poisson process is as follows.

Definition. A counting process $N = \{N(t), t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if

1. $N(0)$=0.

2. The process N has independent increments.

3. For $s < t$, the random variable $N(t) - N(s)$ is Poisson distributed with mean $\lambda(t - s)$.

In this session, thinning and superpositions of Poisson processes, compound Poisson processes and non-homogeneous Poisson processes are discussed.

46 Thinned Poisson Processes

In certain situations it is useful to classify each event of a Poisson process as either a Type 1 or a Type 2 event. This can be accomplished by independently labeling each event as either a Type 1 event with probability p or a Type 2 event with probability $1 - p$. Two *thinned* versions of the underlying Poisson process N may then be formed. These processes are denoted by N_1 for the process counting the numbe of Type 1 events, and N_2 for the process counting the number of Type 2 events. The following result characterizes the processs N_1 and N_2 as well as how they are related to one another.

Result. Suppose that N is an underlying Poisson process with rate λ and that each event of N is classified as either a Type 1 or a Type 2 event with probability p or $1 - p$, respectively. Then, the thinned processes N_1 and N_2 are independent Poisson processes with rates λp and $\lambda(1 - p)$, respectively.

The above result is surprising. Not only are the thinned processes N_1 and N_2 Poisson processes, they are also independent of one another! This can be useful when solving problems.

Example (https://www.math.ucdavis.edu/~gravner/MAT135A/resources/lecturenotes.pdf). Suppose that customers arrive to a store according to a Poisson process at a rate $\lambda$of 10 customers per hour. Each arriving customer is male with probability 1/2 and female with probability 1/2. Assume that 10 male customers arrive to the store between 10 am and 11 am.

- Calculate the probability that 10 female customers also arrive between 10 am and 11 am.

According to the result above, the arrival processes of male and female customers are independent Poisson processes each with a rate of $\lambda(1/2) = 5$ customers per hour. The number of female customers that arrive between 10 and 11 am is therefore Poisson distributed with a mean of 5 and is independent of the number of male customers that arrive. The desired probability is then

$$e^{-5}\frac{5^{10}}{10!}.$$

- Calculate the probability that at least 20 total customers enter bewteen 10 am and 11 am.

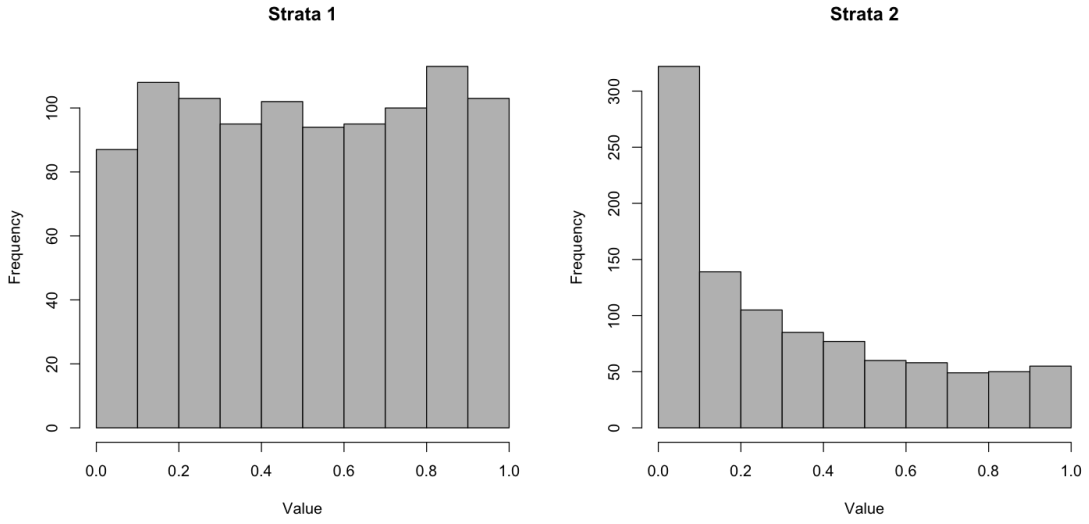
Given that 10 male customers arrive between 10 am and 11 am, at least 20 total customers arrive if 10 or more female customers arrive. The answer is then

$$\sum_{n=10}^{\infty} e^{-5}\frac{5^n}{n!} = 1 - \sum_{n=0}^{9} e^{-5}\frac{5^n}{n!}.$$

The result above provides a method to simulate a thinned Poisson process. Rather than simulating the underlying process N and then thinning it event-by-event, the two thinned Poisson processes N_1 and N_2 can be simulated independently. The following R code implements the function

`PoissonProcessThinned` using this idea. Its input is the arrival rate λ of the underlying Poisson process, the terminal time T and the probabillity p of a Type 1 event, and it returns a list of 3 vectors which provide the event times of the underlying process and the two thinned processes, respectively.

The following R code uses `PoissonProcessThinned` to graph the two thinned processes and the underlying total arrival process over 1 hour of customers to the store in the example above.



Below is a scatterplot of the cumulative number of Type 1 customer arrivals vs. cumulative the number of Type 2 customer arrivals in the example above over the course of 10 hours. The shape of the plot reflects the fact that the numbers of each type of arrivals are independent of one another.

The thinning result above can be extended to an arbitrary number of types of events.

Result. Suppose that N is a Poisson process with rate λ and that each event of N is classified as a Type k event with probability p_k for $k = 1, ..., K$. Then, the thinned processes $N_1, N_2, ..., N_K$ are independent Poisson processes with rates $\lambda p_1, \lambda p_2, ..., \lambda p_K$, respectively.

Example. Suppose that customers arrive to a bank according to a Poisson process with a rate λ of 60 customers per hour. Each arriving customer either uses the ATM with probability $p_1 = 0.4$, closes an account with probability $p_2 = 0.1$, opens a new account with probability $p_3 = 0.2$ or speaks with a mortgage specialist with probability $p_4 = 0.3$.

- Calculate the probability that over the course of one hour: 3 customers arrive to use the ATM, 2 customers arrive to close their account, 4 customers arrive to open a new account, and no customers arrive to speak with a mortage specialist.

According to the result above, the arrival processes of customers to use the ATM, close their account, open a new account, and speak with a mortgage specialist are independent Poisson processes with rates of $24, 6, 12$ and 18 customers per hour, respectively. Moreover, these processes are independent of one another. The desired probability is therefore

$$e^{-24}\frac{24^3}{3!}e^{-6}\frac{6^2}{2!}e^{-12}\frac{12^4}{4!}e^{-18}.$$

- Calculate the probability that over the course of one hour, 2 customers arrive to either open or close their account.

The probability that a customer arrives to either open or close their account is $p_2 + p_3 = 0.1 + 0.2 = 0.3$. Therefore, according to the result above, the arrival process of customers to either open or close their account is a Poisson process with a rate of $(0.3)\lambda = 18$ customers per hour. The desired probability is therefore

$$e^{-18}\frac{18^2}{2!}.$$

Sometimes the probability of an event being classified as a certain type depends on the time at which the event occurs. Suppose for instance that events are classified as either Type 1 or Type 2 and if an event occurs at time $t \geq 0$, the probability that it is a Type 1 event is $p_1(t)$ and the probability that it is a Type 2 event is $p_2(t) = 1 - p_1(t)$. In this case, the thinning processes N_1 and N_2 counting the number of Type 1 and Type 2 events are no longer Poisson processes but we still have the following result.

Result. Suppose that N is a Poisson process with rate λ and that an event of N occurring a time $t \geq 0$ is classified as a Type 1 event with probability $p_1(t)$ or as a Type 2 event with probability $p_2(t) = 1 - p_1(t)$. Then, the number of Type 1 and Type 2 events occuring by some time $T \geq 0$ are independent Poisson random variables with respective means

$$\lambda \int_0^T p_1(t)dt \quad \text{and} \quad \lambda \int_0^T p_2(t)dt.$$

Example. Suppose that customers arrive to a web site according to a Poisson process with a rate of λ. Upon arrival to the site, each customer spends a random amount time with distribution F browsing the site before exiting.

- Assuming that the amounts of time the customers spend browsing the site are independent of another, and that the system is empty at time 0, what is the distribution of the number of customers browsing the site at an arbitrary time $T \geq 0$?

This is an example of an *infinite server queue* and we can answer the question using the result above. Suppose that a customer arrives to the system at some time $0 \leq t \leq T$ and classify the customer as a Type 1 customer if they are still in service at time T. Since the customer will still be in service at time T if and only if their service time is greater than $T - t$, we have the probability that the customer is a Type 1 customer is given by $p_1(t) = 1 - F(T - t)$. It then follows by the result above that the total number of customers in service at time T is Poisson distributed with mean

$$\lambda \int_0^T p_1(t)dt = \lambda \int_0^t (1 - F(T-t))dt.$$

In-Class Exercise. $G/G/\infty$ queue.

47 Superposition of Poisson Processes

Poisson processes can also be combined together. If N_1 and N_2 are separate Poisson processes, then their sum $N = N_1 + N_2$ is referred to as the *superposition* of N_1 and N_2. The following result characterizes the combined process N.

Result. If N_1 and N_2 are independent Poisson proceses with rates λ_1 and λ_2, respectively, then their superposition $N_1 + N_2$ is a Poisson process with rate $\lambda_1 + \lambda_2$.

Example. Two types of subway trains arrive to a subway platform, express trains and local trains. Express trains arrive according to a Poisson proces with a rate of $\lambda_1 = 6$ trains per hour. Local trains arrive according to a Poisson proces with a rate of $\lambda_2 = 8$ trains per hour. The arrival proceses of each type of train are independent of one another.

- Calculate the probability that a total of 16 trains arrive over a 1 hour interval.

Since the arrival processes are independent of one another, the combined process of express and local trains is a Poisson process with a rate of $\lambda_1 + \lambda_2 = 8 + 6 = 14$ trains per hour. The total number of trains that arrive over a 1 hour interval is therefore Poisson distributed with a mean of 14 and so the answer is given by

$$e^{-14}\frac{14^{16}}{16!}.$$

The result above can be extended to the superposition of an arbitrary number of independent Poisson processes.

Result. If $N_1, N_2, ..., N_K$ are independent Poisson proceses with rates $\lambda_1, \lambda_2, ..., \lambda_K$, respectively, then their superposition $N_1 + N_2 + ... + N_K$ is a Poisson process with rate $\lambda_1 + \lambda_2 + ... + \lambda_K$.

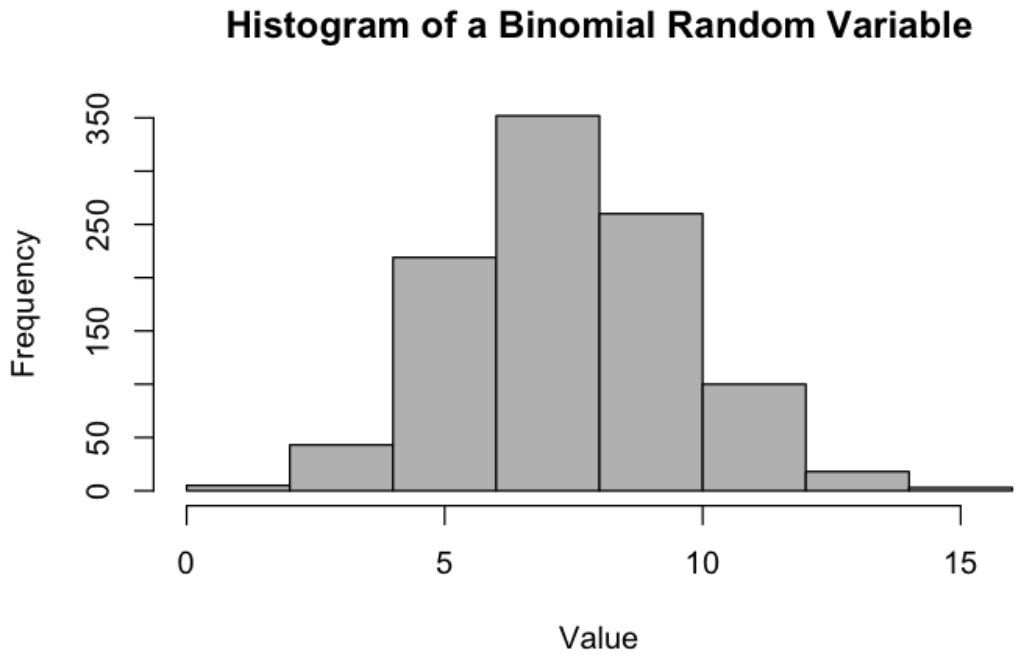
Example. Suppose that visitors arrive to a web page from 100 countries all over the world. From each country, the arrival process is a Poisson proces with a rate λ of 1 visitor per second and the arrival processes from country to country are independent of one another.

- Calculate the probability that no customers arrive over a $1/2$ second interval.

Since there are 100 arrival processes that are independent of one another, and each is a Poisson process with a rate λ of 1 visitor per second, the combined arrival process is a Poisson process with a rate of $\lambda = 100$ visitors per second. The number of visitors that arrive over a $1/2$ second interval is therefore Poisson distribution with a mean of (1/2)100=50 visitors. The desired probability is thus

$$e^{-50}.$$

The R code below generates a random time sampled arrival process to the web page over the course of 1 second. The time sampling occurs every 0.01 seconds. Different countries are assigned different colors in the graph.



In-Class Exercise. Superposition of renewal processes.

48 Nonhomogeneous Poisson Processes

A *nonhomogeneous Poisson process* is a generalization of the Poisson process where the rate at which events occur changes over time. Its definition is as follows.

Definition. A counting process $N = \{N(t), t \geq 0\}$ is said to be a nonhomogeneous Poisson process with rate function $\lambda = \{\lambda(u), u \geq 0\}$ if

1. $N(0)$=0.
2. The process N has independent increments.
3. For $s < t$, the random variable $N(t) - N(s)$ is Poisson distributed with mean

$$\int_s^t \lambda(u)du.$$

Note that unlike an ordinary Poisson process, a nonhomogeneous Poisson process does not have stationary increments. Events are more likley to occur over intervals of time where the intensity function λ is high. This is useful for modeling time varying arrival processes.

Example. (Ross Probabilty Models Example 5.24) Siegbert runs a hot dog stand that opens at 8 a.m. From 8 a.m. until 11 a.m. customers arrive at a steadily increasing rate that starts with an initial rate of 5 customers per hour at 8 a.m. and reaches a maximum rate of 20 customers per hour at 11 a.m. From 11 a.m. until 1 p.m. the rate remains constant at 20 customers per hour. However, the arrival rate then drops stedily from 1 p.m. until closing time at 5 p.m. at which time it has the value of 12 customers per hour.

- If we assume that the numbers of customers arriving at Siegbert's stand during disjoint time periods are independent, then what is a good probability model for the preceding?

A good model to use for this example would be a nonhomogeneous Poisson process with rate function $\lambda = \{\lambda(u), 0 \leq u \leq 9\}$, where

$$\lambda(u) = \begin{cases} 5 + 5u, & \text{for } 0 \leq u < 3, \\ 20, & \text{for } 3 \leq u < 5, \\ 20 - 2(u-5), & \text{for } 5 \leq u \leq 9. \end{cases}$$

- What is the probability that no customers arrive between 8:30 a.m. and 9:30 a.m. on Monday morning?

The number of customers that arrive between 8:30 a.m. and 9:30 a.m. is Poisson distributed with a mean of

$$\int_{0.5}^{1.5} \lambda(u)du = 10.$$

Hence, the probability that no customers arrive during this interval of time is e^{-10}.

- What is the expected number of arrivals in this period?

The expected number of arrivals between 8:30 a.m. and 9:30 a.m. is 10 customers.

There are at least two ways to simulate a nonhomogeneous Poisson process. The first is to thin an ordinary Poisson process in appropriate manner. This method is based on the following result.

Result. Suppose that N is a Poisson process with rate λ and that an event occuring at time t is classified as a Type 1 event with probability $p_1(t)$. Then, the process N_1 counting the number of of Type 1 events is a nonhomogeneous Poisson process with rate function $\lambda = \{\lambda p(t), t \geq 0\}$.

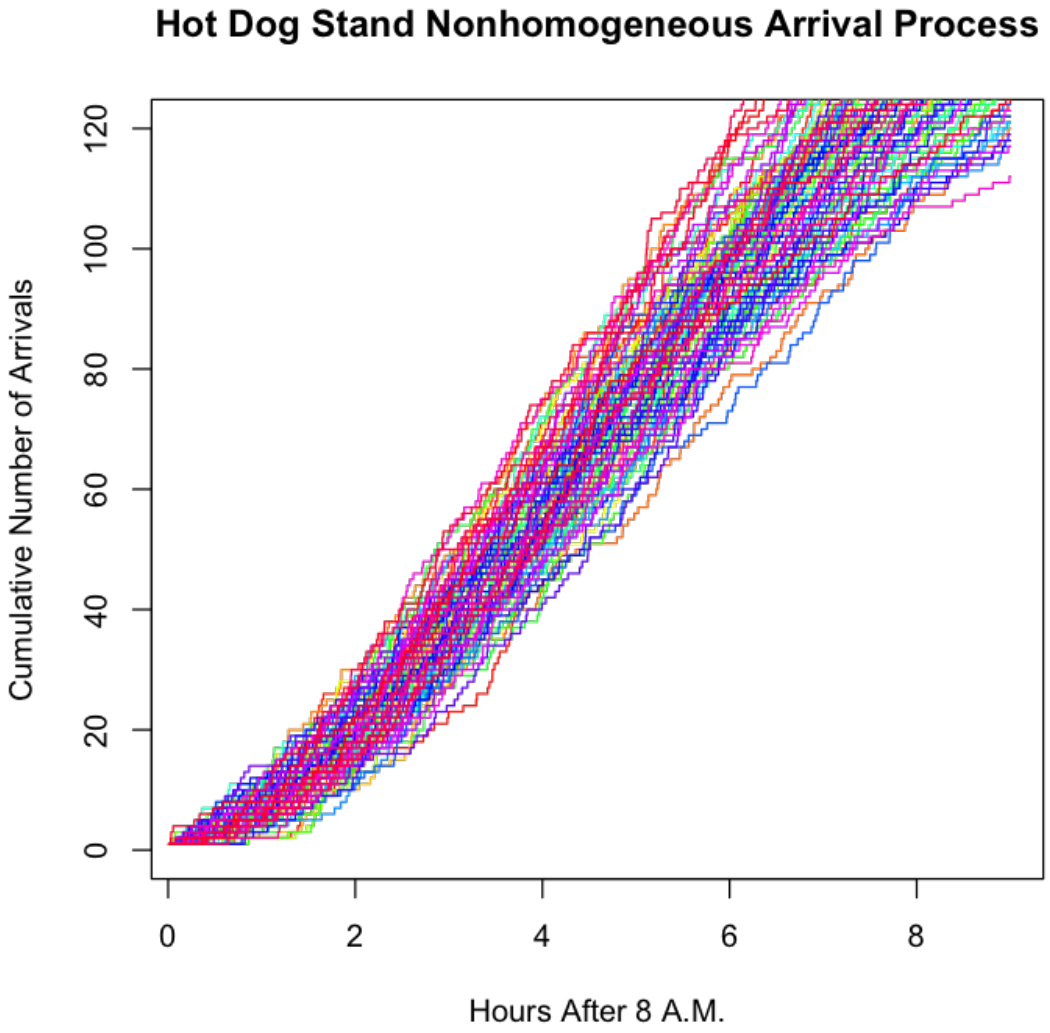
In order to use the above result to simulate a nonhomogeneous Poisson process suppose that N is a nonhomogeneous Poisson process with rate function λ and let λ_{max} be such that $\lambda(t) \leq \lambda_{max}$ for all t. Then, setting

$$p_1(t) = \frac{\lambda(t)}{\lambda_{max}}, \;\; t \geq 0,$$

the above result implies that the thinned process N_1 is a nonhomogeneous Poisson process with rate function λ.

The following R code uses this idea to implement the function `NonhomogeneousOne` which simulates the nonhomogeneous arrival process of the hot dog stand example. The value of λ_{max} is

chosen to be 20 which is the maximum arrival rate to the stand. The function does not take any input and it returns the event times of the arrival process.



A second way to simulate a nonhomogeneous Poisson process up until a time $T \geq 0$ relies on the following result for the distribution of the arrival times given the total number of arrivals up until the time T.

Result. Let $N = \{N(t), t \geq 0\}$ be a nonhomogeneous Poisson process with rate function $\lambda = \{\lambda(u), u \geq 0\}$. Given that $N(T) = n$, the distribution of the vector of event times $(S_1, ..., S_n)$ is the same as that of the order statistic of n independent and identically distributed random variables, each which have the distribution function

$$F(t) = \begin{cases} \frac{m(t)}{m(T)}, & \text{for } 0 \leq t < T, \\ 1, & \text{for } t \geq T, \end{cases}$$

where

$$m(t) = \int_0^t \lambda(u)du, \;\; t \geq 0.$$

The above result implies that given that we know that n events have occured by time T in a Poisson process, we can generate the times at which those events occured by simulating n random variables which are distributed acording to the distribution above. In order to simulate random variables according to this distribution one can use whichever simulation method is most convenient, e.g. inverse transoform method, accpetance-rejection method etc.

Example. Consider again the example of Siegbert's hot dog stand.

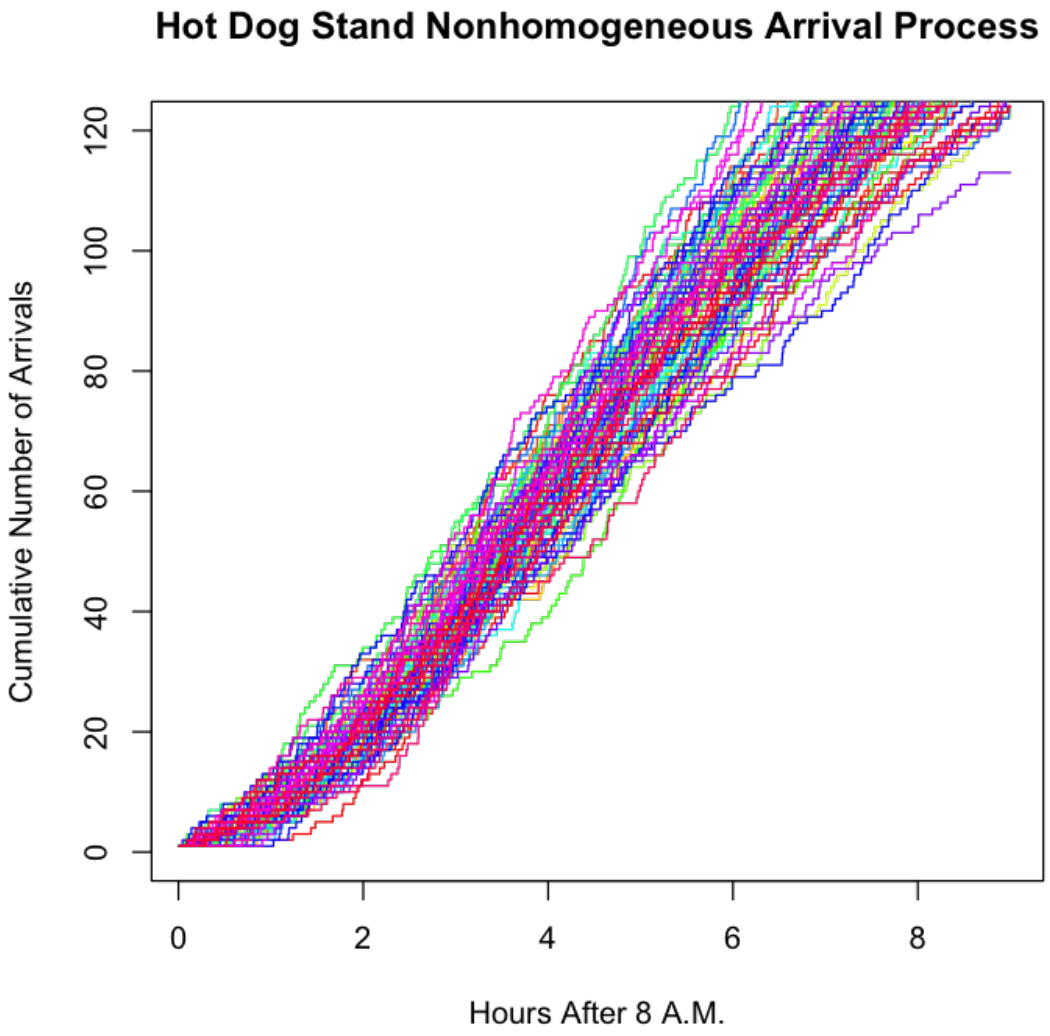
- Given that 50 customers arrived between 8:00 a.m. and 12:00 p.m., what is the probability that 20 customers arrived betwen 8:00 a.m. and 11:00 a.m.?

To answer this question we use the result above. Given that 50 customers arrived betwen 8:00 a.m. and 12:00 p.m., the distribution of the ordered times at which the calls arrived is the same as the order statistic of 50 random variables with distribution function $F(t) = m(t)/m(T)$ for $0 \leq t \leq T$ with $T = 4$. Each of these random variables has a $m(3)/m(4)$ probability of being between 8:00 a.m. and 11:00 a.m. which we can explicitly compute as being equal to 15/23. Hence, the probability that 20 customers arrived betwen 8:00 a.m. and 11:00 p.m. is the same as the probability of a binomial random variables with 50 trials and probability of success 15/23 for each trial being equal to 20. That is, the probability is

$$\frac{50!}{20!30!}\left(\frac{15}{23}\right)^{20}\left(\frac{8}{23}\right)^{30}.$$

In order to simulate the event times of a nonhomogeneous Poisson process with rate function λ up until a time $T \geq 0$, one can first generate a Poisson random variable with a mean of $m(T)$ and then, conditional on the outcome of the Poisson random variable, generate the appropriate number of random variables with distribution function $m(t)/M(T)$ and sort them in order from smallest to largest.

The following R code uses this idea to implement the function `NonhomogeneousTwo` which simulates the nonhomogeneous arrival process of the hot dog stand example. The conditional random variables of the arrival times are generated using the acceptance-rejection method with the uniform distribution between 0 and 9 chosen as the alternative distribution and the value of c set equal to 360/283 The function does not take any input and it returns the event times of the arrival process.



49 Compound Poisson Processes

For modeling purposes it can be helpful for events to be of different sizes. In a *compound Poisson process*, the times at which events occurs is a Poisson process with rate λ but the sizes of each of the events are independent and identically distributed random variables. The compound Poisson process is given by

$$N_X(t) = \sum_{n=1}^{N(t)} X_n, \quad t \geq 0,$$

where N is a Poisson process with rate λ, and $X_1, X_2, ...,$ are independent and identically distributed random variables with distrbution F.

Example. Suppose that groups of customers arrive to a movie theather according to a Poisson process at a rate of 40 groups per hour. The distribution of the size of each group is as follows.

xj	1	2	3	4	5
P(X=xj)	0.10	0.25	0.50	0.10	0.05

- What is the average number of *customers* that arrive to the movie theater each hour?

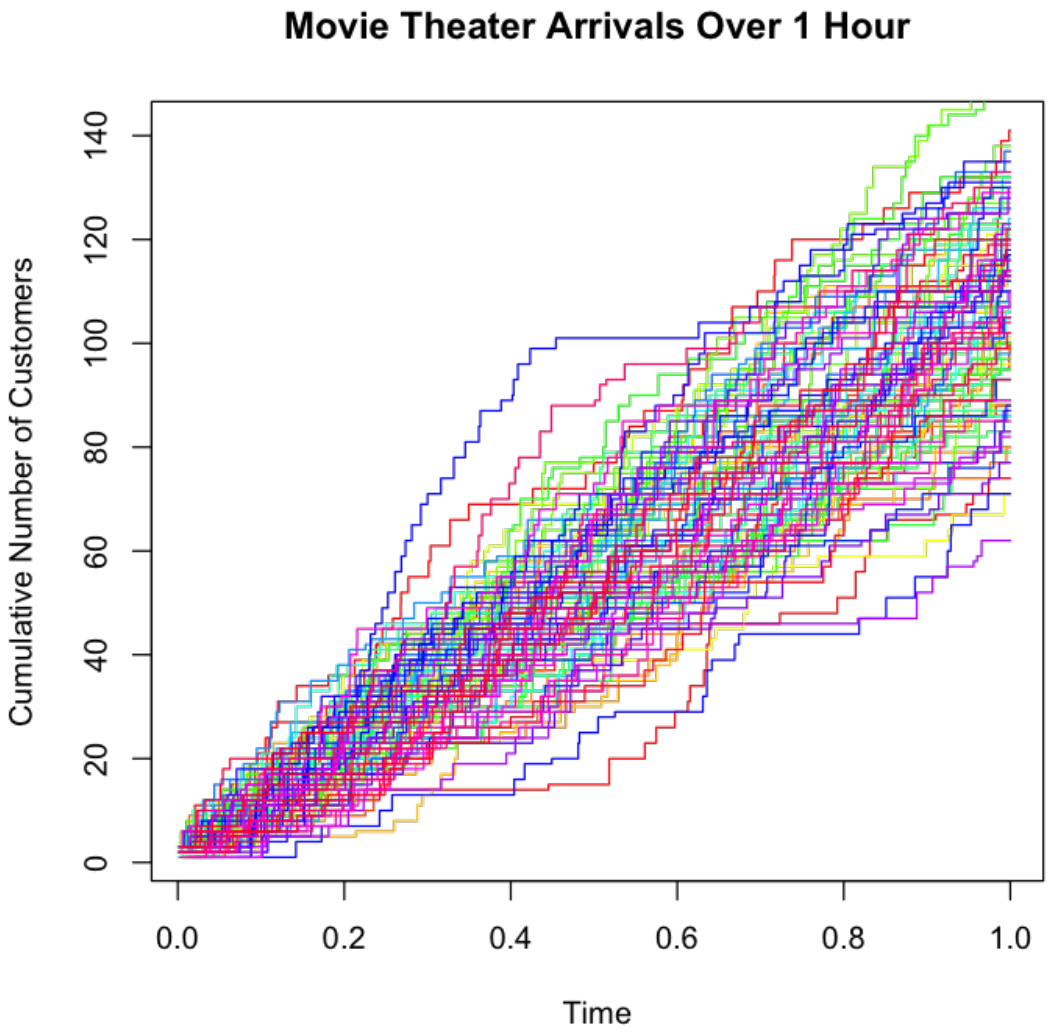
On aveage 40 groups arrive to the movie theater each hour and the average size of each group is

$$1 \times 0.10 + 2 \times 0.25 + 3 \times 0.50 + 4 \times 0.10 + 5 \times 0.05 = 2.75 \text{ customers.}$$

Therefore on averge $40 \times 2.75 = 110$ customers arrive to the theater each hour.

The following R code simulates a compound Poisson process. The function `Distribution` returns a discrete random variable according to the distribution above. The function `CompoundPoissonProcess` simulates a compound Poisson process with event sizes according to the disttribution above. Its input is the rate λ at which events occur and the terminal time T which to simulate up to and its output is a list of 2 vectors. The first vector provides the event times and the second vector the size of each event.

The following R code graphs the output of `CompoundPoissonProcess` over the course of 1 hour assuming the parameters of the example above.

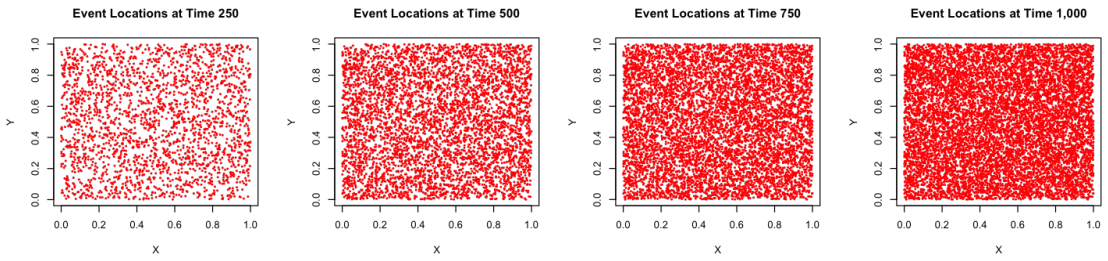


The distribution of a compound Poisson process at a fixed time $t \geq 0$ is a *compound Poisson random variable* which does not admit a closed form for its distribution function. However, if λt is large, then $N_X(t)$ can be approximated by a normal random variable with a mean of $\lambda t E[X]$ and a variance of $\lambda t E[X^2]$.

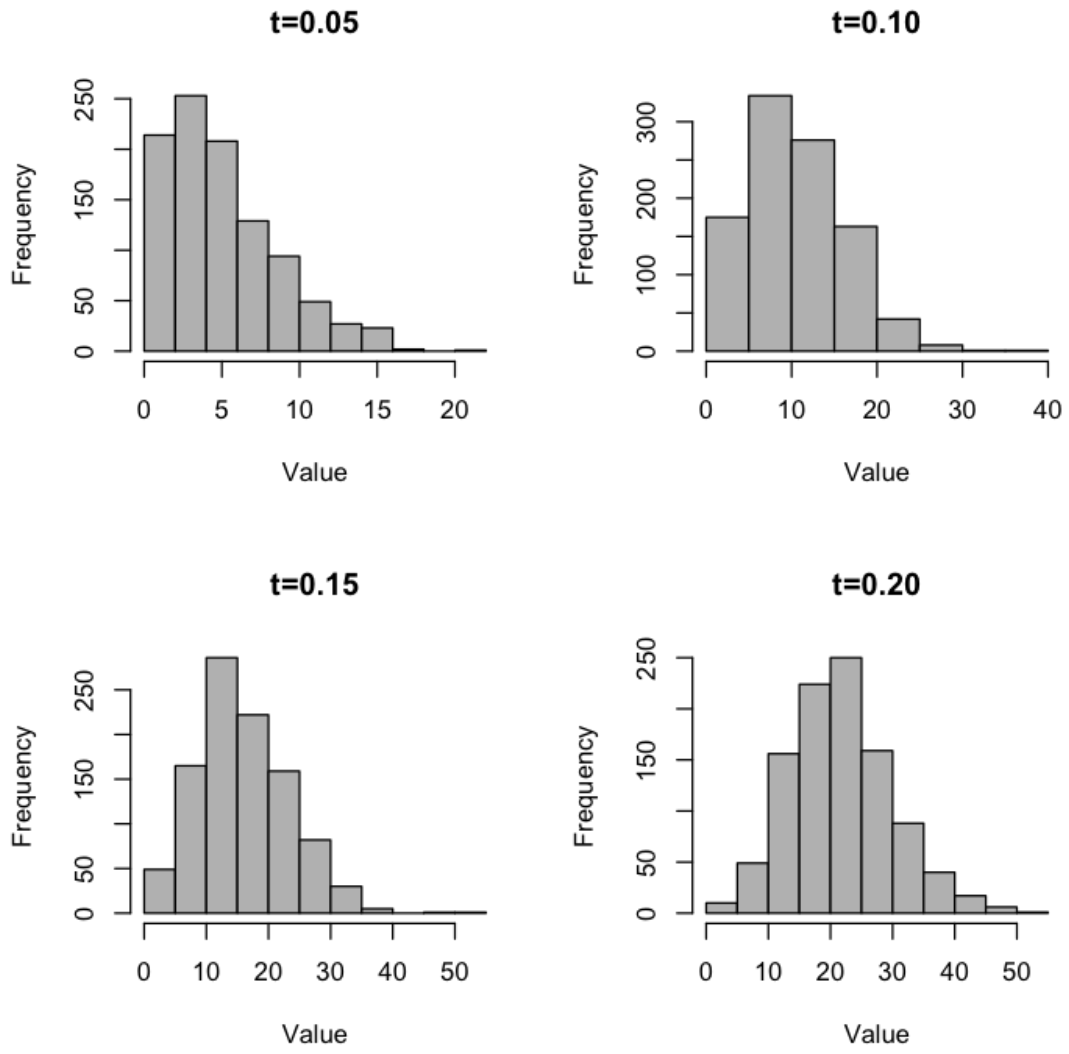
Example. Consider again the movie theater example above.

- What is the probability that 30 or more customers arrive in the first 12 minutes?

We can approximate the distribution of the number of customers that arrive in the first 12 minutes by a normal random variable with a mean of $\lambda t E[X] = 22$ customers and a variance of $\lambda t E[X^2] = 67.6$ customers^2. Hence, the desired probability is approximately $P(N(22, 76.6) > 30)$ which is about 0.17. The true value of the desired probability is close to 0.15. The R code below graphs simulation estimates of the desired probability ranging from 1 to 500 replications.



The display below provides a histogram of 1,000 simulations of $N_X(t)$ for $t = 0.05, 0.10, .0.15$ and 0.20. Note that as t becomes larger, the shape of the histogram looks more like a normal distribtion.



50 An Application to Insurance Risk

One application of compound Poisson processes is to insurance risk modeling. In the *compound Poisson model* of insurance risk theory, an insurance company collects preimiums from its customers at a constant rate and is subject to the random arrivals of claims which must be paid out. The company's capital at time t is modeled by

$$C(t) = c + pt - \sum_{n=1}^{N(t)} X_n.$$

The process $N = \{N(t), t \geq 0\}$ is assumed to be a Poisson process with rate λ and represents the times at which claims arrive. The random variable X_n is the size of the nth claim and it is assumed that $X_1, X_2, ...,$ are independent and identically distributed random variables. The parameter $p > 0$ is the rate at which premiums are paid and c is the company's initial capital.

In order to avoid the company's capital amount from eventually dropping below zero it must be the case that the rate at which premiums come in is greater than the rate at which claims are paid. This condition may be written as

$$p > \lambda E[X].$$

One way to quantify how much excess premium income is being taken in relative to claims liabilities is to write

$$p = (1 + \theta)\lambda E[X].$$

The quanity θ is referred to the *safety loading*. Even if the condition $p > \lambda E[X]$ is satisfied, the company may be unlucky and one large claim or a series of smaller sized claims can cause it to become insolvent. In this case we say that the company is ruined and one of the main focuses of *ruin theory* is to calculate the *ruin probability*.

Now note that the company's capital at time t may be written as $C(t) = c - S(t)$, where

$$S(t) = \sum_{n=1}^{N(t)} X_n - pt.$$

The time of ruin is then given by

$$T_c = \min\{t \geq 0 : S(t) > c\}.$$

There are at least two main quantities of interest in ruin theory. The first is the probability of ruin before some finite time τ. This may be written as $P(T\_c \leq \tau)$. The second is the probability of ruin at any time in the future. This may be written as $P(T_c < \infty)$.

Both of the ruin probabilities mentioned are for the most part difficult to compute and no known closed formulas exist for them. There are however a few special cases in which direct formulas can be obtained.

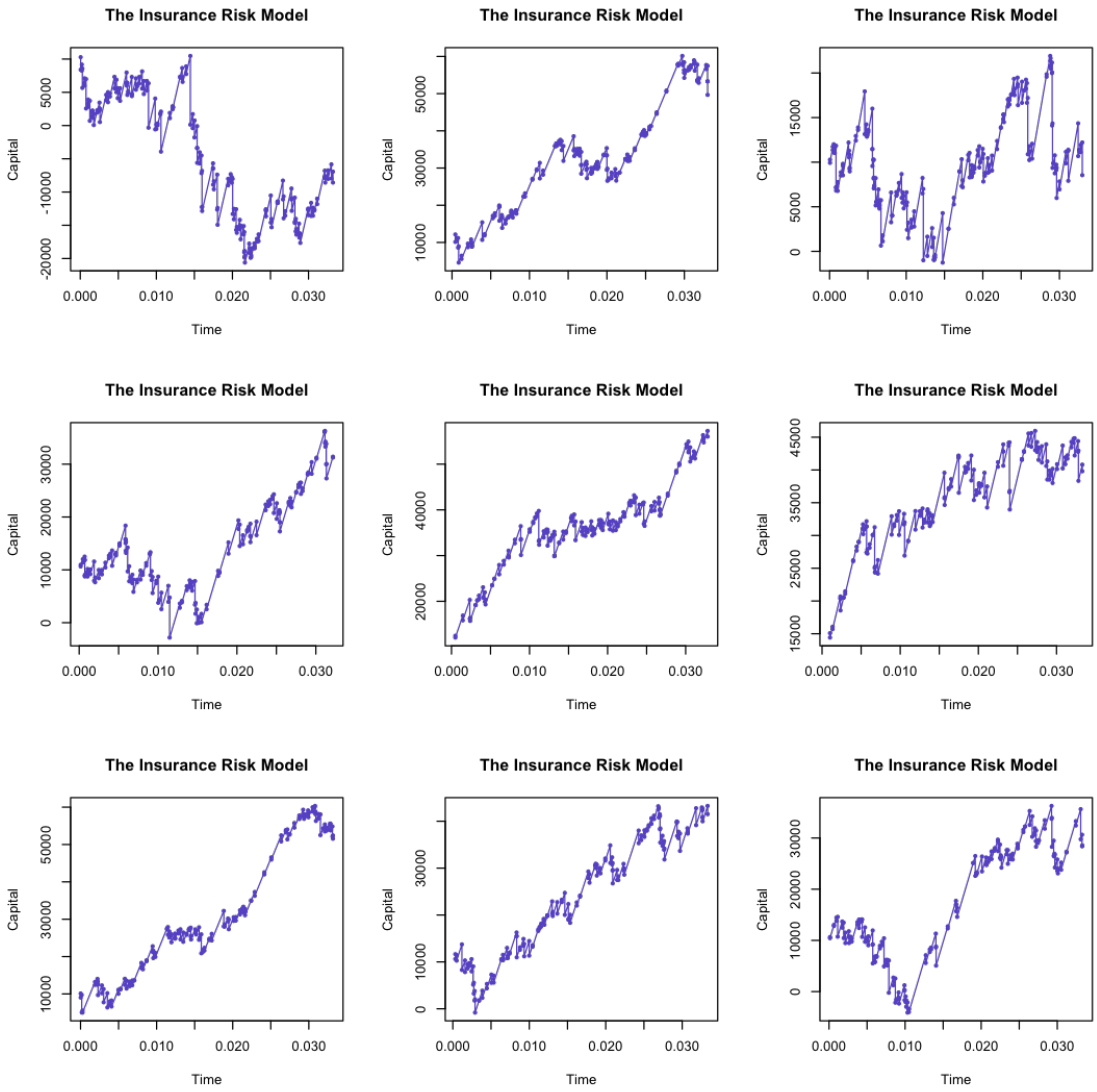
Result. Suppose that in the compound Poisson model of insurance risk claim sizes are exponentially distributed with mean μ. Then, the probability of eventual ruin is given by

$$P(T_c < \infty) = \frac{1}{1+\theta} \exp\left(-\frac{\theta}{1+\theta}\frac{c}{\mu}\right).$$

Example. Suppose that an auto insurnace business receives 3,000 claims a month and the distribution of each claim size is exponentially distributed with an average size of \$1,500. Moreover, suppose that the company receives \$5,000,000 in premium payments a month. In this case, the company's safety loading is $1/9$ and its probability of eventual ruin is

$$P(T_c < \infty) = \frac{9}{10} \exp\left(-\frac{1}{10}\frac{c}{1500}\right).$$

The R code below graphs 9 simulations of one day of a sample path of the company's capital process in the example above assuming \$10,000 of initial capital. Note that on sample paths the company's capital level always stays positive while on others the company becomes insolvent.



Monte Carlo simulation is a useful tool if closed form expressions are not available for the probability of ruin. The probability of ruin before some finite time τ is straightforward to obtain by Monte Carlo. Simulate N replications of the company's capital process up until τ and then compute the proportion of replications in which ruin occured.

Determining the probability of ruin at any time in the future is more difficult. This is because it appears to require knowledge of the company's capital levels infinitely far off into the future. Importance sampling can be used to remedy this but we choose instead to focus on two simple approximations. The first is to compute the probability of ruin up until some very large time T_∞ as an approximation to the probability of eventual ruin. The larger T_∞ is the better this approximation becomes. The second approximation is to select a very large capital level c_∞ and compute the proportion of simulation runs in which ruin occurs before c_∞ is reached. The idea behind this approximation is that once the capital level reaches c_∞ ruin is unlikely to occur.

Example. Suppose again that an auto insurnace business receives 3,000 claims a month and the distribution of each claim size is exponentially distributed with an average size of \$1,500. Moreover, suppose that the company receives \$5,000,000 in premium payments a month. In this case, the company's safety loading is $1/9$ and its probability of eventual ruin is

$$P(T_c < \infty) = \frac{9}{10} \exp\left(-\frac{1}{10}\frac{c}{1500}\right).$$

The R code below graphs the Monte Carlo estimate of $P(T_c < \infty)$ for various levels of initial capital up to \$50,000 and using the approximation of simulating up until T_∞=1 day. For each initial capital level we perform 1,000 simulation runs. In this case, we have chosen T_∞ to be only equal to 1 day because the safety loading $\theta = 1/9$ is large. The dashed red line in the curve is the true value of $P(T_c < \infty)$ using the formula above. Notice that the simulation estimates of the probabilities of ruin are consisently lower than the true values. This is because there is a positive probability that the company becomes insolvent only after 1 day. Increasing T_∞ to be larger than 1 day reduces the difference.

The R code below graphs the Monte Carlo estimate of $P(T_c < \infty)$ for various levels of initial capital up to \$50,000 and running each simulation only up until the capital level c_∞=\$60,000 is reached. For each initial capital level we perform 1,000 simulation runs. In this case, we have chosen c_∞ to be only equal to \$60,000 because the safety loading $\theta = 1/9$ is large. The dashed red line in the curve is the true value of $P(T_c < \infty)$ using the formula above. Notice that the simulation estimates of the probabilities of ruin exhibit more variability than in the case above of simulating up until T_∞. For the most part they are below the true ruin probability as would be expected but every once in a while they fall above due to the variance of the simulation estimate.

51 Session 9: Discrete Time Markov Chains

Summer 2019 - Instructor: Josh Reed

Teaching Assistant: TBD

In this session, discrete time Markov chains are studied.

52 Introduction to Discrete Time Markov Chains

A *discrete time Markov chain* is a stochastic process where conditional on the current state of the process, the distribution of the future states of the process is independent of its past states. This property is referred to as the *Markov property* of the process after the mathematician Andrey Markov. If $X = \{X_n, n = 0, 1, 2, ...\}$ is a Markov chain with a discrete state space then the Markov property may be written as

$$P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = n_0) \quad = \quad P(X_{n+1} = i_{n+1} | X_n = i_n) \tag{7}$$

for $n \geq 0$. The dynamics of most of the Markov chains that we study in this class will be time homogeneous so that the probability on the righthand side above does not change with time. Because of this we can simply write

$$P(X_{n+1} = j | X_n = i) \quad = \quad P_{i,j} \tag{8}$$

for all i and j in the state space. The matrix formed by the P_{ij} is referred to as the *1-step transition matrix* of the chain.

Example. Each day Peter is either happy (H), so-so (S) or unhappy (U).

- If Peter is happy today, then he will be H, S or U tomorrow with probabilities 0.5, 0.4, 0.1, respctively.
- If Peter is so-so today, then he will be H, S or U tomorrow with probabilities 0.3, 0.4, 0.3, respctively.
- If Peter is unhappy today, then he will be H, S or U tomorrow with probabilities 0.2, 0.3, 0.5, respctively.

If X_n denotes the Peter's mood on day n, then $\{X_n, n \geq 0\}$ is a Markov chain with state space $S = \{H, S, U\}$ and one-step transition matrix

$$P \quad = \quad \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}.$$

Example. Consider a blackjack gambler who starts out with \$$c$. The player gambler \$1 each hand with probability p and loses \$1 each hand probability $1 - p$. Moreover, each hand is independent of the others. Once the gambler has reached either \$0 or \$$N$, he quits. Model the winnings of the gambler as a Markov chain.

If X_n denotes the winnings of the gambler after n hands, then $\{X_n, n \geq 0\}$ is a Markov process with state space $S = \{0, 1, 2, ..., N\}$. The transition matrix P of X is given by $P_{0,0} = P_{NN} = 1$ and

$$P_{i,i+1} = 1 - P_{i,i-1} = p \text{ for } i = 1, ..., N-1.$$

The picture below illustrates the probabilities of transition between the various states of the gambler's Markov chain for the case of an upper limit of 5 hands. Each circle or *node* corresponds to a state and is labeled with the gambler's winnings for that state. The arrows represent potential transitions between states and are labeled with their respective probabilities. This type of picture is referred to as a *state transition diagram*.

Example. Consider a user whose is browsing the internet. We represent the internet as a graph (G, E) with N vertices. Each vertice in the graph is a webpage and a directed edge from vertice i to vertice j is a link from webpage i to webpage j. When a user is finished reading a webpage, with probability d she picks a webpage uniformly at random from the enitre internet to go to next, or with probability $1 - d$ she picks a page uniformly at random out of the pages that the current

page she is reading links to. The probability that a user transitions from webpage i to webpage j can be written as

$$P_{ij} = d \cdot \frac{1}{N} + (1-d) \cdot \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}},$$

where $0/0 = 0$ and $A_{ij} = 1$ if webpage i links to webpage j. The process keeping track of the order in which the user visits webpages is a discrete time Markov chain with transition matrix P. This model served as the basis for Google's original pagerank algorithm. The following is an early graph of the internet on June 29th, 1999 created by the Internet Mapping Project.

Example. Suppose that the weather today depends on the weather for the past two days. If it rained for the past two days, then the probability it will rain today is 0.9. If it rained yesterday but not the day before, then the probability of rain today is 0.7. If it did not rain yesterday but rained two days ago, then the probability of rain today is 0.5. Finally, if it did not rain for the past two days, then the probability of rain today is 0.4.

The process $X = \{X_n, n \geq 1\}$, where X_n is either 1 or 0, depending on whether it rained on day n or not, respectively, is not a Markov chain since the probability of rain depends on what happened for the past two days. However, if we enlarge the state space and define states 1 through 4 by setting

$$\begin{aligned}
\text{state } 1 &= \text{rain for the past two days,} \\
\text{state } 2 &= \text{rain yesterday, but no rain two days ago,} \\
\text{state } 3 &= \text{no rain yesterday, but rain two days ago,} \\
\text{state } 4 &= \text{no rain for the past two days,}
\end{aligned}$$

and let X keep track of the weather for the past two days according to states 1 through 4 above, then it is straightforward to verify that X is a Markov chain with one-step transition matrix

$$P = \begin{pmatrix} 0.9 & 0 & 0.1 & 0 \\ 0.7 & 0 & 0.3 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0.4 & 0 & 0.6 \end{pmatrix}.$$

A state transition diagram for this example is given below.

53 Calculating Transition Probabilities

One quantity of interest when studying discrete Markov chains is the probability that the chain will move from one state to the next. The one-step transition matrix provides the probability that the chain will move from state i to state j is 1 step. Next, consider the probability that the chain will move from state i to state j in 2 steps. In this case, we can condition on the first step that the chain takes to write

$$\begin{aligned} P(X_2 = j|X_0 = i) &= \sum_{k\in\mathcal{S}} P(X_2 = j, X_1 = k|X_0 = i) \\ &= \sum_{k\in\mathcal{S}} P(X_2 = j|X_1 = k, X_0 = i)P(X_1 = k|X_0 = i) \\ &= \sum_{k\in\mathcal{S}} P_{ik}P_{kj}. \end{aligned}$$

In matrix form the above identity states that the 2-*step transition matrix* is given by P^2. In general, for a discrete time Markov chain the *n-step transition matrix* is given by the 1-*step transition matrix* P raised to the power n, that is P^n.

Example. Suppose that Peter in the example above is happy today. Calculate the probability that Peter is unhappy 3 days from now. In order to calculate the probability that Peter is unhappy 3 days from we first calculate the 3-step transition marix P^3. This can be carried out in R as follows.

```
0.356 0.378 0.266
0.336 0.370 0.294
0.322 0.364 0.314
```

The probability that Peter is unhappy 2 days from now given that he is happy today is provided by the $(1,2)$nd entry of the 3-step transition matrix, which in this case is equal to 0.378.

Example. Consider the gambler in the second example above. Determine the distribution of gambler's winnings after playing 50 hands of blackjack. Assume that the gambler starts out with \$8 and stops upon reaching either \$0 or \$35, and that the probability of winning each hand is 60%.

Assuming that the gambler starts out with \$8, the distribution of the gambler's winnings after 50 hands is given by the 8th row the 50-step transition matrix. This may be found by raising the 1-step transition matrix to the 50th power. The following R code performs this computation and graphs the distribution as a bar chart. Notice the relatively high probabilities of the gambler being either going bankrupt or reaching the upper limit of \$35. This is because once these states are reached the gambler never leaves them. Such states are referred to as absorbing states.

Example. Consider the weather example above. Assuming that it rained for the past 2 days, calculate the probability that it will be sunny 4 days from today. In order to determine this probability we need to be mindful of the state space description for this example. If it rained for the past 2 days, the current state is state 1. Now consider the event that it is sunny 4 days from now. This implies that *5 days from now* the weather is either in state 3 or 4. The desired probability is therefore the probability of transitioning from state 1 to either state 3 or 4. The 5-step transition matrix for this problem can be calculated in R as follows.

```
0.71439 0.08515 0.09671 0.10375
0.67697 0.09945 0.09833 0.12525
0.59605 0.13220 0.09945 0.17230
0.58100 0.13784 0.10020 0.18096
```

The probability that it will be sunny 4 days from today is provided by the sum of the $(1,3)$rd and $(1,4)$th entries of the 5-step transition matrix, which in this case is equal to 0.20046.

54 Simulating Markov Chains

So far the examples we have encountered have not required us to simulate a Markov chain. Instead we were able to compute the desired probabilities by raising the 1-step transition matrix to a certain power. However for calculating expectations related to a Markov chain simulation can be a useful tool.

Example. Consider the gambler in the example above. Suppose that we would like to calculate the average number of hands played before either going bankrupt or reaching the upper limit of \$35. This quantity can be approximated from the transition matrix but it is illustrative to see how simulation can be used too.

In the R code below we simulate 10,000 replications of the gambler's winnings over the course of 400 hands assuming that the gambler starts out with \$8. Notice that in most cases the gambler has finished playing by 400 hands but there are a few scenarios in which he has not either gone bankrupt or reached the upper limit by then.

Now for each scenario we find the time at which the gambler finished playing and then plot the results as a histogram.

It turns out that the general formula for the average time until the gambler finishes playing is messy but if $p \neq 1/2$ can written as

$$\frac{c}{1-2p} - \frac{N}{1-2p} \cdot \frac{\left(\frac{1-p}{p}\right)^c - 1}{\left(\frac{1-p}{p}\right)^N - 1}.$$

If $p = 1/2$ it is simpler and given by $c(N-c)$. In this case since $p = 0.6$ we use the formula above and obtain a value of approximately 128.2. The average of the value returned by the Monte Carlo simulation is given above. The result obtained by simulation is biased downwards slighlty since a small percent of the sample paths do not terminate by 400 hands.

Example. Consider the internet example above assume that the internet is modeled by a 100 vertice graph randomly generated according to the Barabasi-Albert preferential attachment model. In this model the graph is "grown" one vertice at a time and new nodes "prefer" to connect to existing nodes that have a larger number of incoming edges.

Now suppose that an internet user browses the graph above for several hours according to pagerank algorithm given above. One quantity of interest is the number of times that each page is visited. This metric can serve as an estimate of how much traffic each page receives. The following R code simulates 500 scenarios of a user browing 1,000 pages in succession assuming that the initial page is uniformly distributed across the graph. The parameter d is set equal to 0.05 which implies that the user on average visits 20 pages linked together before returning to a search engine and randomly selecting a new page.

55 Classification of States

One question which is often asked above a Markov chain is whether every state is reachable from every other state? The answer to this question can be found by looking at the state transition

diagram for the chain. If there is a path from every state in the state space to every other state in the state space then we say that the Markov chain is *irreducible*. Irreducible Markov chains possess many nice properties as we will see.

Because not all Markov chains are irreducible we also need some terminology to describe this situation. If on the state transition diagram there exists a path from a state A to a state B, then we say that B is *accessible* from A. If it turns out that A is also accessible from B, then we say that A and B *communicate* with each other. An irreducible Markov chain turns out just to be a chain in which all states communicate with one another.

Example. Consider the example of Peter and his moods. From the 1-step transition diagram it is clear that every states is accessible from every other state in only 1 step. Hence, the Markov chain is irreducible in this case.

Example. Consider the gambler example above. In this case if the gambler begins with $\$c$ where $0 < c < N - 1$, then he will either end up with \$0 or $\$N$ but along the way he can visit every state in between. Thus, every state is accessible from state i for $0 < i < N$. On the other hand, once the gambler enters states 0 or N, then the gambler remains there forever and so states 0 and N only communicate with themselves. Because of this the Markov chain in this example is not irreducible.

Example. Consider the internet example above. It may not always be the case that every webpage is accessible from every other webpage simply by following a series of links from one page to the next. If this were the case we would say that the graph of the internet is *strongly connected*. However, recall that after visiting each page the user with probability d selects a page uniformly at random to go to next. This means that even if the underlying graph is not strongly connected, the Markov chain is still irreducible, so long as $d > 0$.

Example. Consider the weather example above. By viewing the outside in the state transition diagram, it can be seen the Markov chain is irreducible.

56 Steady State Distributions

Consider again the example above of the Markov chain for Peter's moods and suppose that we ask ourselves the following question. If Peter is happy today, what is the probability that he is unhappy 5 days now? How about 10 days from now? 20 days from now? Because people's moods change from day to day, it is reasonable to assume that the fact that Peter is happy today bears little relationship to whether he will be unhappy 5 days from now. And maybe even less so for 10 or 20 days from now. We can attempt to verify this claim by directly calculating the probability that Peter is unhappy 5, 10 and 20 days from now given that he is happy today. The results are as follows.

$$\begin{matrix} 0.34072 & 0.37180 & 0.28748 \\ 0.33840 & 0.37084 & 0.29076 \\ 0.33676 & 0.37016 & 0.29308 \end{matrix}$$

$$\begin{matrix} 0.3387190 & 0.3709716 & 0.2903094 \\ 0.3387082 & 0.3709671 & 0.2903246 \\ 0.3387006 & 0.3709640 & 0.2903354 \end{matrix}$$

0.3387097 0.3709677 0.2903226
0.3387097 0.3709677 0.2903226
0.3387097 0.3709677 0.2903226

Recall that in the Markov chain for Peter's moods state 1 corresponds to being happy and state 3 to being unhappy. From the calculations above we see that the probability that Peter is unhappy 5, 10 and 20 days from now given that he is happy today are 0.28748, 0.2903094 and 0.2903226. The numbers are all very close to one another and seem to be converging to a value of 0.2903 up to 4 significant digits. Moreover, notice that another interesting thing is happening. Regardless of Peter's mood was today, he has the same probability of being unhappy 10 days from now, namely 0.2903226. He also has the same probabilities of being unhappy and so-so, namely 0.3387097 and 0.3709677, respectively. These 3 probabilities together form an example of a *steady state distribution* for the Markov chain.

A Markov chain may have a unique steady state distribution, multiple steady state distributions, or no steady state distribution at all. If a Markov chain is *ergodic* it will always have a unique steady state distribution. We will not go into the technical details of the definition of an ergodic Markov chain but instead specify directly whether a chain is ergodic or not. When a unique steady state distribution exists we denote it by π.

One way to compute an ergodic Markov chain's steady state distribution is to raise its transition matrix to a very high power and then look at the probabilities of being in each state. If the probabilities do not vary by the initial state, then a steady state distribution has been found. A second way to find the steady state distribution of an ergodic Markov chain is to solve the following system of equations.

$$\pi_j = \sum_{i \in S} \pi_i P_{ij} \text{ for each } j \in S, \tag{9}$$

$$\sum_{i \in S} \pi_i = 1. \tag{10}$$

Example. Consider the example above of Peter's moods. In this case the Markov chain is ergodic and the equations for its steady state distribution are as follows.

$$\begin{aligned} \pi_1 &= 0.5\pi_1 + 0.3\pi_2 + 0.2\pi_3 \\ \pi_2 &= 0.4\pi_1 + 0.4\pi_2 + 0.3\pi_3 \\ \pi_3 &= 0.1\pi_1 + 0.3\pi_2 + 0.5\pi_3 \end{aligned}$$

together with

$$\pi_1 + \pi_2 + \pi_3 = 1.$$

Notice that the above system consits of 4 equations but only 3 unknowns. It turns out that one of the first 3 equations is redundant and can be eliminated. The final equation must always be kept.

1. 0.338709677419355 2. 0.370967741935484 3. 0.290322580645161

Example. Consider next the gambler's example. In this case the corresponding Markov chain is not ergodic. This can be seen by raising its 1-step transition matrix to a high power and analyzing the resulting structure.

Only the first and last columns of the $1,000$-step transition matrix are positive. The rest of the columns are zero. This is because after 1,000 hands the gambler has either gone bankrupt or reached his upper limit. Also, the probability of going bankrupt decreases as the gambler's initial cash increases. This is further evidence that a steady state distribution does not exist.

Example. The Markov chain corresponding to the page rank algorithm is ergodic so long as $d > 0$. In this case, the system of equations characterizing the stationary distribution π is

$$\pi_i = \frac{d}{|V|} + \frac{1-d}{|V|} \cdot \sum_{j \in M(i)} \frac{1}{L(j)} \cdot \pi_j, \text{ for each } i \in S, \tag{11}$$

$$\tag{12}$$

where $|V|$ is the number of vertices in the graph G, and $M(i) = \{k : A_{ki} = 1\}$ is the set of pages pointing to page i, and

$$L(j) = \sum_{k=1}^{|V|} A_{jk}.$$

Also it is required that $\pi_1 + ... + \pi_{|V|} = 1$. One way to interpret π_i is as the long-run fraction of time that a user surfing the web will spend on page i. Because of this, pages with a higher value of π_i receive more traffic and in this context π_i is referred to as the *PageRank* of page i. The above equations for π imply that a page's PageRank is equal to a constant plus a weighted sum of the PageRank's that point to that page i, where linking pages with less total links receive a large weight.

In general it is difficult solve the system of equations and so two options are available. The first is to attempt to raise the matrix P to a very high power which could be difficult is there are many vertices in the graph. The second is to simulate a user browsing the web and then compute the fraction of time they spend on each page. This was done in the internet example above and the resulting bar chart provides an approximation to each page's PageRank.

Example. Consider the weather example above. In this case the Markov chain describing the weather for the past 2 days is ergodic and the steady state distribution is the solution to the system of equations

$$\begin{aligned}
\pi_1 &= 0.9\pi_1 + 0.7\pi_2 \\
\pi_2 &= 0.5\pi_3 + 0.4\pi_4 \\
\pi_3 &= 0.1\pi_1 + 0.3\pi_2 \\
\pi_4 &= 0.5\pi_3 + 0.6\pi_4
\end{aligned}$$

together with

$$\pi_1 + \pi_2 + \pi_3 + \pi_4 \quad = \quad 1.$$

Solving this system of equations in R one obtains the following.

1. 0.682926829268293 2. 0.0975609756097561 3. 0.0975609756097561 4. 0.121951219512195