University of Alabama at Birmingham
Department of Mathematics

Probability Theory

Lecture Notes for MA 485/585
(1995–2013)

Dr Nikolai Chernov

July 2013

Chapter 1

Combinatorics

Five Cards

Five cards are labeled $1, 2, 3, 4, 5$. They are shuffled and lined up in an arbitrary order. How many ways can this be done? What is the chance that they line up in the proper order, i.e., as $1, 2, 3, 4, 5$?

$\boxed{2}\boxed{1}\boxed{5}\boxed{4}\boxed{3}$ or $\boxed{3}\boxed{5}\boxed{4}\boxed{2}\boxed{1}$ or $\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{5}$ or ...

Solution: If we line the cards one by one, we will have to choose one card out of five for the first place, then one out of (the remaining) four for the second place, etc. So the total number of line-ups will be

$$5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

The proper line-up is unique, so the chance of it happening is 1:120. If one one lines up the cards 120 times, then its is fair to expect the proper line-up would to occur just once.

Extra question: Suppose one plays a game betting \$1, lining up the cards and winning x dollars should the proper lineup occur. What should x be so that the game would be fair?

Answer: Obviously, $x = 120$. Indeed, in 120 rounds the player will bet \$120 and should expect to win x dollars once. If $x = 120$, then the losses and gains would cancel out.

First Rule for Permutations

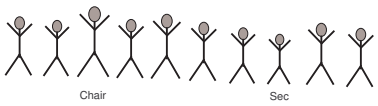
The number of ways to line up n objects is

$$P_n = n \cdot (n-1) \cdots 2 \cdot 1 = n! \qquad (\text{“n factorial”})$$

It is called the number of *permutations* of n objects.

Committee Choosing a Chair and a Secretary

A committee of 10 members decides to choose a chair and a secretary arbitrarily (at random). What is the chance that the *tallest* member becomes the chair and *shortest* – the secretary?



Solution: We can choose a chair from all the 10 members and then a secretary from the remaining 9 members. So the total number of choices will be $10 \cdot 9 = 90$. The chance that the tallest member is the chair and the shortest member the secretary is now 1:90.

Second Rule for Permutations

The number of ways to select and order (line up) m objects from a pool of n objects is

$$P_{n,m} = \underbrace{n \cdot (n-1) \cdots (n-m+1)}_{m} = \frac{n!}{(n-m)!}$$

It is called the number of *permutations* of n objects taken m at a time (or the number of m-element permutations of n objects).

Deck of Cards with Two Aces

A deck of 10 cards contains two aces. We pick two cards arbitrarily (at random). What is the chance that both are aces?

Solution: The number of ways to choose an *ordered* pair of cards is $10 \cdot 9 = 90$. The number of ways to choose an *unordered* pair of cards is $90/2 = 45$. Here we divide by two because each pair can be ordered in two ways.

Rule for Combinations

The number of ways to select m objects (without ordering) from a pool of n objects is

$$C_{n,m} = \binom{n}{m} = \frac{n \cdot (n-1) \cdots (n-m+1)}{m!} = \frac{n!}{(n-m)!\ m!}$$

It is called the number of *combinations* of n objects taken m at a time (or the number of m-element combinations of n objects).

Note that

$$\binom{n}{0} = 1 \quad \binom{n}{1} = n \quad \binom{n}{2} = \frac{n(n-1)}{2} \quad \dots \quad \binom{n}{n-1} = n \quad \binom{n}{n} = 1$$

(A standard convention is $0! = 1$). The above row is symmetric, i.e.,

$$\binom{n}{m} = \binom{n}{n-m}$$

Partitions

In how many ways one can divide (partition) a pool of n objects into two groups: one of m objects and the other of $n - m$ objects?

Solution: One just needs to choose m objects for the first group and leave the rest in the second group. So the number of choices is $\binom{n}{m}$.

In how many ways one can partition a pool of n objects into two groups of arbitrary sizes?

First solution: Since the size of the first group may take values $m = 0, 1, \dots, n$, and for each m the number of partitions is $\binom{n}{m}$ then the total is

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n-1} + \binom{n}{n}$$

Second solution: Each object in the pool can be put into the first or the second group, i.e. there are two choices for each object. Hence, there are

$$\underbrace{2 \cdot 2 \cdots 2}_{n} = 2^n$$

ways to create a partition.

Comparing the above solutions, we arrive at a useful formula:

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n \tag{1}$$

Newton's formula

The numbers $\binom{n}{m}$ are called *binomial coefficients*. They are involved in the famous Binomial expansion theorem, also called Newton's formula:

$$(a+b)^n = \sum_{m=0}^{n} \binom{n}{m} a^{n-m} b^m. \tag{2}$$

Note that (1) on p. 3 is a particular case of Newton's formula, obtained by the substitution $a = b = 1$. Another substitution, $a = 1$ and $b = -1$, gives one more remarkable formula:

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots \pm \binom{n}{n} = 0$$

(here the signs alternate). It can be written as

$$\binom{n}{0} + \binom{n}{2} + \cdots = \binom{n}{1} + \binom{n}{3} + \cdots$$

(all even values of m on the left, all odd values of m on the right). This is useful for one of homework exercises.

Pascal's triangle

The binomial coefficients $C_{n,m} = \binom{n}{m}$ can be nicely arranged in the form of a triangle:

						1								
					1		1							
				1		2		1						
			1		3		3		1					
		1		4		6		4		1				
	1		5		10		10		5		1			
1		6		15		20		15		6		1		
1		7		21		35		35		21		7		1
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Here the n-th row contains the numbers $\binom{n}{m}$ for $0 \le m \le n$. A 'magic' property of this triangle is that each number is the sum of the two closest numbers on the row directly above it.

100 Coin Tossings

A coin is tossed 100 times. What is the chance one observes exactly 50 Heads and 50 Tails?

Solution: The result of 100 tosses can be recorded by a string of H's and T's, for example

$$\text{HTTHHTHHH}\cdots\text{HTH}$$

of length 100. How many such strings do we have? It is

$$\underbrace{2\cdot 2\cdots 2}_{100}=2^{100}$$

since each letter is either H or T (two possibilities). Now, how many strings contain exactly 50 H's and 50 T's? It is same as the number of ways to pick 50 positions out of 100 available (say, we pick 50 positions for H's, filling the rest with T's). This number is $\binom{100}{50}$. Hence, the chance to observe 50 Heads is

$$\frac{1}{2^{100}}\binom{100}{50}.$$

More generally:

Coin Tossing Formula

If one tosses a coin n times, the chance to observe exactly m Heads is

$$\frac{1}{2^n}\binom{n}{m} \tag{3}$$

The numbers like above are very hard to compute for large m and n. One of the goals of probability theory is to find efficient ways to compute such numbers approximately.

Question: Guess what the number $\frac{1}{2^{100}}\binom{100}{50}$ is, approximately.

Possible answers: One naïve idea goes as follows. A fair coin must land on Heads and Tails the same number of times, so the chance to observe equal number of Heads and Tails must be high, close to 100%.

Another naive idea: the number of Heads in 100 tosses may be 0,1,2,...,100. Since 50 is one of them, the chance is 1:101, i.e. about 1%. Both naive guesses are way off mark.

A better idea: there are some very likely values for the number of Heads, such as 50 and those close to 50, and very unlikely values, those far from 50, which can be ignored. If the number of very likely values is, say, 10, then the chance is 1:10, or 10%. A good guess!

(The exact answer is 7.96%, we will arrive at it in Chapter 15.)

Team from a Group of Employees

A small company employs 10 men and 10 women. It forms a team of three employees for a special project by picking three employees at random. What is the chance that all the members of the team are women?

Solution: A quick idea is that each member of the team is a women with probability 1/2. Then all the three members would be women with probability $(1/2)^3 = 1/8$. Right? Wrong!

There are exactly $C_{20,3}$ ways to select three employees out of 20, and $C_{10,3}$ ways to select three women out of 10 available. So, the chance to pick three women is

$$\frac{C_{10,3}}{C_{20,3}} = \frac{2}{19}.$$

This is close to 1/8, but somewhat smaller. To see why it must be smaller, let us select team members one by one. After one woman is selected for the team, the balance is broken, and there are fewer women available (only 9) than men (still all the 10).

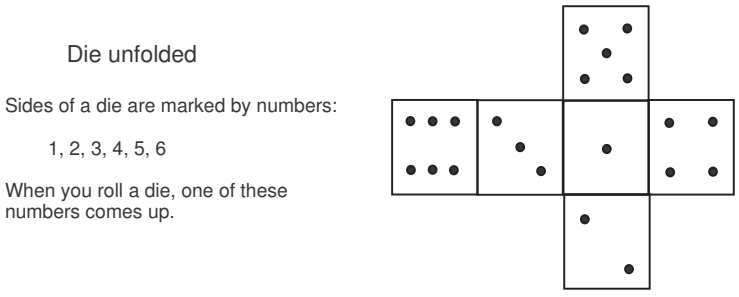
The Sum of Two Dice

Two dice are rolled. What is the chance that the sum of the numbers shown equals 9?

Solution: Each die has six faces and shows a number from 1 to 6, see below. Two dice show a pair of numbers from 1 to 6. There are $6 \times 6 = 36$ such pairs. One can make a chart of all pairs and locate the pairs that sum to 9:

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

There are 4 pairs that sum to 9, so the chance is 4/36=1/9.



Urn Problem

An urn contains 10 white balls and 20 black balls. Four balls are taken from the urn at random. What is the probability that two white and two black balls are taken?

Solution: There are $C_{30,4}$ ways to choose four balls out of 30 available. Now, there are $C_{10,2}$ ways to pick two white balls and $C_{20,2}$ ways to pick two black balls, so there are $C_{10,2} \cdot C_{20,2}$ ways to pick two white and two black balls from the urn. The probability is then

$$\frac{C_{10,2}\, C_{20,2}}{C_{30,4}} = \frac{190}{609} \approx 0.312.$$

Committee and Chairman

A group of n people is going to form a committee of k persons with a chairman. How may ways can this be done?

Solution: There is $C_{n,k}$ ways to form a committee and then k ways to select a chairman from within the committee. So, the total number is $k\, C_{n,k}$.

Committee of Variable Size (optional material)

Assume now that the size of the committee, k, is not fixed, i.e. it can take any value from 1 to n. Then the total number of ways to select a committee (of arbitrary size) with a chairman is

$$\sum_{k=1}^{n} k\, C_{n,k}.$$

Alternatively, one can form a committee with a chairman as follows. Pick a chairman first from the entire group of n people, and then allow the chairman to select members for his/her committee. The chairman will select a committee from the remaining $n-1$ people, thus partitioning them into two groups – the committee per se and the rest of the group. We already know that there are 2^{n-1} ways to partition a group of $n-1$ people into two parts. Thus, the total number of ways to select a chairman and a committee is $n\, 2^{n-1}$. Comparing this to the formula above, we arrive at another remarkable formula:

$$\sum_{k=1}^{n} k\, C_{n,k} = n\, 2^{n-1}. \tag{4}$$

Chapter 2

Probability Space

Probability theory studies experiments (procedures, games, etc.) whose results cannot be completely calculated (predicted), so that they may end up with more than one possible outcome.

Three Coin Tossings

Toss a coin three times. What are possible outcomes? What is the chance to observe exactly two Heads?

Solution: We know from Chapter 1 that the result of three tosses can be recorded by a string of H's and T's of length three. There are 8 such strings:

$$\begin{array}{ccccc} & \text{HHT} & \text{HTT} & & \\ \text{HHH} & \text{HTH} & \text{THT} & \text{TTT} & \\ & \text{THH} & \text{TTH} & & \end{array}$$

Three strings (in the second column) contain exactly two Heads, so the chance to observe two Heads is 3/8.

Stubborn Coin Flipper

A stubborn person tosses a coin until it lands heads-up. What are possible outcomes? What is the chance that three or more tosses will be necessary?

Solution: Clearly, possible outcomes are:

$$\text{H},\ \ \text{TH},\ \ \text{TTH},\ \ \text{TTTH},\ \ \text{TTTTH},\ \ldots$$

The corresponding probabilities are

$$1/2,\ \ 1/4,\ \ 1/8,\ \ 1/16,\ \ 1/32,\ \ldots$$

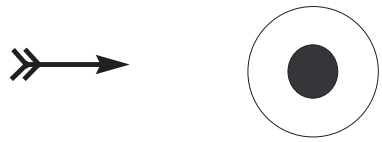
One knows from calculus that the sum of these numbers equals one, i.e.

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^n}+\cdots=1.$$

The probability that three or more tosses are necessary is found by summation

$$\frac{1}{8}+\frac{1}{16}+\cdots+\frac{1}{2^n}+\cdots=\frac{1}{4}. \tag{5}$$

Note differences between the above two examples. In the first, we could list all the outcomes, and their probabilities were equal. In the second, the number of possible outcomes is infinite, and they all have different probabilities.



Archery

One shoots at a target, which is a round disk of radius 30 inches. Assuming that the arrow lands anywhere in the target arbitrarily, what is the chance that the bull's-eye, the inner disk of radius 10 inches, will be hit?

Solution: An outcome of this experiment is the spot (point) on the target surface where the arrow lands. All the points on the surface are possible outcomes. It is important to note: one cannot assign positive probabilities to individual points (outcomes). Instead, one associates the probability to hit any region on the target surface with the area of that region. So, the probability to hit the bull's-eye is proportional to its area, or more precisely it is the *relative* area of the bull's-eye within the target:

$$\frac{\pi\, 10^2}{\pi\, 30^2} = \frac{1}{9}.$$

(Recall: the area of a disk of radius r equals πr^2.)

The last example is similar to the previous one, as there are again infinitely many possible outcomes. However, there is a big difference: the probability of *each* outcome is now zero. So the summation rule used in (5) would not work here: you cannot add zeros and get something other than zero. Positive probabilities are now assigned to whole regions within the target, not to individual outcomes.

We now review common features of the above three examples and generalize them. A random experiment always has more than one possible outcome. The collection (set) of all possible outcomes can be described and represented by a list, chart or a geometric figure. In probability theory, one is interested in probabilities of certain parts of that collection of outcomes, or subcollections (subsets) of outcomes. The probability is a number between 0 and 1.

Probability space

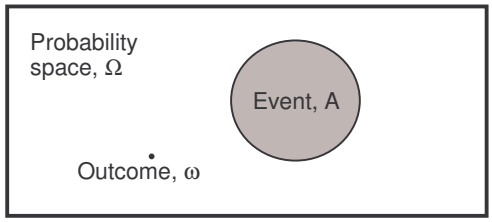
The set of all possible outcomes of a random experiment is called *probability space*. We denote it by Ω ("capital omega"). Its elements, or points, are called *outcomes*, they are denoted by ω ("little omega"). The result of the random experiment is always one point ω of Ω, i.e., $\omega \in \Omega$.

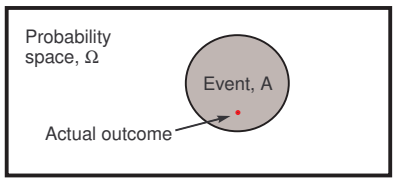
Example: if we choose a letter from the word `ABBA` randomly, then possible choices are `A` and `B`, so our probability space will consist of *two* (not four!) elements: `A` and `B`. We write $\Omega = \{\texttt{A},\texttt{B}\}$.

An *event* is a part of Ω (called a subset of Ω). It is often characterized by a certain condition (such as "two Heads are observed in three tosses" or "the bull's-eye is hit"). Events are denoted by A, B, C, etc.

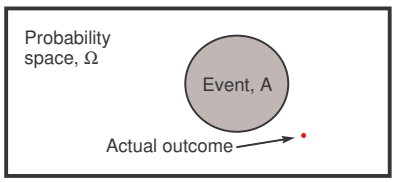
We say that an event A *occurs* if the random experiment results in an outcome ω that belongs in A, i.e., $\omega \in A$. If ω happens to be outside of A, i.e., $\omega \notin A$, the event A *does not occur*.

To visualize these concepts, we usually draw a rectangle that represents the probability space. Its points represent outcomes. Various regions within the rectangle (usually, shown as disks) represent events.



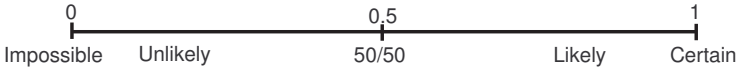


Event A does occur



Event A does **not** occur

Each event has *probability*, which is a number between 0 and 1; it represents the likelihood of A. The probability of A is denoted by $\mathbb{P}(A)$.



Rules for events and probabilities

(a) The entire Ω is called *certain* event. It always occurs because it contains every possible outcome ω. Its probability is one: $\mathbb{P}(\Omega) = 1$.

(b) There is a special symbol, $\emptyset$, used for events that never occur. They contain no outcomes (they are empty sets). Their probability is zero, $\mathbb{P}(\emptyset) = 0$. An event with no outcomes is said to be *impossible*.

(c) If A is an event, then the rest of Ω is called the *complement* of A and denoted by A^c. If A occurs, A^c does not, and vice versa. The probability of A^c is related to that of A by the rule $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(d) If A is a part of B, we write $A \subset B$ (*inclusion*). This means that A implies B (i.e., if A occurs, then B also occurs). Their probabilities satisfy the rule $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(e) The common part of two events, A and B, is called their *intersection*, denoted by $A \cap B$, or just AB. It occurs whenever both A **and** B occur.

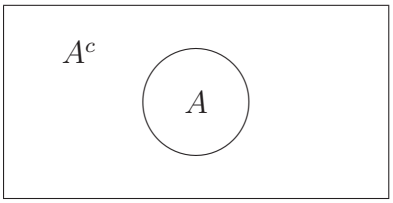
(f) The event consisting of all the outcomes that are either in A or in B is called the *union* of A and B, denoted by $A \cup B$. It occurs whenever A **or** B occurs.

(g) If two events A and B have no common part (no common outcomes; note that in this case $A \cap B = \emptyset$), then A and B are said to be *disjoint*, or *mutually exclusive*. They cannot occur simultaneously. In this case we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

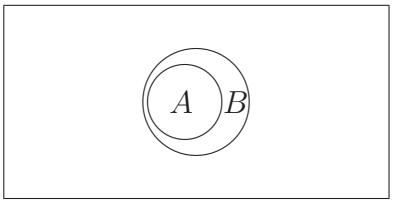
In the Stubborn Coin Flipper example, it is possible (at least theoretically) that the coin always lands on Tails and the flipping will never stop. So, there is one outcome that we have overlooked: TTT... (infinitely many T's). This outcome has probability zero.

Events that have probability zero, even if they do contain some outcomes, are often called *impossible*, too. If we want to distinguish them from impossible events that contain no outcomes (as described in (b) above), we can say that an event is *physically impossible* if it contains no outcomes. Example: if you roll two dice, then it is physically impossible to have their sum equal 14.

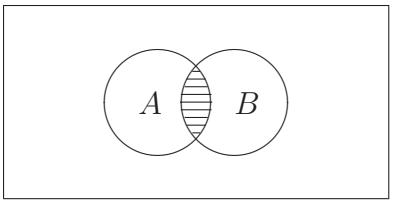
An event is physically possible but *probabilistically impossible* if it contains some outcomes, but its probability is zero. Example: infinitely many T's in the Stubborn Coin Flipper case. Another example: arrow landing right at the center of the bull's-eye in the Archery example. Yet another example: arrow landing right on the border of the bull's-eye in the Archery example. (Note: the border is a circle, which is just an extremely thin line; its thickness is zero, so its is area equals its length times zero, which equals zero.)



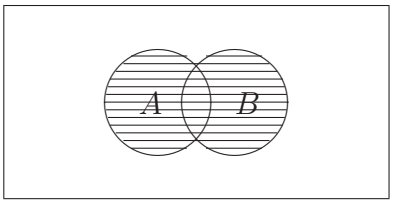
$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$



$A \subset B; \quad \mathbb{P}(A) \le \mathbb{P}(B)$



$A \cap B, \ AB; \quad A$ and B



$A \cup B; \quad A$ or B

Venn's diagrams

The above diagrams illustrate the rules of probability. The big rectangle always represents the probability space Ω. The disks inside the rectangle represent events A, B, etc. Such nice pictures are called *Venn's diagrams*.

De Morgan's laws

The following De Morgan's laws may be useful:

$$(A \cup B)^c = A^c \cap B^c \qquad \text{and} \qquad (A \cap B)^c = A^c \cup B^c.$$

They are easy to verify by examining Venn's diagrams.

Distributive laws

The following distributive laws may be useful:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad \text{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

To verify these, draw three overlapping circles representing A, B, C and shadow the related areas.

Summation rules

For two events, A and B, we have

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

For three events, A, B, C, we have

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

Similarly, for four events $A_1, \ldots, A_4$ we have

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup A_4) = \sum_i \mathbb{P}(A_i) - \sum_{i \neq j} \mathbb{P}(A_i \cap A_j) \\ + \sum_{i \neq j \neq k} \mathbb{P}(A_i \cap A_j \cap A_k) - \mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4).\end{aligned}$$

This type of formulas are called inclusion-exclusion formulas.

Two Dice

Two dice are rolled. What is the chance that at least one six will be shown?

Solution: One can use a chart as in Example 1.15:

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

By direct count, there are 11 outcomes where at least one die shows six. So the chance is 11/36. A more elegant solution is obtained as follows: let $A = \{\text{The first die shows 6}\}$ and $B = \{\text{The second die shows 6}\}$. Then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}.$$

• Note: in the Stubborn Coin Flipper example, we can apply the Complementary Event Rule as follows:

$$\mathbb{P}(\geq 3 \text{ tosses}) = 1 - \mathbb{P}(1 \text{ toss}) - \mathbb{P}(2 \text{ tosses}) = 1 - \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

Two Simple Examples

1. Let $\mathbb{P}(A) = 0.5$, $\mathbb{P}(B) = 0.3$ and $\mathbb{P}(A \cap B) = 0.1$. Find $\mathbb{P}(A^c \cap B^c)$ and $\mathbb{P}(A^c \cup B^c)$. Answers: 0.3 and 0.9, respectively. (Just draw a Venn's diagram to see this.)

2. Let $\mathbb{P}(A) = 0.8$, $\mathbb{P}(B) = 0.7$. Find the minimum possible value for $\mathbb{P}(A \cap B)$. Answer: 0.5. See solution below.

Solution to Example 2: Note that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) = 1.5 - \mathbb{P}(A \cup B).$$

Since $\mathbb{P}(A \cup B)$ cannot exceed 1, you cannot subtract more than the unity from 1.5. So, we obtain $\mathbb{P}(A \cap B) \geq 1.5 - 1 = 0.5$. This is the minimum value.

Birthday Problem

A class has 30 students. What is the chance that some two students have birthdays on the same day?

Solution: Intuitively, the coincidence of two birthdays seems to be very unlikely. If so, our intuition must be misleading. Because the chance of coincidence is actually very high.

To solve the problem, notice that $\mathbb{P}(\text{coincidence}) = 1 - \mathbb{P}(\text{no coincidence})$, and the latter probability is

$$\mathbb{P}(\text{no coincidence}) = \frac{P_{365,30}}{365^{30}} = \frac{365 \cdot 364 \cdots 336}{365 \cdot 365 \cdots 365} \approx 0.2937.$$

Hence, the chance of coincidence is $1 - 0.2937 = 0.7063$, i.e., over 70%.

Here is an explanation to the above formula: 365^{30} is the number of ways 30 students may have birthdays, and $P_{365,30}$ is the number of ways 30 students may have birthdays on 30 *distinct* days. (The day of February 29 in leap years is ignored, for simplicity.)

Why was our intuition so misleading? Well, it is because we compared a small number of students, 30, to a large number of days, 365. Instead, we should have thought of the number of *pairs* of students, which is $C_{30,2} = 435$. This number is *larger* than the number of days in a year.

Chapter 3

Conditional Probability and Independence

Three Coin Tosses Again

A friend tosses a coin three times. You accidentally notice that the first time the coin lands head-up (but you do not see how it lands two more times). What is the chance that the friend observes 2 Heads in all the three tosses?

Solution: In Chapter 2, we found all possible outcomes. There are eight of them. Now, with the additional information at our disposal, we can exclude those starting with a T. That leaves us with four possible outcomes:

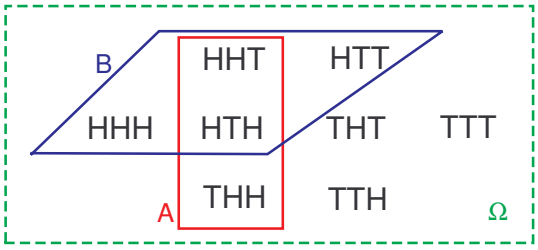
$$\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}$$

Two of them contain exactly two Heads. So, the chance is 2/4=1/2.

Note that we have two events around:

$$A = \{\text{2 Heads are observed}\} \qquad \text{and} \qquad B = \{\text{First toss is Heads}\}$$

We already know that $\mathbb{P}(A) = 3/8$, from Chapter 2. Now, the event A is considered under the condition that the event B has occurred. Then the conditional probability of A, given B, is found by calculating the fraction of A within B, i.e. the fraction of $A \cap B$ within B. See the illustration.



The event A covers three outcomes, out of eight total, within the entire space Ω. But it covers only two outcomes, out of four total, within the event B.

In a sense, our additional knowledge (that the first time the coin landed heads-up) made us reduce the probability space, so that the event B now plays the role of a new, reduced, probability space.

Conditional probability

The conditional probability of an event A, given an event B, is

$$\mathbb{P}(A/B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{6}$$

Multiplication rule

The above formula can be rewritten as

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

A symmetric formula also holds:

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

Two Spades from a Deck of Cards

A deck of 52 cards has 13 spades. If two cards are drawn from the deck at random, what is the chance that both are spades?

Solution: Let $A = \{\text{First card is a spade}\}$ and $B = \{\text{Second card is a spade}\}$. Clearly, $\mathbb{P}(A) = 13/52 = 1/4$. If the first card is a spade, then the chance to draw another spade is $12/51$ (the remaining deck of 51 cards has 12 spades left). This means that $\mathbb{P}(B/A) = 12/51$. Hence,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B/A) = \frac{1}{4} \cdot \frac{12}{51} = \frac{12}{204} = \frac{1}{17}.$$

Extended multiplication rule

If $A_1, A_2, \ldots, A_n$ are events, then

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2/A_1) \cdot \mathbb{P}(A_3/A_1 \cap A_2) \times \\ \times \cdots \times \mathbb{P}(A_n/A_1 \cap \cdots \cap A_{n-1}).\end{aligned}$$

Birthday Problem Revisited

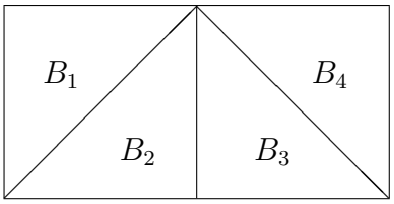
The birthday problem of Chapter 2 now can be solved by using the extended multiplication rule:

$$\mathbb{P}(\text{no coincidence}) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{336}{365}.$$

Here we take students one by one, and multiply the conditional probabilities that the birthday of each student is different from the birthdays of the previously taken students.

Partition

Let $B_1, \ldots, B_n$ be disjoint (i.e., mutually exclusive) events; i.e. $B_i \cap B_j = \emptyset$ for all $i \neq j$. Let $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$, i.e. these events cover (exhaust) the entire probability space. We call $\{B_1, \ldots, B_n\}$ a *partition* of Ω.



A partition of Ω into four events: B_1, B_2, B_3, B_4

Note that every outcome ω belongs to one and only one of $B_1, \ldots, B_n$. In other words, exactly one of these events occurs.

Law of Total Probability

Let $B_1, \ldots, B_n$ be a partition of Ω, as defined above. Let A be an event. Then

$$\mathbb{P}(A) = \mathbb{P}(B_1) \cdot \mathbb{P}(A/B_1) + \cdots + \mathbb{P}(B_n) \cdot \mathbb{P}(A/B_n) \qquad (7)$$

One can think of $B_1, \ldots, B_n$ as conditions under which the event A may occur. The events $B_1, \ldots, B_n$ are often called *hypotheses*.

Alex Goes to School

Alex goes to school by bus or train, whichever comes first. He notices that the bus comes first with probability 30% and the train with probability 70%. When Alex takes a train, he arrives late to school with probability 5%. When he takes a bus, he is late to school with probability 20%. Without knowing what mode of transportation he will take tomorrow, find the probability that he will be late to school?

Solution: The event in question here is $A = \{\text{Alex is late to school}\}$. This may happen under two conditions (hypotheses): $B_1 = \{\text{Alex takes bus}\}$ and $B_2 = \{\text{Alex takes train}\}$. Hence,

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(B_1) \cdot \mathbb{P}(A/B_1) + \mathbb{P}(B_2) \cdot \mathbb{P}(A/B_2) \\ &= 0.3 \times 0.2 + 0.7 \times 0.05 = 0.095.\end{aligned}$$

Two-Stage Experiment

Amanda rolls a die and then flips a coin the number of times that she sees on the die when it lands. What is the chance she observes two Heads?

Solution: In the first stage, the die shows one of the six numbers $1, \ldots, 6$. These are six events, which we denote by $B_1, \ldots, B_6$. They are disjoint and exhaust all the possibilities, so they make a partition. In the second stage, the event $A = \{\text{Two Heads are observed}\}$ may (or may not) occur, and its probability depends on the number shown on the die. Applying the law of total probability gives

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(B_1) \cdot \mathbb{P}(A/B_1) + \cdots + \mathbb{P}(B_6) \cdot \mathbb{P}(A/B_6) \\ &= \frac{1}{6} \cdot 0 + \frac{1}{6} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{3}{8} + \frac{1}{6} \cdot \frac{C_{4,2}}{2^4} + \frac{1}{6} \cdot \frac{C_{5,2}}{2^5} + \frac{1}{6} \cdot \frac{C_{6,2}}{2^6} = \frac{33}{128}.\end{aligned}$$

Note that we used the formula (3) from page 5 to find the probability of observing 2 Heads in n tosses for $n = 2, \ldots, 6$.

Two Dice Again

Roll a die twice. If the first roll is a six, what is the chance the second roll will be a six?

Solution: Let $A = \{\text{The second roll is a six}\}$ and $B = \{\text{The first roll is a six}\}$. Then

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

Note that $\mathbb{P}(A) = 1/6$, so that

$$\mathbb{P}(A/B) = \mathbb{P}(A).$$

In other words, the probability of A does not change when the event B occurs, the event B does not affect the chance of A to occur.

Independent Events

Two events, A and B, are said to be *independent* if

$$\mathbb{P}(A/B) = \mathbb{P}(A).$$

By using (6), we can rewrite this equation as

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \tag{8}$$

and also as

$$\mathbb{P}(B/A) = \mathbb{P}(B).$$

All these three equations mean the same – independence of A and B.

Note: The equation $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$ is better than the other two: it is symmetric. It also works when $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$, while the other two may fail. So, it is preferred for practical purposes.

Tossing Two Coins

Suppose we flip two coins. Let

$$A = \{\text{First coin shows Head}\}, \qquad B = \{\text{Both coins show the same face}\}$$

Are A and B independent?

Solution: One easily finds that $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 1/2$ and $\mathbb{P}(A \cap B) = 1/4$. Then we just check that $1/2 \times 1/2 = 1/4$. So, yes, they are independent.

Note: Sometimes the independence is obvious, like in the previous example with two dice (because there is no way the first die can affect the second). Sometimes the independence is harder to recognize, as it is in the above example with two coins. One can explain the independence here noting that the second coin may or may not show the same face as the first one with probability 1/2, no matter what face the first coin shows.

Note: If two events A, B are independent, then A^c, B^c are also independent, i.e. $\mathbb{P}(A^c \cap B^c) = \mathbb{P}(A^c)\,\mathbb{P}(B^c)$. Moreover, A and B^c are independent, i.e, $\mathbb{P}(A \cap B^c) = \mathbb{P}(A)\,\mathbb{P}(B^c)$. Similarly, A^c and B are independent.

Independence of Three Events

Three events A, B, C are said to be mutually (or jointly) independent if
(a) every pair of them are independent in the sense of (8), and
(b) the following holds:

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

Note: neither condition (a) and (b) alone is enough for joint independence. One needs to check *both* (a) and (b) to verify joint independence of A, B, C.

Tossing Two Coins Again

Suppose we flip two coins. Let

$$A = \{\text{First coin shows Head}\}, \qquad B = \{\text{Both coins show the same face}\}$$

and

$$C = \{\text{Second coin shows Head}\}$$

Are A, B, C jointly independent?

Solution: We have seen already that A and B are independent. Similarly, B and C are independent. Obviously, A and C are independent. So, the requirement (a) above holds. We can say that A, B, C are *pairwise independent*.

On the other hand, $\mathbb{P}(A \cap B \cap C) = 1/4$, and $1/2 \times 1/2 \times 1/2 \neq 1/4$, so the requirement (b) fails. Thus the events A, B, C are *not* jointly independent.

Tossing Three Coins

Suppose we flip three coins. Let

$$A = \{\text{First coin shows Head}\}, \qquad B = \{\text{Second coin shows Tail}\}$$

and

$$C = \{\text{At least two coins show Head}\}$$

Are A, B, C jointly independent?

Solution: One easily finds that $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$. Also, $\mathbb{P}(A \cap B \cap C) = 1/8$, so requirement (b) above holds. But A and C are dependent, so requirement (a) fails.

Independence of Several Events

Several events $A_1, \ldots, A_n$ are said to be mutually (or jointly) independent if for any subcollection $A_{i_1}, \ldots, A_{i_k}$ of them the following holds:

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k})$$

Rocket with Redundant System

A rocket has a built-in redundant system. It has three components, K_1, K_2, K_3 that can automatically replace each other. If component K_1 fails, it is bypassed and component K_2 takes over, etc. So, as long as one component works the system is functioning. Suppose that the probabilities of failure of these components are 10%, 20% and 5%, respectively. Find the probability that the entire system works.

Solution: First, note: $\mathbb{P}(\text{system works}) = 1 - \mathbb{P}(\text{system fails})$. The system fails if all the three components fail. The failures are mutually independent events, so

$$\mathbb{P}(\text{system fails}) = 0.1 \cdot 0.2 \cdot 0.05 = 0.001$$

So, the entire system will function with probability 99.9%. A remarkably high reliability!

An additional note: it is more difficult to find the probability that exactly two components fail, because they can fail in various combinations: $\{1, 2\}$, $\{1, 3\}$, and $\{2, 3\}$. In each case the remaining component is assumed to be working. Therefore, the probability that two components fail is

$$\begin{aligned}\mathbb{P}(\text{two components fail}) &= \mathbb{P}(\text{1st and 2nd fail}) \\ &\quad + \mathbb{P}(\text{1st and 3rd fail}) + \mathbb{P}(\text{2nd and 3rd fail}) \\ &= 0.1 \cdot 0.2 \cdot 0.95 + 0.1 \cdot 0.8 \cdot 0.05 + 0.9 \cdot 0.2 \cdot 0.05 \\ &= 0.032.\end{aligned}$$

A useful note: If several events $A_1, \ldots, A_n$ are independent, one can replace any number of them by their complements (e.g., A_1 by A_1^c, etc.), and the new collection of events will be also independent.

Recall the law of total probability (7). Suppose we need to compute $\mathbb{P}(B_i/A)$ for some $i = 1, \ldots, n$. Our basic formulas give

$$\mathbb{P}(B_i/A) = \frac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B_i) \cdot \mathbb{P}(A/B_i)}{\mathbb{P}(A)}$$

Let us now replace the denominator $\mathbb{P}(A)$ by (7). Then we obtain

Bayes Formula

$$\mathbb{P}(B_i/A) = \frac{\mathbb{P}(B_i) \cdot \mathbb{P}(A/B_i)}{\mathbb{P}(B_1) \cdot \mathbb{P}(A/B_1) + \cdots + \mathbb{P}(B_n) \cdot \mathbb{P}(A/B_n)}$$

Note: the numerator is one of the terms that appear in the denominator.

Interpretation of Bayes Formula

We first recall how we interpreted the law of total probability (7). An event A can occur under different conditions (hypotheses) $B_1, \ldots, B_n$. The probability that A occurs under each condition B_i is known. The likelihood of each condition to take place is known, too. Then we can compute the total (or unconditional) probability of A by (7).

Now, the situation "turns around". Suppose we know that the event A has occurred. But we are not aware of under what condition this happened. So we need to estimate the probability that the condition B_i had taken place before the event A occurred. This is what the Bayes formula does.

Who is Bayes?

Bayes formula was named after the Reverend Thomas Bayes (1701-1761) who derived and used it first.

Three Factories and Defective Chip

Three factories, F_1, F_2, and F_3, produce computer chips. The factory F_1 produces 50% of all the chips in the market, the factory F_2 accounts for 40% of the market, and the factory F_3 for 10%. It is known that 1% of the chips made by the factory F_1 are defective. For the factories F_2 and F_3 the rates of defective chips are 2% and 3%, respectively. Suppose Betty's computer has a defective chip. Betty wonders: which factory is most likely to have produced it?

Solution: First, Betty assumes that it should be F_1, which accounts for most of the chips in the market. On second thought, Betty assumes that it is F_3, whose chips are the least reliable. The exact solution shows that it is F_2.

Let B_1, B_2, B_3 denote the events that Betty's computer chip was produced by F_1, F_2, and F_3, respectively. Denote by A the event that the chip turns out defective. Then $\mathbb{P}(A/B_1) = 0.01$, $\mathbb{P}(A/B_2) = 0.02$, $\mathbb{P}(A/B_3) = 0.03$. Now, by the Bayes formula, we have

$$\mathbb{P}(B_1/A) = \frac{0.5 \cdot 0.01}{0.5 \cdot 0.01 + 0.4 \cdot 0.02 + 0.1 \cdot 0.03} = \frac{5}{16}$$

$$\mathbb{P}(B_2/A) = \frac{0.4 \cdot 0.02}{0.5 \cdot 0.01 + 0.4 \cdot 0.02 + 0.1 \cdot 0.03} = \frac{8}{16}$$

$$\mathbb{P}(B_3/A) = \frac{0.1 \cdot 0.03}{0.5 \cdot 0.01 + 0.4 \cdot 0.02 + 0.1 \cdot 0.03} = \frac{3}{16}$$

Well, the highest chance is shown for the factory F_2. So, Betty should blame the factory F_2, it was most likely to have made her defective chip.

Smoking on Plane

Statistics show that 3% of men smoke but only 1% of women do. During a non-smoking flight, one passenger is smoking in the restroom. There are 40 male and 60 female passengers on the plane. What is the chance that the person smoking in the restroom is a man?

Solution: Let M, F denote the events that an arbitrarily chosen passenger is a man or a women, respectively. On this airplane, $\mathbb{P}(M) = 0.4$ and $\mathbb{P}(F) = 0.6$. Let S denote the event that a passenger is a smoker. Then $\mathbb{P}(S/M) = 0.03$ and $\mathbb{P}(S/F) = 0.01$. By the Bayes formula we have

$$\mathbb{P}(M/S) = \frac{0.4 \cdot 0.03}{0.4 \cdot 0.03 + 0.6 \cdot 0.01} = \frac{12}{18} = \frac{2}{3}$$

so it is twice more likely that the smoker is a man than a woman.

The last two examples are optional. They are somewhat counterintuitive and may provoke creative thinking and lively discussions. But they are not necessary for developing basic skills.

Brothers and Sisters

Suppose for simplicity that the number of children in a family is 1, 2, or 3, with probability 1/3 each, and boys and girls appear equally likely. Little Bobby has no brothers. What is the probability that he is an only child?

Solution: Let B_1, B_2, B_3 be the events that a family has one, two, or three children. Let A be the event that a family has only one boy. We assumed that $\mathbb{P}(B_1) = \mathbb{P}(B_2) = \mathbb{P}(B_3) = 1/3$. Now it is simple to find $\mathbb{P}(A/B_1) = 1/2$, $\mathbb{P}(A/B_2) = 1/2$ and $\mathbb{P}(A/B_3) = 3/8$. Then

$$\mathbb{P}(B_1/A) = \frac{1/3 \cdot 1/2}{1/3 \cdot 1/2 + 1/3 \cdot 1/2 + 1/3 \cdot 3/8} = \frac{4}{11}.$$

Surprise?

Let us change the previous example a bit. Suppose now that little Bobby has no sisters. What is the probability that he is an only child?

Solution: One might assume that this probability is exactly the same as in the previous example, i.e. 4/11. Why?

Remember, boys and girls appear equally likely. In the previous example you knew that Bobby could only have sisters, now you know that Bobby can only have brothers, right? What difference does it make, whether Bobby has siblings of one sex or the other?

Well, apparently it does make a difference if we look at numbers. Let again B_1, B_2, B_3 be the events that a family has one, two, or three children. Let A be the event that a family has no girls. Then it is simple to find $\mathbb{P}(A/B_1) = 1/2$, $\mathbb{P}(A/B_2) = 1/4$ and $\mathbb{P}(A/B_3) = 1/8$. Then

$$\mathbb{P}(B_1/A) = \frac{1/3 \cdot 1/2}{1/3 \cdot 1/2 + 1/3 \cdot 1/4 + 1/3 \cdot 1/8} = \frac{4}{7}.$$

Surprise? Yes, it is not so easy to understand why it is more likely that a boy with no sisters is the only child than a boy with no brothers... Go figure...

Chapter 4

Discrete Random Variables

Counting Heads

Suppose again a coin is tossed three times. And suppose you play a game in which you win \$1 each time the coin shows Head. Then your total win is X dollars where X is the number of Heads in three flips. Possible values of X are 3, 2, 1, 0, with respective probabilities 1/8, 3/8, 3/8, 1/8.

Note that X may take 4 distinct values, as opposed to 8 distinct outcomes in this game. The number of values is lower than the number of outcomes. This is so because we do not care in which order Heads and Tails come, all we care about is the total number of Heads. Hence, for example, three outcomes HHT, HTH, THH are not distinguishable, they are "combined" into one value of X (which is 2).

Success/Failure Trials

Let us generalize the above example. Suppose that you perform three trials, in each of which you may succeed or fail. For example, you take three tests on the pass/fail basis. Or you throw a basketball. Or roll a die in a game where you win \$2 if the die shows 5 or 6 and lose \$1 otherwise. Each trial has two possible outcomes: success (S) and failure (F). The experiment consisting of 3 trials has 8 outcomes. We arrange them in the same format as Heads and Tails of a coin:

	SSF	SFF	
SSS	SFS	FSF	FFF
	FSS	FFS	

In many cases, we only care about the total number of successes, let us call it X. Then X takes values 3, 2, 1, 0.

The essential difference of this situation from the coin tosses is that the probability of success is not necessarily equal to 1/2. Assume that the probability of success in every trial is the same, call it p. Then the probability of failure is $1-p$, we denote it by q. So, p and q take values between 0 and 1 and are related by $p+q=1$. Suppose also that successes and failures in individual trials are independent. Then the probabilities of outcomes in our experiment can be found by a simple multiplication rule, for example $\mathbb{P}(\mathrm{HHT}) = ppq = p^2q$, $\mathbb{P}(\mathrm{THT}) = qpq = pq^2$, etc. This way we can find the probabilities that X takes values $0, 1, 2, 3$. We summarize them in the table below:

values of X	0	1	2	3
probabilities	q^3	$3pq^2$	$3p^2q$	p^3

Note that $X=1$ combines three outcomes, all with the same probability pq^2. Similarly, $X=2$ combines three outcomes, all with the same probability p^2q. Finally, note that by Newton's formula (2) of Chapter 1

$$q^3 + 3pq^2 + 3p^2q + p^3 = (q+p)^3 = 1^3 = 1$$

so the probabilities sum up to one, as they should.

Bernoulli Trials

Any simple trial with only two possible outcomes is called *Bernoulli*[1] *trial*. It is customary to label the outcomes by "success" and "failure" (S and F). Suppose we perform n simple trials where in each trial success has probability p (the same from trial to trial) and the outcomes of trials are mutually independent. This experiment is called a *sequence of Bernoulli trials*.

An outcome of a sequence of n Bernoulli trials can be represented by a string of S's and F's of length n. The probability of an outcome given by a sequence of S's and F's can be found simply by multiplying the corresponding p's and q's (here and everywhere $q=1-p$ is the probability of Failure). Hence, if the string has k Successes (S's) and $n-k$ Failures (F's), its probability is p^kq^{n-k}.

[1]Named after Swiss scientist Jacob Bernoulli (1655–1705).

Binomial Random Variable

Suppose n Bernoulli trials are performed. The total number of successes is often what one needs to know. Call it X. Possible values of X are $0, 1, \ldots, n$. The value of X depends on the outcome of the experiment. This way we can consider X as a function on Ω, with numerical values. For example, if we perform $n = 3$ trials, then $X(\text{SSS}) = 3$, $X(\text{FSF}) = 1$, $X(\text{FFF}) = 0$, etc. We will continue this discussion after a brief digression.

Random Variables

A *random variable* is a function on the probability space Ω, whose values are numbers. We denote random variables by X, Y, Z, U, V, etc. Hence, if X is a random variable then for every outcome ω its value $X(\omega)$ is a number that can be computed.

Note that a random variable X can take the same value on several (or many) distinct outcomes, i.e. $X(\omega) = X(\omega')$ for some $\omega \neq \omega'$. So, knowing the value of X it may not be possible to identify the outcome ω. Thus, a random variable provides an incomplete information about the outcome of the experiment. This is not bad. In many cases a random variable simply suppresses unnecessary details (such as the order in which Heads and Tails come, if we only care about the number of Heads).

Binomial Random Variable (continued)

We call X described above a *binomial random variable*, resulted from n Bernoulli trials. Next we find the probability that $X = k$ for each $0 \leq k \leq n$.

Solution: We note that $X = k$ when the outcome of the experiment is a string that contains k S's and $n - k$ F's. Each such string has probability $p^k q^{n-k}$. There are $C_{n,k}$ of such strings, as we know from Chapter 1. Hence,

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}, \qquad 0 \leq k \leq n. \tag{9}$$

This formula is good for all $k = 0, 1, \ldots, n$. Note that the sum of these probabilities is

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

by the Binomial theorem (page 4). This is why X is called the *binomial* random variable.

Note that the probability $\mathbb{P}(X = k)$ above depends on two quantities, n and p (because q is merely a shorthand for $1 - p$). We call n and p the *parameters* of the binomial random variable X and say that X is binomial(n, p) or shortly $b(n, p)$.

Geometric Random Variable

Generalizing the Stubborn Coin Flipper example of Chapter 2, consider independent Bernoulli trials that are performed until a success occurs. The outcomes in this experiment are

$$\text{S}, \quad \text{FS}, \quad \text{FFS}, \quad \text{FFFS}, \ \ldots, \text{FFFFS}, \ \ldots$$

Let X be the number of trials performed. We call X a *geometric random variable*. It takes values $1, 2, \ldots, n, \ldots$. Each value $X = n$ is taken on exactly one outcome, $\underbrace{\text{F}\ldots\text{F}}_{n-1}\text{S}$. The probability of this outcome is $\underbrace{q\cdots q}_{n-1}p = pq^{n-1}$. Hence,

$$\mathbb{P}(X = n) = pq^{n-1} \qquad \text{for all} \quad n \geq 1. \tag{10}$$

Note that the sum of these probabilities is

$$\sum_{n=1}^{\infty} pq^{n-1} = p\,(1 + q + q^2 + \cdots) = p \cdot \frac{1}{1-q} = \frac{p}{p} = 1 \tag{11}$$

so the probabilities sum to one, as they should. The probability $\mathbb{P}(X = n)$ above involves the only parameter p. We denote X by geometric(p).

Note: in Calculus, the infinite sum $1 + q + q^2 + \cdots = \frac{1}{1-q}$ is called *geometric series*. This explains the name of our random variable.

Exercise with Geometric Random Variable

For a geometric random variable X, compute $\mathbb{P}(X > k)$.

Solution: We have

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

Discrete Random Variables

Generalizing the above examples, we say that a random variable X may take some values $x_1, x_2, \ldots$ with corresponding probabilities $p_1, p_2, \ldots$. The list of values may be finite as for Binomial R.V. or infinite as for Geometric R.V. It is sometimes convenient to put them all in a table:

X	x_1	x_2	x_3	$\cdots$
$\mathbb{P}$	p_1	p_2	p_3	$\cdots$

Random variables that allow such representation are said to be *discrete*. We will see a different type of random variables in Chapter 5.

Total Probability Rule

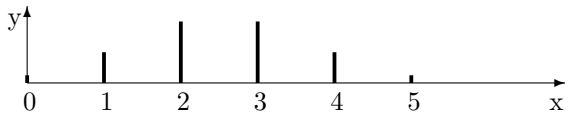
The probabilities must always sum to one:

$$p_1 + p_2 + p_3 + \cdots = 1$$

Probability Mass Function

The function that assigns the probability p_k to the value x_k is called *probability mass function* (p.m.f.) or sometimes just *probability function*. For example, the formulas (9) on page 27 and (10) on page 28 give probability mass functions for binomial and geometric random variables, respectively.

Probability mass function of a binomial r.v. $b(5, 1/2)$:



Uniform Discrete Random Variable

Let $n \geq 1$. A very simple random variable takes values $1, 2, \ldots, n$ with equal probabilities, $1/n$. Its probability mass function is

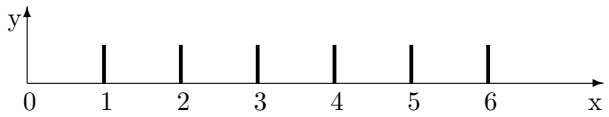
$$\mathbb{P}(X = k) = 1/n \qquad \text{for all} \quad 1 \leq k \leq n.$$

We call X a *uniform* discrete random variable.

Example: Die

An example of a uniform discrete random variable is the number shown when a die is rolled: there we had $n = 6$ and $\mathbb{P}(X = k) = 1/6$ for $k = 1, \ldots, 6$.

Probability mass function of a uniform r.v. with $n = 6$:



Notational Remark

It is important to show the range of the variable in the formula for the probability mass function, such as $0 \leq k \leq n$ in (9) on page 27 and $n \geq 1$ in (10) on page 28. It is assumed that the probability is zero for all other values of k. For example, if X is $b(n, p)$, then $\mathbb{P}(X = -1) = 0$, $\mathbb{P}(X = n + 5) = 0$, $\mathbb{P}(X = 1.4) = 0$, etc.

Special binomials: n Large, p Small

In some practical situations we have a binomial random variable with very large n and very small p. Here are a few examples:

(a) The number of calls taken by an operator. Here the number of people (customers) who may call is usually very large, but the probability that an individual customer calls at a particular moment is usually very small.

(b) The number of customers arriving at a convenience store or a car shop on a given day. Again, we have a large number of potential customers and a small probability that any particular customer will visit that store (shop) on that day.

(c) The number of lottery tickets that win if you buy a huge number of them (each ticket wins with a very low probability).

(d) The number of defective items found by a quality control in a production line. Usually, the fraction of defective items is small (say 1% or lower), and a few dozens or hundreds of items are being taken for a test.

Poisson Approximation to Binomial

Let X be a binomial random variable, $b(n,p)$, with small p and large n, as described above. The exact formula (9) on page 27 for $\mathbb{P}(X=k)$ is practically useless because it involves huge numbers such as $n!$ and tiny numbers such as p^k that may cause trouble even if you use a good computer. Our goal is to approximate $\mathbb{P}(X=k)$ by a formula that only involves "reasonable" numbers. We assume that n is huge, p is tiny, and k is reasonably small: $k=0,1,2\ldots$.

First, we note that

$$\mathbb{P}(X=k)=\frac{n(n-1)\cdots(n-k+1)}{k!}\,p^k(1-p)^{n-k}.$$

Since n is huge, we have $n-k\approx n$, and so

$$\mathbb{P}(X=k)\approx\frac{n^k}{k!}\,p^k(1-p)^n=\frac{(np)^k}{k!}\left[(1-p)^{1/p}\right]^{np}.$$

There is a useful formula in calculus:

$$\lim_{x\to 0}(1-x)^{1/x}=e^{-1}$$

based on which we approximate $(1-p)^{1/p}$ by e^{-1}. We also denote the product np by λ. Hence,

$$\mathbb{P}(X=k)\approx\frac{\lambda^k}{k!}e^{-\lambda}.$$

Note that $\lambda=np$ is the product of a huge number n and a tiny number p, so usually λ is a "reasonable" number.

Conclusion: if X is a binomial random variable $b(n,p)$ with large n and small p, one can compute the probability $\mathbb{P}(X=k)$ for small k by

$$\mathbb{P}(X=k)\approx\frac{\lambda^k}{k!}e^{-\lambda}\qquad\text{with}\qquad\lambda=np. \tag{12}$$

This formula is called *Poisson*[2] *approximation to binomials.*

Note: the value $\lambda=np$ has the (intuitively understandable) meaning of the *average number* of successes in n trials.

[2]Named after French mathematician Siméon Denis Poisson (1781-1840).

Poisson Random Variable

Motivated by (12) on the previous page, we introduce *Poisson random variable* X that takes values $0, 1, 2, \ldots$ with probabilities

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!}\, e^{-\lambda} \qquad \text{for all} \quad k \geq 0.$$

Here $\lambda > 0$ is a parameter. We denote it by $X = \text{poisson}(\lambda)$ or shortly $X = p(\lambda)$. One can check that the probabilities sum up to one:

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = e^0 = 1.$$

Here we used the Taylor expansion for e^x known from calculus:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

The Poisson approximation (12) works very well in practical applications.

Defective Items

In a production line, 0.4% of items are defective. If $n = 500$ items are taken randomly for quality control, what is the probability that 0 (or 1, or 2) of them are found defective?

Solution: Clearly, the number X of defective items in the group of 500 is binomial(500,0.004). The Poisson approximation gives

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

with

$$\lambda = 500 \cdot 0.004 = 2.$$

Hence,

$$\mathbb{P}(X = 0) \approx e^{-2}, \qquad \mathbb{P}(X = 1) \approx 2e^{-2},$$
$$\mathbb{P}(X = 2) \approx 2e^{-2}, \qquad \mathbb{P}(X = 3) \approx \frac{4}{3} e^{-2}, \ \ldots$$

All these numbers are easily computable with any simple calculator.

Role of λ

Poisson approximation (12) only requires the average number of successes $\lambda = np$, the values of n and p separately are not involved. They need not even be known!

Coffee Break

An operator knows that she receives about 5 calls per hour, on the average. She decides to take a 10 minute coffee break. What is the chance that somebody calls during her coffee break?

Solution: Here the average number of calls is 5 per hour, so it is 5/6 per 10 minute period. Hence, $\lambda = 5/6$. Then

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - e^{-\lambda} = 1 - e^{-5/6} = 0.5654.$$

So, it is more likely than not that her coffee break will be interrupted by a call.

Note: the original average of 5 call was given per our. We had to adjust it as our time slot was only 10 minutes. Beware that the average may have to be adjusted in other examples, too.

Chapter 5

Continuous Random Variables

Here we study random variables that can take any value in the real line or any value in a given interval.

Archery Again

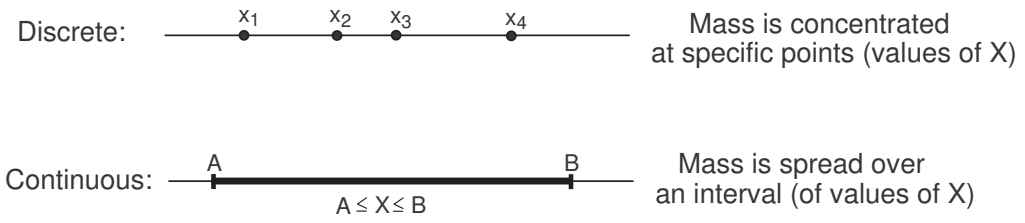
Continuing the Archery example of Chapter 2, let X be the distance from the hit point to the center of the target. Then X takes any value between 0 and 30 (inches). By using a formula, $0 \leq X \leq 30$.

Lifetime

Let X be the lifetime of a brand new TV (or a rand new car). Since it is totally unpredictable when the TV (or the car) dies, we have to assume that X may take any positive value, i.e., $X \geq 0$.

Range of Values

In the above examples, we have random variables that take values in a certain (finite or infinite) interval. Clearly, we cannot list all possible values in any table or chart, in the way we did for discrete random variables in Chapter 4. This is a new type of random variables, that we will call *continuous*.



Probabilities of Values

Another novelty in the above examples: for any possible value x, the probability $\mathbb{P}(X = x)$ is zero (in the Archery example, the value $X = x$ is taken on a circle of radius x, which is just a curve on the target surface, and the area of any curve is zero.) Since we have $\mathbb{P}(X = x) = 0$ for every individual value x, we have to think of how to describe the random variable X in a meaningful way. Instead of individual values of X we will care about *intervals* of values of X, i.e. we will consider probabilities $\mathbb{P}(a < X < b)$ for various $a < b$. Such probabilities are usually positive, and they describe the random variable X in a meaningful and complete way.

Probabilities of Intervals

For a given random variable X, the probability $\mathbb{P}(a < X < b)$ depends on both a and b, so it is a function of two variables. Fortunately, it can be reduced to a function of one variable by the following trick:

$$\mathbb{P}(a < X < b) = \mathbb{P}(X < b) - \mathbb{P}(X \leq a)$$

where $\mathbb{P}(X < b)$ and $\mathbb{P}(X \leq a)$ depend on one variable each.

Distribution Function

Given a random variable X, its *distribution function* $F_X(x)$ is defined by

$$F_X(x) = \mathbb{P}(X \leq x).$$

Note that X denotes the random variable, and x is the argument of the function F_X, i.e. a real variable $-\infty < x < \infty$.

Archery (continued)

Let us compute the distribution function for the random variable X in the Archery example.

Solution: The random variable X takes values $0 \le X \le 30$. Hence, if $x < 0$, then $X \le x$ is an impossible event and $\mathbb{P}(X \le x) = 0$. If $x > 30$, then $X \le x$ is always true (it is a certain event), hence $\mathbb{P}(X \le x) = 1$. If $0 \le x \le 30$, then the event $\{X \le x\}$ occurs if the hit point lies in the inner disk of radius x. Now, as we described in Chapter 2, we have

$$\mathbb{P}(X \le x) = \frac{\pi x^2}{\pi\, 30^2} = \frac{x^2}{900}.$$

Finally, we have

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x^2/900 & \text{if } 0 \le x \le 30 \\ 1 & \text{if } x > 30 \end{cases} \tag{13}$$

Properties of Distribution Function

It is clear that $F_X(x)$ always has the following properties:

- $0 \le F_X(x) \le 1$ (since it equals the probability of an event).
- $F_X(x)$ is monotonically increasing, i.e. $F_X(x_1) \le F_X(x_2)$ whenever $x_1 \le x_2$ [this is because of the inclusion $\{X \le x_1\} \subset \{X \le x_2\}$, which implies $\mathbb{P}(X \le x_1) \le \mathbb{P}(X \le x_2)$].
- If X has a maximum value, $X_{\max}$, i.e. $X \le X_{\max}$, then $F_X(x) = 1$ for all $x \ge X_{\max}$. Also, if X has a minimum value, $X_{\min}$, i.e. $X \ge X_{\min}$, then $F_X(x) = 0$ for all $x < X_{\min}$.
- Generalizing the previous observation, we have

$$\lim_{x\to\infty} F_X(x) = 1 \quad \text{and} \quad \lim_{x\to-\infty} F_X(x) = 0.$$

It is interesting that any function $y = F(x)$ that satisfies the first, second, and fourth of the above properties and, in addition, is continuous from the right at every point x, is a distribution function for some random variable X. We will not need that fact, though.

Computation of Probabilities

For any interval (a, b) we have

$$\mathbb{P}(a < X \leq b) = F(b) - F(a).$$

Also,

$$\mathbb{P}(X \leq b) = F(b) \qquad \text{and} \qquad \mathbb{P}(X > a) = 1 - F(a).$$

Archery (continued)

By the above formulas, we can compute

$$\mathbb{P}(1 < X < 3) = F(3) - F(1) = \tfrac{3^2}{900} - \tfrac{1^2}{900} = \tfrac{8}{900},$$

$$\mathbb{P}(X > 20) = 1 - F(20) = 1 - \tfrac{20^2}{900} = 1 - \tfrac{4}{9} = \tfrac{5}{9},$$

$$\mathbb{P}(10 < X < 40) = F(40) - F(10) = 1 - \tfrac{10^2}{900} = 1 - \tfrac{1}{9} = \tfrac{8}{9}.$$

Note that $F(40) = 1$, not $\frac{40^2}{900}$; see rule (13) on page 36.

Continuous Random Variables

Here is official definition: A random variable X is said to be *continuous* if for any real number x we have $\mathbb{P}(X = x) = 0$. In this case the distribution function $F_X(x)$ is a continuous function (it has no jumps or interruptions).

Casual Treatment of Endpoints

Note: if X is continuous, then in all the formulas for Computation of Probabilities above we have $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$. Thus, it does not matter whether one uses exclusive inequality ($<$) or inclusive one ($\leq$) in those formulas. In particular, we have

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = F(b) - F(a).$$

We will only use distribution function for continuous random variables, so we will be rather casual in some formulas like $\mathbb{P}(a < X < b)$ or $\mathbb{P}(a \leq X \leq b)$ including or excluding the endpoints a and b at will. This will not make any difference. When working with continuous random variables it does not matter if one includes endpoints of intervals or not.

Probability Density Function

The expression $F(b) - F(a)$ in the previous formulas reminds us of the fundamental theorem of calculus:

$$F(b) - F(a) = \int_a^b f(x)\,dx \qquad \text{where} \quad f(x) = F'(x).$$

The function $f(x) = F'(x)$ is called the *probability density function*. Now we can compute probabilities in terms of $f(x)$:

$$\mathbb{P}(a < X < b) = \int_a^b f(x)\,dx.$$

Also,

$$\mathbb{P}(X < b) = \int_{-\infty}^b f(x)\,dx \qquad \text{and} \qquad \mathbb{P}(X > a) = \int_a^\infty f(x)\,dx.$$

Note also that $F(x)$ can be computed itself in terms of $f(x)$:

$$F(x) = \int_{-\infty}^x f(u)\,du \tag{14}$$

That is, F is an antiderivative of f.

Properties of Density Function

- $f(x) \geq 0$.
- If $X \leq X_{\max}$, then $f(x) = 0$ for all $x \geq X_{\max}$.
- If $X \geq X_{\min}$, then $f(x) = 0$ for all $x \leq X_{\min}$.

Normalization Rule

The total integral of $f(x)$ equals one:

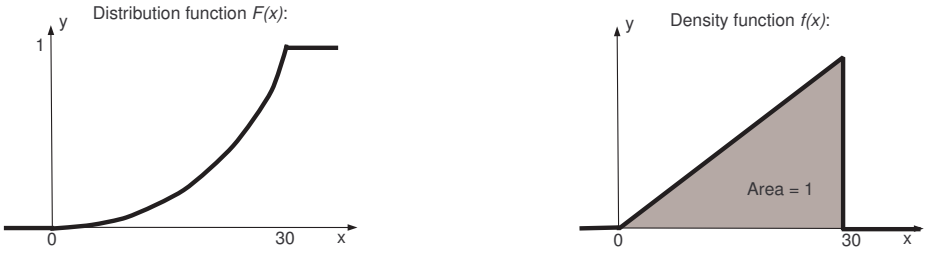
$$\int_{-\infty}^\infty f(x)\,dx = 1$$

Min/Max Rules

If X has a maximum value, $X_{\max}$, then $F(x) = 1$ and $f(x) = 0$ for all $x \geq X_{\max}$. If X has a minimum value, $X_{\min}$, then $F(x) = 0$ and $f(x) = 0$ for all $x < X_{\min}$. So, both functions $F(x)$ and $f(x)$ take trivial values outside the interval $[X_{\min}, X_{\max}]$, on which the random variable takes all its values. It is therefore customary to only specify $F(x)$ and/or $f(x)$ on the 'essential' interval $[X_{\min}, X_{\max}]$ and omit their description beyond that interval.

Archery (continued)

In the archery example, we can just say $F(x) = x^2/900$ for $0 < x < 30$. It is implicitly assumed that $F(x)$ is trivial elsewhere, i.e. $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > 30$. The corresponding density function is $f(x) = F'(x) = x/450$ for $0 < x < 30$. Again, it is understood that $f(x) = 0$ beyond the interval $0 < x < 30$. Note that it is necessary to specify the interval on which the formulas for $F(x)$ and $f(x)$ hold!



Archery example

Quiz Questions

Which of the following functions are distribution functions? For those that are, find the density function.

(1) $F(x) = x$ for $-1 < x < 1$. [No, since $F(x)$ is negative for $-1 < x < 0$.]

(2) $F(x) = x^2$ for $-1 < x < 1$. [No, since $F(x)$ decreases for $-1 < x < 0$.]

(3) $F(x) = 1 - x^{1-\rho}$ for $x > 1$ (here $\rho > 1$ is a constant). Yes. [The density is $f(x) = (\rho - 1)x^{-\rho}$. Random variables with this density are said to have *power law*.]

Use of Normalization Rule

Suppose X has probability density function $f(x) = cx$ for $1 < x < 4$ and 0 elsewhere. Find the value of c.

Solution: To find c, we use the normalization rule on page 38:

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \int_1^4 cx\,dx = cx^2/2\Big|_1^4 = 15c/2.$$

From this equation we get $c = 2/15$. Hence, $f(x) = 2x/15$.

More Exercises

In the previous example, compute $\mathbb{P}(X > 2)$, $\mathbb{P}(2 < X < 3)$, $\mathbb{P}(X > 0)$, $\mathbb{P}(X \geq 4)$, $\mathbb{P}(X = 2)$ and $\mathbb{P}(X > 2/X < 3)$.

Solution: First, we find the distribution function:

$$F(x) = \int_{-\infty}^{x} f(u)\,du = \int_1^x \frac{2}{15}u\,du = \frac{1}{15}u^2\Big|_1^x = \frac{1}{15}(x^2 - 1)$$

for all $1 \leq x \leq 4$. Note that we actually integrate from 1 (the minimum of the random variable X) to x. It should be also understood that $F(x) = 0$ for $x < 1$ and $F(x) = 1$ for $x > 4$, but this need not be shown explicitly.

Now,

$$\mathbb{P}(X > 2) = 1 - F(2) = 1 - 3/15 = 4/5,$$
$$\mathbb{P}(2 < X < 3) = F(3) - F(2) = 8/15 - 3/15 = 1/3,$$
$$\mathbb{P}(X > 0) = 1 - F(0) = 1 - 0 = 1,$$
$$\mathbb{P}(X \geq 4) = 1 - F(4) = 1 - 15/15 = 0.$$

Next, $\mathbb{P}(X = 2) = 0$ since X is a continuous random variable. Lastly, $\mathbb{P}(X > 2/X < 3)$ is a conditional probability, so

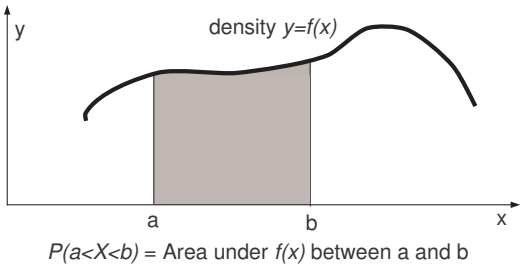
$$\mathbb{P}(X > 2/X < 3) = \frac{\mathbb{P}(2 < X < 3)}{\mathbb{P}(X < 3)} = \frac{F(3) - F(2)}{F(3)}$$
$$= \frac{(3^2 - 1) - (2^2 - 1)}{3^2 - 1} = \frac{5}{8}.$$

Probability Density Function and Area

The probability

$$\mathbb{P}(a < X < b) = \int_a^b f(x)\,dx$$

equals the area under the graph of the density function. Since the total (combined) probability of all possible values for any random variable equals one, the area under the entire graph of the density function $y = f(x)$ equals one (this is exactly the normalization rule given on page 38).



P(a<X<b) = Area under *f(x)* between a and b

Another Interpretation of Density

Let $(c, c + d)$ be a small interval near a point c. Assume that the interval is so small that the density $f(x)$ is almost constant on it, i.e. $f(x) \approx f(c)$. Hence

$$\mathbb{P}(c < X < c + d) = \int_c^{c+d} f(x)\,dx \approx f(c) \cdot d$$

So,

$$f(c) \approx \mathbb{P}(c < X < c + d)/d$$

for small d. Hence, the density is the ratio of the probability that X takes value in a small interval and the length of that interval. In other words, the density is the "probability per unit length".

Note: the higher $f(x)$, the more likely the value x and nearby values are taken by the random variable. In the archery example, the density $f(x) = x/450$ increases as x goes from 0 to 30. So, the least likely values are those near 0, and the most likely values are those near 30. This makes perfect sense, because to get $X \approx 0$ we need to hit a small area around the center of the target. The values $X \approx 30$ correspond to hitting a much larger area all around the outer edge of the target.

Uniform Random Variable

The simplest type of a continuous random variable is a variable that takes values in an interval (a, b) where all values are "equally likely". That is, the density $f(x)$ is constant on (a, b), i.e. $f(x) = c$, where c is some constant.

To find c, we can use the normalization rule $\int_a^b f(x)\,dx = 1$, i.e.,

$$\int_a^b c\,dx = c\int_a^b dx = c(b-a) = 1$$

Solving this equation for c gives $c = 1/(b-a)$. Hence,

$$f(x) = \frac{1}{b-a} \qquad \text{and} \qquad F(x) = \frac{x-a}{b-a} \tag{15}$$

for $a < x < b$; and these functions take their trivial values outside the interval (a, b). We denote this random variable by $X = U(a, b)$. Note that the graph of the density function $f(x)$ is a rectangle over the interval (a, b), this is why uniform distribution is sometimes referred to as *rectangular*.

Uniform random variable *U(a,b)*

One can think of the value of a uniform random variable $U(a, b)$ as a (completely) randomly selected point from the interval (a, b).

According to the same principle as in the archery example, the probability to hit any smaller interval (u, v) inside (a, b) is proportional to its length, i.e.

$$\mathbb{P}(u < X < v) = \frac{v-u}{b-a}$$

This entirely agrees with the above formulas for $f(x)$ and $F(x)$.

Special Uniform, U(0,1)

One uniform random variable, $X = U(0,1)$, plays an exclusive role in probability theory. This will be made clear below and in subsequent chapters. We only note here that its density and distribution functions are given by very simple formulas:

$$f(x) = 1 \quad \text{and} \quad F(x) = x \qquad \text{for } \ 0 < x < 1. \tag{16}$$

Random Number Generators (RNG)

The uniform random variable $U(0,1)$ plays an important role in computer programming. Many computer software packages (such as MATLAB, Maple, Mathematica) have built-in random number generators that, upon request, produce numbers between 0 and 1 which are supposed to be completely random. Each time an RNG is called it returns a new random number between 0 and 1. This way a computer RNG simulates the uniform random variable $U(0,1)$. In MATLAB, for example, you can call the RNG by typing `x=rand`, then `x` will be a random number between 0 and 1.

Chapter 6

Exponential Random Variables

Waiting Time

As in the Coffee Break example on page 33, suppose an operator receives $\lambda > 0$ calls per hour, on the average. Let T denote the waiting time until she gets her next call. Find its distribution function F_T.

Solution: Obviously, T takes positive values, $T > 0$. Now we have

$$F_T(x) = \mathbb{P}(T \leq x) = 1 - \mathbb{P}(T > x)$$

The event $T > x$ means that no calls arrive during the time interval $(0, x)$. By the logic used on page 33, the average number of calls during this interval is λx, so we have $\mathbb{P}(T > x) = \mathbb{P}(k = 0) = e^{-\lambda x}$. Thus

$$F_T(x) = 1 - \mathbb{P}(T > x) = 1 - e^{-\lambda x} \qquad (\text{for all} \quad x > 0)$$

Exponential Random Variable

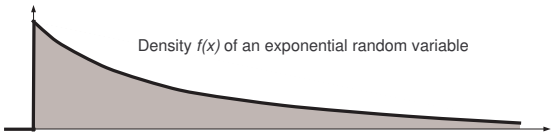
An exponential random variable X takes positive values, $X > 0$, and its distribution function is

$$F_X(x) = 1 - e^{-\lambda x} \qquad \text{for all} \quad x > 0.$$

By differentiating, we obtain the density function

$$f(x) = \lambda e^{-\lambda x} \qquad \text{for all} \quad x > 0.$$

The constant $\lambda > 0$ is the parameter of the exponential random variable. We denote this variable by exponential(λ).



Poisson Process

We see that the waiting time for the next call is an exponential random variable with parameter $\lambda > 0$ that represents the average number of calls per unit time. A similar example: the waiting time until the next customer arrives in a convenience store or a car shop (as was mentioned on page 30).

This is a common feature of any process that involves events occurring randomly, at random times, in which the average number of events per unit time remains unchanged (stable). Such events can be marked by points on a line (time axis), their locations correspond to times when the events occur. The locations of points are random, and the number of points in any given interval is random, too.

There are other examples of that sort where the line is *not* a time axis. Consider failures in a long cable/phone line. Or accidents on a long highway. Or state trooper patrol cars deployed on a long highway. In all these examples the locations of failures/accidents/patrol cars are random, and even their number in any given interval of the line is random.

Any sequence of random points on a line of the above type is known as *Poisson process*. The average number of those points per unit length is denoted by $\lambda > 0$ and called the density or *rate*; it is a numerical parameter of the whole process. The interval/distance from any given point on the line to the next point of the process is an exponential random variable with parameter λ.

Time to Failure

Suppose a company is using a machine or device (an airplane, a ship, a truck, a boat, etc.) which occasionally fails and needs repair or requires service. Then the time to (the next) failure is again an exponential random variable with parameter $\lambda > 0$ that represents the average number of failures per unit time.

This is a common interpretation of exponential random variable X, so we will stick to it. We will interpret X as the *time to failure*, and the parameter $\lambda > 0$ as the *failure rate* (the mean number of failures per unit time).

Memoryless Property (No Aging)

Let X be an exponential random variable. Consider two events

$$A = \{X > a\} \qquad \text{and} \qquad B = \{X > a + x\}$$

for some $a > 0$ and $x > 0$. Note: A means that the time to failure exceeds a, i.e. the machine functions properly at least a units of time; B means that the machine functions properly at least $a + x$ units of time. Let us compute the conditional probability $\mathbb{P}(B/A)$. To state the question differently: given that the machine has worked without failures a units of time, what is the probability that it will work another x units of time without failure?

Recall that

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

Note that the event $B = \{X > a + x\}$ implies $A = \{X > a\}$, i.e. $B \subset A$, so $B \cap A = B$. Hence,

$$\mathbb{P}(B/A) = \frac{\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(X > a + x)}{\mathbb{P}(X > a)} = \frac{e^{-\lambda(a+x)}}{e^{-\lambda a}} = e^{-\lambda x}.$$

Compare this to the probability

$$\mathbb{P}(X > x) = 1 - F(x) = e^{-\lambda x}$$

They are the same!

Conclusion: The chances that the machine will work without failures another x units of time are independent of how long the machine has already worked since the last failure. The chances are the same as for a brand new machine. This property is usually called *no aging* (the machine is not getting any older, the chances of its failure are always the same), or *lack of memory* (the machine does not "remember" when it failed last time, so its chances to fail again are independent of the past history of failures).

The "no aging/memoryless" property is characteristic for exponential random variables – actually, no other continuous random variable has this property. We will not need that last fact, though.

Radioactive Decay

Machines (airplanes, cars, boats) in real life certainly deteriorate in time (get older and have "memory" of past failures and repairs), so the exponential random variable can only describe the time to failure approximately. There is, however, a natural phenomenon that is characterized by an ideal memoryless/no aging attribute. It is radioactive decay.

Radioactive atoms can explode (disintegrate) accidentally at any time. Since nothing is happening to the atom during its life, it certainly does not "remember" how long it has lived, and it cannot be getting any "older". The decay time (or the lifetime) of an atom is an exponential random variable.

The process of decay can be illustrated as follows. Suppose a piece of radioactive material contains N atoms (usually, N is huge, of order 10^{30} or so). We will look at it at regular intervals of t units of time. During the first interval of t units of time, each atom can explode with probability $\mathbb{P}(X < t) = 1 - e^{-\lambda t}$, so it will survive with probability $p = 1 - \mathbb{P}(X < t) = e^{-\lambda t}$. Hence, approximately $(1 - p)N$ atoms disintegrate during the first interval, and pN atoms survive. During the next interval of time, each surviving atom has the same chance to disintegrate, that is again $1 - p$. So, approximately $(1-p)pN$ atoms will disintegrate and p^2N atoms will survive. After k intervals of time, p^kN atoms will survive.

Another way to look at it is to wait until half of the atoms disintegrate, i.e. assume that $p = e^{-\lambda t} = 1/2$ at time t. Then, if we wait the same period of time again (t units of time), what happens? Will the other half of the atoms disintegrate? No! Actually, half of the remaining atoms will disintegrate, so only 25% of the original atoms will survive. When another t units of time elapse, only 12.5% of the original atoms will remain, etc.

Half-life

The period of time t it takes for half of the radioactive atoms to disintegrate is called *half-life*. It is denoted by $t_{1/2}$. It is characterized by the formula $e^{-\lambda t_{1/2}} = 1/2$, from which

$$\lambda t_{1/2} = \ln 2 \approx 0.693$$

This is the relation between λ and $t_{1/2}$. The value of $t_{1/2}$ is a standard technical characteristic of radioactive atoms, it can be found in reference books. Given $t_{1/2}$, one can compute λ by $\lambda = \frac{\ln 2}{t_{1/2}}$. Note that

$$\mathbb{P}(X > t_{1/2}) = 1/2, \quad \mathbb{P}(X > 2t_{1/2}) = 1/4, \quad \mathbb{P}(X > 3t_{1/2}) = 1/8, \quad \text{etc.}$$

Example

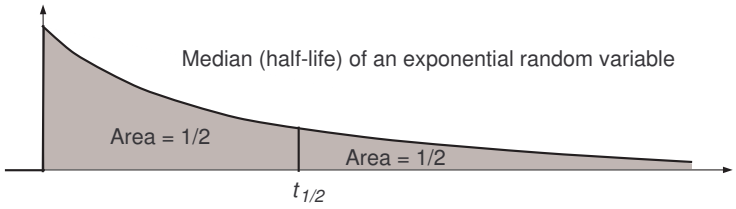
Let X be an exponential random variable with half-life $t_{1/2} = 4$. Find λ and compute $\mathbb{P}(X > 8)$ and the conditional probability $\mathbb{P}(X > 143/X > 135)$.

Solution: We have $\lambda = \frac{\ln 2}{4} \approx 0.173$. Now, $8 = 2t_{1/2}$, so $\mathbb{P}(X > 8) = (1/2)^2 = 1/4$. Next, by the lack of memory, $\mathbb{P}(X > 135 + 8/X > 135) = \mathbb{P}(X > 8) = 1/4$.

Median

Let X be an arbitrary random variable. The value of x such that $F_X(x) = 1/2$ is called the *median* of the random variable X. It is denoted by m, so that $F_X(m) = 1/2$. Note that $\mathbb{P}(X \leq m) = \mathbb{P}(X > m) = 1/2$. In this sense, m exactly divides the probability distribution of X in half.

The half-life $t_{1/2}$ is the median for the exponential random variable.



Example

New York Times reported in 1999 that the median of the prices of houses in the South of the United States was \$135,000. What does this mean?

This means that half of the houses are sold below \$135,000 and half of the houses are sold for more than \$135,000.

Percentiles (optional material)

One can characterize a probability distribution by other dividing points, which are called *percentiles*. The $(100p)$th percentile, $0 < p < 1$, is a point π_p such that

$$\mathbb{P}(X \leq \pi_p) = p \qquad \text{and} \qquad \mathbb{P}(X > \pi_p) = 1 - p$$

So, π_p is the solution of the equation $F(\pi_p) = p$.

The most important percentiles are the median, $m = \pi_{1/2}$, and the quartiles, $q_1 = \pi_{1/4}$ and $q_3 = \pi_{3/4}$ (called the first and third quartiles, respectively).

Chapter 7

Functions of Random Variables

Warning: Experience shows that many students have difficulties with this section. A careful reading of all examples is advised.

Function of a Random Variable

Let X be a random variable, and $y = g(x)$ a function. Then $Y = g(X)$ is another random variable. We will see how to find the distribution function and density function of Y, if those of X are given.

Example

Let X be uniform $U(0,1)$ and $Y = 12X - 6$. Find F_Y and f_Y.

Solution: We have

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(12X - 6 \le y) = \mathbb{P}\Big(X \le \frac{y+6}{12}\Big) = F_X\Big(\frac{y+6}{12}\Big).$$

Since X is $U(0,1)$, its distribution function is $F_X(x) = x$ according to (16). Thus we have $F_Y(y) = (y+6)/12$. Differentiating gives $f_Y(y) = 1/12$.

Of course, one needs to specify were these formulas for F_Y and f_Y are valid. One simply needs to find the values that the variable Y takes. Since $0 < X < 1$, we have $0 < 12X < 12$ and $-6 < Y < 6$. So the final answer must look like this:

$$F_Y(y) = \frac{y+6}{12} \quad \text{and} \quad f_Y(y) = \frac{1}{12} \qquad \text{for} \quad -6 < y < 6$$

Note that we found F_Y and f_Y first and then determined the range of values of the new variable Y. This is, generally, *not* a good idea.

It is advisable to find the range (all possible values) of the random variable Y first, before computing F_Y and f_y; this may simplify calculations. See examples below.

Method

Given X and $Y = g(X)$, the following method should be used to compute the distribution function F_Y and density function f_Y of Y:

(1) Find the range (interval of possible values) for the variable Y.

(2) Start with $F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\big(g(X) \le y\big)$, then solve the inequality $g(X) \le y$ for X (this is the most tricky and confusing part!).

(3) Express the resulting probability in terms of the distribution function F_X.

(4) Use the given (known) formula for F_X to obtain a formula for F_Y.

(5) Differentiate F_Y to get f_Y.

(6) Record the final answer: give formulas for F_Y and f_Y and specify the range of values where they are valid.

Example

Let X be uniform $U(-1, 1)$ and $Y = 1/(X + 1)$. Find F_Y and f_Y.

Solution: Since $-1 < X < 1$, we have $0 < X+1 < 2$ and so $\frac{1}{2} < \frac{1}{X+1} < \infty$. Now compute $F_Y(y)$ for $\frac{1}{2} < y < \infty$:

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\Big(\frac{1}{X+1} \le y\Big) = \mathbb{P}\Big(X + 1 \ge \frac{1}{y}\Big).$$

Note: the inequality $\frac{1}{X+1} \le y$ can be transformed into $X + 1 \ge \frac{1}{y}$ only because both sides are positive, i.e., $\frac{1}{X+1} > 0$ and $y > 0$ (otherwise the inequality might have been reversed – dividing or multiplying both sides by a negative number reverses the inequality sign). But how do we know that both sides are positive? It is because we have determined the range of values! We have actually established that $\frac{1}{X+1} > \frac{1}{2}$ and $y > \frac{1}{2}$.

Now recall that $F(x) = \frac{x+1}{2}$ by (15) and complete the calculation:

$$F_Y(y) = \mathbb{P}\Big(X \ge \frac{1}{y} - 1\Big) = 1 - F\Big(\frac{1}{y} - 1\Big) = 1 - \frac{1}{2}\left(\frac{1}{y} - 1 + 1\right) = 1 - \frac{1}{2y}.$$

for $\frac{1}{2} < y < \infty$. By differentiating, $f_Y(y) = \frac{1}{2y^2}$. So the final answer is:

$$F_Y(y) = 1 - \frac{1}{2y} \quad \text{and} \quad f_Y(y) = \frac{1}{2y^2} \qquad \text{for } \frac{1}{2} < y < \infty$$

Example

Let X be uniform $U(0,1)$ and $Y = -\frac{1}{\lambda}\ln(1-X)$ for some constant $\lambda > 0$. Find F_Y and f_Y.

Solution: Since $0 < X < 1$, we have $0 < 1 - X < 1$, then $-\infty < \ln(1-X) < 0$, and so $0 < Y < \infty$. Now, for all $0 < y < \infty$ we have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \le y) = \mathbb{P}\Big(-\frac{1}{\lambda}\ln(1-X) \le y\Big) = \mathbb{P}\big(\ln(1-X) \ge -\lambda y\big) \\ &= \mathbb{P}\big(1 - X \ge e^{-\lambda y}\big) = \mathbb{P}\big(X \le 1 - e^{-\lambda y}\big) = F_X\big(1 - e^{-\lambda y}\big) = 1 - e^{-\lambda y}. \end{aligned}$$

Amazingly, this is the distribution function from Chapter 6. So, Y is an exponential random variable. Now, by differentiating, $f_Y(y) = \lambda e^{-\lambda y}$.

Generating Exponential Random Variables

The last example shows how to generate an exponential random variable by a computer. Simply call a standard random number generator (RNG) that returns a value of X in the interval $(0,1)$, then compute $Y = -\frac{1}{\lambda}\ln(1-X)$. In fact, one can generate any random variable by using the RNG; see also below.

Example

Let X be exponential(λ) and $Y = \sqrt{X}$. Find F_Y and f_Y.

Solution: Since $0 < X < \infty$, we have $0 < \sqrt{X} < \infty$, then $0 < Y < \infty$. Now, for all $0 < y < \infty$ we have

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\big(\sqrt{X} \le y\big) = \mathbb{P}(X \le y^2) = F_X(y^2) = 1 - e^{-\lambda y^2}.$$

Lastly, by differentiating, $f_Y(y) = 2\lambda y e^{-\lambda y^2}$.

Example

Let X be uniform $U(-1,1)$ and $Y = X^2$. Find F_Y and f_Y.

Solution: This is somewhat tricky! First, $-1 < X < 1$, then $0 < X^2 < 1$, hence $0 < Y < 1$. So, for all $0 < y < 1$ we have

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\big(X^2 \le y\big)$$

Now, solving $X^2 \le y$ for X we need to remember that $0 < y < 1$ and $-1 < X < -1$, hence the solution is $-\sqrt{y} \le X \le \sqrt{y}$ (inspect this carefully!). Therefore,

$$\begin{aligned} F_Y(y) &= \mathbb{P}\big(-\sqrt{y} \le X \le \sqrt{y}\big) = F_X\big(\sqrt{y}\big) - F_X\big(-\sqrt{y}\big) \\ &= \frac{1+\sqrt{y}}{2} - \frac{1-\sqrt{y}}{2} = \sqrt{y} \end{aligned}$$

and $f_Y(y) = F_Y'(y) = \frac{1}{2\sqrt{y}}$.

Linear Transformation

Let X be any random variable with distribution function F_X and density f_X. Let $Y = a + bX$, where a and $b > 0$ are constants. Find F_Y and f_Y.

Solution: We have

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(a + bX \le y) = \mathbb{P}\Big(X \le \frac{y-a}{b}\Big) = F_X\Big(\frac{y-a}{b}\Big). \tag{17}$$

Note: we have used the fact that $b > 0$ when solving the inequality for X. By differentiating and using the chain rule,

$$f_Y(y) = \frac{1}{b}\, f_X\Big(\frac{y-a}{b}\Big). \tag{18}$$

Fahrenheit vs Celsius

The variable $Y = aX + b$ is called a linear transformation of X. It is simply the rescaling and shifting of the values of X. Such transformations are common in practice. For example, if X is the temperature in Celsius, then $Y = 1.8X + 32$ is the temperature in Fahrenheit.

Special Example (optional material)

Let X be an arbitrary continuous random variable with distribution function F_X. Find the distribution function of $Y = F_X(X)$.

Solution: First, recall that $0 \le F_X(X) \le 1$ for any distribution function. Now we have for $0 \le y \le 1$

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\big(F_X(X) \le y\big).$$

Solving $F_X(X) \le y$ gives $X \le F_X^{-1}(y)$, where F_X^{-1} denotes the inverse function. Then

$$F_Y(y) = \mathbb{P}\big(X \le F_X^{-1}(y)\big) = F_X\big(F_X^{-1}(y)\big) = y$$

for $0 < y < 1$. Hence, Y is a uniform random variable on the interval $(0, 1)$, i.e. $U(0, 1)$. Note: one more special appearance of $U(0, 1)$!

Generating continuous random variables by computer. Let X be an arbitrary continuous random variable. According to the above, the variable $Y = F_X(X)$ is $U(0, 1)$. In other words, if Y is $U(0, 1)$, then $X = F_X^{-1}(Y)$ has the distribution function F_X. This is the basis for generating (by computer) any continuous random variable: call an RNG to get a number Y between 0 and 1, then compute $X = F_X^{-1}(Y)$. Practically, this amounts to solving the equation $Y = F_X(X)$ for X. If the formula for F_X is simple, the exact solution can be found by algebra. If F_X is complicated, one can find an approximate solution by using numerical algorithms.

Chapter 8

Normal Random Variables

Normal random variables (also called *Gaussian random variables*) are the most important random variables in probability theory. We first introduce one of them – the *standard normal random variable*.

Standard Normal Random Variable

This random variable is usually denoted by Z. Its density function is

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad \text{for all} \quad -\infty < x < \infty$$

Note that $f_Z(x)$ is positive for all $-\infty < x < \infty$, hence Z takes on all real values, it does not have a minimum or a maximum, its range is the entire real line. Also note that $f_Z(x)$ is an even function, i.e. $f_Z(x) = f_Z(-x)$.

The graph of $f_Z(x)$ is a bell-shaped curve (see the next page), symmetric about the y-axis. This curve is called *gaussian curve*. Its maximum is attained at $x = 0$, then it decreases on both sides of its top point. Actually, it decreases very fast. One can easily check that

$$\begin{aligned}
f_Z(0) &\approx 0.399, \\
f_Z(1) &\approx 0.242, \\
f_Z(2) &\approx 0.054, \\
f_Z(3) &\approx 0.0044, \\
f_Z(4) &\approx 0.00013, \\
f_Z(5) &\approx 0.000001, \\
f_Z(6) &\approx 0.000000006 \ldots
\end{aligned}$$

For larger x, the function $f_Z(x)$ keeps decreasing at a dramatic rate. For all practical purposes, one can think that $f_Z(x)$ vanishes for all $|x| > 6$.

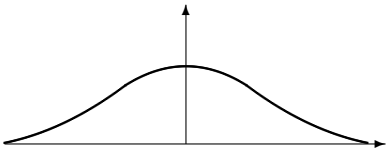
Gaussian random variables are named after
Carl Friedrich Gauss (1777-1855)
who is regarded by many as the greatest
mathematician of all times

The Φ Function

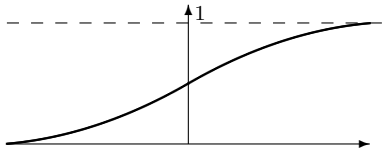
The distribution function of Z is denoted by $\Phi(x)$. According to general formula (14) on page 38

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du$$

Unfortunately, there is no simpler formula for $\Phi(x)$ (the above integral cannot be expressed in terms of elementary functions). The above integral formula is the best one can write for $\Phi(x)$.



Density function $f_Z(x)$



Distribution function $\Phi(x)$

Since there is no convenient formula for $\Phi(x)$, the values of this function cannot be easily computed. We will use a table on a separate page to find the values of the function $\Phi(x)$.

Table of $\Phi(x)$

Let us learn how to use the table. For $0 \le x \le 3.99$ this is obvious. For negative x, specifically for $-3.99 \le x \le 0$, we can use the symmetry rule:

$$\Phi(-x) = 1 - \Phi(x) \qquad \text{for all} \quad x > 0.$$

This rule follows from the symmetry of the density function $f_Z(x)$ about zero: $\Phi(-x) = \mathbb{P}(Z < -x) = \mathbb{P}(Z > x) = 1 - \Phi(x)$.

Finally, for all $x > 3.99$ we will simply set $\Phi(x) = 1$ and for all $x < -3.09$ we will set $\Phi(x) = 0$. It is clear from the end of Table for $\Phi(x)$ that this is accurate, up to four digits after the decimal point.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Examples

Compute $\Phi(1)$, $\Phi(2.36)$, $\Phi(-1.25)$, $\Phi(4.7)$.

Answers:

$$\Phi(1) = 0.8413 \qquad \text{(from the table)}$$

$$\Phi(2.36) = 0.9909 \qquad \text{(from the table)}$$

$$\Phi(-1.25) = 1 - \Phi(1.25) = 1 - 0.8943 = 0.1057$$

(by using the symmetry)

$$\Phi(4.7) = 1 \qquad \text{(because 4.77 exceeds 3.99)}$$

Examples

Compute $\mathbb{P}(Z < 2.87)$, $\mathbb{P}(Z > 0.76)$, $\mathbb{P}(Z < -0.76)$, $\mathbb{P}(Z > -2)$, $\mathbb{P}(-0.6 < Z < 1.3)$, $\mathbb{P}(|Z| < 2)$, $\mathbb{P}(|Z| < 3)$, $\mathbb{P}(|Z| > 4)$.

Solution. We have

$$\mathbb{P}(Z < 2.87) = \Phi(2.87) = 0.9979$$

$$\mathbb{P}(Z > 0.76) = 1 - \Phi(0.76) = 1 - 0.7764 = 0.2236$$

$$\mathbb{P}(Z < -0.76) = \Phi(-0.76) = 1 - \Phi(0.76) = 0.2236$$

$$\mathbb{P}(Z > -2) = 1 - \Phi(-2) = \Phi(2) = 0.9772$$

$$\begin{aligned}\mathbb{P}(-0.6 < Z < 1.3) &= \Phi(1.3) - \Phi(-0.6) = \Phi(1.3) - 1 + \Phi(0.6) \\ &= 0.9032 - 1 + 0.7257 = 0.6289\end{aligned}$$

$$\mathbb{P}(|Z| < 2) = \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 = 2 \times 0.9772 - 1 = 0.9544$$

$$\mathbb{P}(|Z| < 3) = \Phi(3) - \Phi(-3) = 2\Phi(3) - 1 = 2 \times 0.9986 - 1 = 0.9972$$

$$\mathbb{P}(|Z| > 4) = 1 - \mathbb{P}(|Z| < 4) = 1 - (2\Phi(4) - 1) = 0$$

Normal Random Variables

Now we are ready to introduce a (general) normal random variable. It has two parameters: μ and $\sigma > 0$. It can be defined in terms of the standard normal random variable Z by

$$Y = \mu + \sigma Z.$$

In other words, Y is obtained by rescaling and shifting (multiplying by σ and adding μ) of the standard normal random variable Z. The normal random variable Y is denoted by $\mathcal{N}(\mu, \sigma^2)$, i.e. we say that Y is $\mathcal{N}(\mu, \sigma^2)$. Note also that Z is $\mathcal{N}(0, 1)$, is this notation.

Density and Distribution Function of $\mathcal{N}(\mu, \sigma^2)$

By (18) on page 52, the density of $Y = \mathcal{N}(\mu, \sigma^2)$ is

$$f_Y(x) = \frac{1}{\sigma} f_Z\Big(\frac{x-\mu}{\sigma}\Big) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

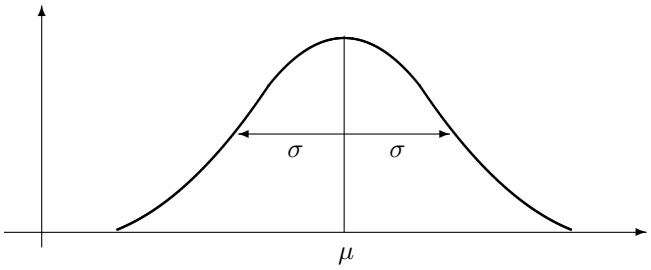
and by (17) on page 52, the distribution function is

$$F_Y(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

for all $-\infty < x < \infty$.

Note: the density function $f_Y(x)$ is positive for all x, has a peak at $x = \mu$ and goes down on both sides of its peak. It is a bell-shaped curve, just like $f_Z(x)$, but it is shifted so that its center is at the point μ. The other parameter, σ affects the shape of the curve: for smaller σ, the peak is taller and thinner, for larger σ the peak is shorter (lower) and thicker (wider). See illustration on the next page.

Remember that, in any case, the total area under the graph of $f_Y(x)$ is the same, it is equal to one by the normalization rule on page 38.



Density function $f_Y(x)$ of a normal r.v. $Y = \mathcal{N}(\mu, \sigma^2)$

Example

Let Y be $\mathcal{N}(5, 4)$. Compute $\mathbb{P}(Y < 7)$ and $\mathbb{P}(3 < Y < 6)$.

Solution: We have $\mu = 5$ and $\sigma^2 = 4$, hence $\sigma = 2$. Now

$$\mathbb{P}(Y < 7) = F_Y(7) = \Phi\left(\frac{7-5}{2}\right) = \Phi(1) = 0.8413$$

and

$$\begin{aligned}\mathbb{P}(3 < Y < 6) &= F_Y(6) - F_Y(3) \\ &= \Phi\left(\frac{6-5}{2}\right) - \Phi\left(\frac{3-5}{2}\right) \\ &= \Phi(0.5) - \Phi(-1) \\ &= 0.6915 - 1 + 0.8413 = 0.5328.\end{aligned}$$

Manipulations with a Normal Random Variable

Let $Y = \mathcal{N}(\mu, \sigma^2)$ be a normal random variable. What can we say about $W = a + bY$, if a and b are some constants? It turns out that W is also a normal random variable

$$W = \mathcal{N}(a + b\mu, b^2\sigma^2)$$

Indeed, $Y = \mu + \sigma Z$ where Z is a standard normal random variable, so

$$W = a + b(\mu + \sigma Z) = \underbrace{a + b\mu}_{\text{new } \mu} + \underbrace{b\sigma}_{\text{new } \sigma} Z.$$

Note: If Y is $\mathcal{N}(\mu, \sigma^2)$, then the variable $W = -Y$ is normal $\mathcal{N}(-\mu, \sigma^2)$ (this follows from the above rule with $a = 0$ and $b = -1$).

Rule of Three Sigmas

We have seen that $\mathbb{P}(|Z| < 3) = 99.72\%$. Now, for any normal random variable $Y = \mathcal{N}(\mu, \sigma^2)$ we have

$$\mathbb{P}(\mu - 3\sigma < Y < \mu + 3\sigma) = \mathbb{P}(-3 < Z < 3) = 99.72\%.$$

Hence, it is almost certain that Y takes values in the interval $(\mu-3\sigma, \mu+3\sigma)$. In many practical applications one takes it for granted that the normal random variable **must** be within the distance 3σ from μ. This is known as the "rule of 3σ", or the "three-sigma rule".

Rule of Two Sigmas

In statistics, on the other hand, a "rule of two sigmas" is popular: values of a normal random variable $\mathcal{N}(\mu, \sigma^2)$ in the interval between $\mu - 2\sigma$ and $\mu+2\sigma$ are called 'typical' or 'usual'. Values beyond this interval are regarded as 'unusual'. We have seen that $\mathbb{P}(|Z| < 3) = 95\%$. Hence 95% of values are 'usual' and 5% are 'unusual'.

Standard Normal Variable Squared

Find the distribution and density function of $W = Z^2$.

Solution. This is similar to an example on page 51. Obviously, $W > 0$. Now we have, for all $x > 0$,

$$F_W(x) = \mathbb{P}(W \le x) = \mathbb{P}(Z^2 \le x) = \mathbb{P}(-\sqrt{x} < Z < \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}).$$

By differentiating and using the chain rule we get

$$f_W(x) = \frac{1}{2\sqrt{x}}\, f_Z(\sqrt{x}) + \frac{1}{2\sqrt{x}}\, f_Z(-\sqrt{x}) = \frac{1}{\sqrt{x}}\, f_Z(\sqrt{x}) = \frac{1}{\sqrt{2\pi x}}\, e^{-\frac{x}{2}}$$

for $x > 0$. The random variable $W = Z^2$ is called a χ^2 variable with one degree of freedom (it is used in statistics).

Role of Normal Random Variables

The importance of normal random variables will be demonstrated later, in Chapter 15. Right now we can just say that many random variables in practical applications are normal or approximately normal.

For instance, let X be the height (or weight) of a randomly selected adult male in a large population (city, state, nation). Naturally, most of the heights (weights) of adult men are grouped near the statistical average, but there are some that are farther away from the average – and the farther away the fewer of them can be found. Plotting the density function will give something close to a bell-shaped curve, which represents a normal distribution.

Error Function (optional material)

In some older textbooks and physical and engineering applications, another function is used instead of $\Phi(x)$. It is called the *error function* and given by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2}\, dy$$

To find the relation between $\operatorname{erf}(x)$ and $\Phi(x)$, one can change variable $u = \sqrt{2}y$ in the expression for $\Phi(x)$ and arrive at

$$\Phi(x) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right).$$

This is a useful conversion formula.

Approximations to $\Phi(x)$

In some very precise calculations, one needs accurate values of $\Phi(x)$ for $x > 4$. The following formula gives a very good approximation:

$$\Phi(x) \approx 1 - \frac{1}{x\sqrt{2\pi}}\, e^{-x^2/2}.$$

For example, $\Phi(5) \approx 1 - 7.45 \times 10^{-7} = 0.999999255$.

Normal Random Variables in Physics (optional)

A gas (the air in the room, for example) consists of billions and billions of molecules (something like 10^{25} or 10^{30} molecules). They move all the time, at various speed and in various directions. If we pick one molecule at random, then its velocity vector $\mathbf{v} = (v_x, v_y, v_z)$ will be a random vector, its components v_x, v_y, v_z will be random variables. They have the same distribution, since there is apparently no difference between the x- y- and z-direction in a homogeneous gas. Moreover, if one rotates the coordinate frame (i.e. redirects the coordinate axes), then the new components v_x, v_y, v_z, even though measured differently, will have the same distribution. Another law of physics says that v_x, v_y, v_z must be independent – this seems intuitively quite reasonable. So let us fix these features: the components v_x, v_y, v_z must be independent from each other and their distribution must be the same in any coordinate system. It turns out, quite surprisingly, that the *only* distribution with these two features is normal! This is called the *Maxwell law* in physics.

Summary

The following chart represents all basic types of continuous random variables:

	density $f(x)$	distribution $F(x)$	range
uniform $U(0,1)$	1	x	$0 < x < 1$
uniform $U(a,b)$	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$a < x < b$
exponential(λ)	$\lambda e^{-\lambda x}$	$1 - e^{-\lambda x}$	$x > 0$
st.normal $\mathcal{N}(0,1)$	$\frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$	$\Phi(x)$	$-\infty < x < \infty$
normal $\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$	$-\infty < x < \infty$

Chapter 9

Joint Distributions

Here we deal with situations that involve two (or more) random variables.

Example

Let X be a discrete random variable that takes values 0, 1, 2 with the following probabilities:

values of X	0	1	2
probabilities	0.2	0.5	0.3

Let Y be another discrete random variable that takes values -1, 0, 2 with the following probabilities:

values of Y	-1	0	2
probabilities	0.1	0.4	0.5

Assume that X and Y are independent.

Which pairs of values (X, Y) are possible? What are their probabilities? Find $\mathbb{P}(X = Y)$. Find $\mathbb{P}(X < Y)$.

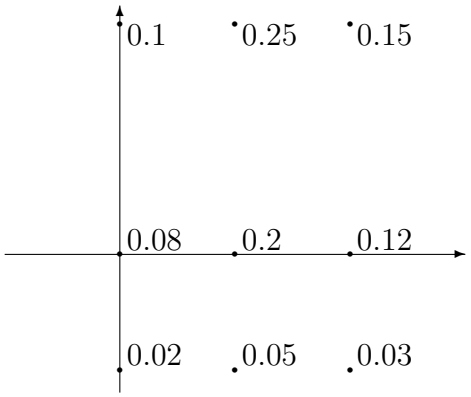
Solution. Possible pairs of values are $(0, -1)$, $(0, 0)$, $(0, 2)$, $(1, -1), \ldots, (2, 2)$. The probabilities are computed by the multiplication rule

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

which applies because of independence. The following table lists all the pairs with the corresponding probabilities

(x, y)	(0,-1)	(0,0)	(0,2)	(1,-1)	(1,0)	(1,2)	(2,-1)	(2,0)	(2,2)
prob.	0.02	0.08	0.1	0.05	0.2	0.25	0.03	0.12	0.15

One can also mark all the pairs (x, y) as points on the xy plane and write the probability next to each point. This completely characterizes the distribution of the pair of random variables, which is called the *joint distribution* of X and Y.



Now, the event $\{X = Y\}$ contains all the points on the diagonal $x = y$. In our example it contains two points: $(0, 0)$ and $(2, 2)$. Hence, $\mathbb{P}(X = Y) = 0.08 + 0.15 = 0.23$.

The event $\{X < Y\}$ contains all the points *above* the diagonal $x = y$. In our example it contains two points: $(0, 2)$ and $(1, 2)$. Hence, $\mathbb{P}(X < Y) = 0.1 + 0.25 = 0.35$.

Note: a similar table of pairs of values we had in Example 1.15 (rolling two dice). In that example, we had $6 \times 6 = 36$ pairs, each taken with probability $1/36$.

Discrete Pair of Random Variables

A discrete pair of random variables X, Y can be characterized by the list of all possible pairs of values, with the corresponding probabilities.

Joint Distribution Function

Any pair of random variables X, Y can be characterized by a *joint distribution function*. This is a function of two variables, $F(x, y)$, defined by

$$F_{X,Y}(x, y) = \mathbb{P}(X \le x \text{ and } Y \le y).$$

Here X, Y denote the random variables, and x, y are the arguments of the function.

Example (continued)

Here are some values of the joint distribution function in the previous example:

- $F(1.1, 0.8) = 0.35$ (the quadrant to the left and below the point $(1.1, 0.8)$ covers four pairs of (X, Y), with the total probability of 0.35)
- $F(5, -0.4) = 0.1$ (the quadrant to the left and below the point $(5, -0.4)$ covers three pairs of (X, Y), with the total probability of 0.1)
- $F(4, 7) = 1$, $F(-2, 8) = 0$, etc.

We will not attempt to describe this function completely, it is not of much use in this example.

Example

Let $X = U(0, 1)$ and $Y = U(0, 1)$ be two uniform random variables that are independent. Find the joint distribution function $F_{X,Y}(x, y)$.

Solution: Because of independence, we can use the multiplication rule:

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbb{P}(X \le x \text{ and } Y \le y) \\ &= \mathbb{P}(X \le x) \cdot \mathbb{P}(Y \le y) = F_X(x) F_Y(y). \end{aligned}$$

We know that $F_X = x$ for $0 < x < 1$ and $F_Y(y) = y$ for $0 < y < 1$, by (16) on page 43. Hence, $F_{X,Y}(x, y) = xy$ for all $0 < x, y < 1$. For other values of x, y the function $F_{X,Y}(x, y)$ is not interesting, because those values are not taken by the pair of random variables X, Y.

Joint Density Function

The joint density function $f_{X,Y}(x, y)$ of a pair of random variables is

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \, \partial y}.$$

This is the second order (mixed) partial derivative of $F_{X,Y}$ with respect to x and y.

Note: the students who have not taken Calculus III, should not worry too much. We will use elements of multivariate calculus (partial derivatives and double integrals) only of the simplest forms.

Previous Example (continued)

Since $F_{X,Y}(x, y) = xy$, we have $f_{X,Y}(x, y) = 1$ for $0 < x, y < 1$ (and zero elsewhere, since other values of x, y are not taken by the pair X, Y).

Computation of Probabilities

For any region R in the xy plane

$$\mathbb{P}\big\{(X, Y) \text{ is in region } R\big\} = \iint_R f_{X,Y}(x, y) \, dx \, dy$$

This is a double integral of the function $f_{X,Y}$ over the region R.

Rule for Constant Density Functions

Let the joint density function be constant: $f_{X,Y}(x, y) = c$ over the region R (as in the previous example, where $f(x, y) = 1$ over the unit square). Then the above double integral simply equals c times the area of R. Hence,

$$\mathbb{P}\big\{(X, Y) \text{ is in region } R\big\} = c \cdot \text{Area}(R)$$

In all our examples and test problems in MA 485 involving the computation of probabilities, the joint density function will be constant. So, all our double integrals can be computed by this simple rule.

Previous Example (continued)

(a) Find the probability $\mathbb{P}(X + Y < 1)$.

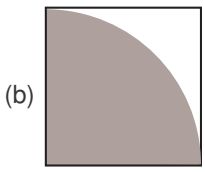
Solution: The part of the region $\{x + y < 1\}$ that lies within the unit square $0 < x, y < 1$ is the left lower triangle, i.e. half of the square. Its area is $1/2$, hence $\mathbb{P}(X + Y < 1) = 1/2$.

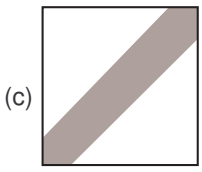
(b) Find $\mathbb{P}(X^2 + Y^2 < 1)$.

Solution: The region $\{X^2 + Y^2 < 1\}$ is the unit circle. Within the unit square $0 < x, y < 1$, it makes just a quarter of the circle, so its area is $\pi/4$. Hence, $\mathbb{P}(X^2 + Y^2 < 1) = \pi/4$.

(c) Find $\mathbb{P}\big(|X - Y| < 0.1\big)$.

Solution: The region $\{|X - Y| < 0.1\}$ is a strip around the diagonal line $y = x$. Within the unit square $0 < x, y < 1$, it stretches from the bottom left cornet to the top right corner. To find its area, it is convenient to subtract the total area of the two remaining triangles from the area of the square. Hence, the area of the strip is $1 - (0.9)^2 = 0.19$, so $\mathbb{P}(|X - Y| < 0.1) = 0.19$.





Using Probabilities to Compute $\pi = 3.14159\ldots$

The problem (b) above suggests a method of determining the number π to any precision (at least theoretically). One can generate pairs of random numbers (x, y) by a random number generator, every time check the condition $x^2 + y^2 < 1$, and in the end the fraction of pairs satisfying this condition gives you the number $\pi/4$. In Chapter 15 we will learn how many pairs of random numbers one needs to generate if one wants to obtain k correct digits of the number $\pi/4$.

Example

Let X, Y have the joint density function $f(x, y) = 2$ for $0 < y < x < 1$. Find $\mathbb{P}(X - Y > 0.4)$. See illustration on the previous page.

Solution: Note that the density is constant (=2) over the triangle $0 < y < x < 1$ (and, by default, $f(x, y) = 0$ elsewhere). The region $x - y > 0.4$ makes a smaller triangle within it, see illustration. The area of the smaller triangle is $\frac{1}{2}(0.6)^2 = 0.18$. Hence, $\mathbb{P}(X - Y > 0.4) = 2 \times 0.18 = 0.36$.

General Advice

In examples like above it is advisable to sketch the region where $f(x, y) \neq 0$, and, within it, the subregion corresponding to the given event.

Strange Example (optional material)

Let X be a uniform random variable on $(0, 1)$, i.e. $X = U(0, 1)$, and $Y = X^2$. Describe the distribution of the pair X, Y.

Solution: Since $Y = X^2$, all possible pairs of X, Y lie on the parabola $y = x^2$, more precisely on the stretch of it from the point $(0, 0)$ to the point $(1, 1)$. This is not a discrete pair of random variables, since for every single point (x, y) we have $\mathbb{P}(X = x \text{ and } Y = y) = 0$. On the other hand, it does not have a joint density function. This is a very unusual example, and we will not study it in detail.

Multiplication Rule for Independent Random Variables

Let X and Y be two independent random variables. Then we have simple multiplication rule

$$F_{X,Y}(x, y) = F_X(x)\, F_Y(y)$$

for distribution functions and

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$$

for density functions (if those exist).

The first rule is already explained in one of the previous examples, the second follows by differentiation.

Three and More Random Variables

Let $X_1, \ldots, X_n$ be n random variables. One can also call $(X_1, \ldots, X_n)$ a random vector with n components. In the same way as above, we can define the joint distribution function and the joint density function for the variables $X_1, \ldots, X_n$. The previous multiplication rule works for any number of independent random variables.

Min/max of Two Random Variables

Let X and Y be two independent random variables, and F_X and F_Y their distribution functions. Let $V = \max\{X, Y\}$ and $W = \min\{X, Y\}$. Find the distribution functions of V and W.

Solution: Note that $V \leq x$ whenever both $X \leq x$ and $Y \leq x$. Hence,

$$F_V(x) = \mathbb{P}(V \leq x) = \mathbb{P}(X \leq x, Y \leq x) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq x) = F_X(x)F_Y(x)$$

Similarly, note that $W > x$ whenever both $X > x$ and $Y > x$. Hence,

$$F_W(x) = 1 - \mathbb{P}(W > x) = 1 - \mathbb{P}(X > x)\mathbb{P}(Y > x) = 1 - (1 - F_X(x))(1 - F_Y(x))$$

Special case: If X and Y have the *same* distribution function F, then

$$F_V(x) = F^2(x) \quad \text{and} \quad F_W(x) = 1 - (1 - F(x))^2.$$

Independent Identically Distributed (i.i.d.) Random Variables

Let $X_1, \ldots, X_n$ be independent random variables that have the same distribution function F. In this case we call them *independent identically distributed (i.i.d.)* random variables. Examples: tossing a coin n times or rolling a die n times produces a sequence of n results (numbers). These results are independent and have the same probability distribution. Whenever the same experiment is repeated n times independently, and each time one records a numerical output, one gets a sequence of i.i.d. random variables. This sort of situation is the most basic and most common in probability theory.

Min/max of n i.i.d. Random Variables

Let $X_1, \ldots, X_n$ be i.i.d. random variables. Let $V = \max\{X_1, \ldots, X_n\}$ and $W = \min\{X_1, \ldots, X_n\}$. Find the distribution functions of V and W.

Solution: Very much like in the case of two variables, we obtain

$$F_V(x) = F^n(x) \quad \text{and} \quad F_W(x) = 1 - (1 - F(x))^n$$

where F is the common distribution function of the variables $X_1, \ldots, X_n$. If the density $f(x) = F'(x)$ exists, then V and W also have density functions. They can be found by differentiation and the chain rule:

$$f_V(x) = F_V'(x) = nF^{n-1}(x)F'(x) = nF^{n-1}(x)f(x)$$

and similarly

$$f_W(x) = n(1 - F(x))^{n-1}f(x).$$

Example

Let $X_1, \ldots, X_n$ be i.i.d. random variables, each of them being uniform on $(0, 1)$, i.e., $U(0, 1)$. Find the distribution and density functions of $V = \max\{X_1, \ldots, X_n\}$ and $W = \min\{X_1, \ldots, X_n\}$.

Solution: The common distribution function of $X_1, \ldots, X_n$ is $F(x) = x$ (for $0 < x < 1$), and the common density function is $f(x) = 1$. Hence,

$$F_V(x) = x^n \quad \text{and} \quad f_V(x) = nx^{n-1}$$

for $0 < x < 1$. Also,

$$F_W(x) = 1 - (1 - x)^n \quad \text{and} \quad f_W(x) = n(1 - x)^{n-1}$$

for $0 < x < 1$.

If we graph the density functions $f_V(x)$ and $f_W(x)$, we will see that $f_V(x)$ has a tall peak at $x = 1$ and is very low near $x = 0$. On the contrary, $f_W(x)$ has a tall peak at $x = 0$ and is very low near $x = 1$.

This is due to the fact that V, the maximum of $X_1, \ldots, X_n$, most likely takes values close to 1. On the contrary, W, the minimum of $X_1, \ldots, X_n$, most likely takes values close to 0.

Time to Failure of a Multicomponent System

A system consists of n identical components which may fail independently of each other. Denote by X_i, $1 \le i \le n$, the lifetime (time to failure) of the ith component. Then $X_1 \ldots, X_n$ are independent random variables with a common distribution function $F(x)$. Let T be the lifetime (time to failure) of the entire system and $F_T(x)$ its distribution function.

Here we consider two types of systems. One uses connection of components "in series". This type of system is *fragile*, it works only if all the components work, so that $T_{\text{fragile}} = \min\{X_1, \ldots, X_n\}$. The other uses connection "in parallel". That type of system is *robust*, it works if at least one component is functioning, so that $T_{\text{robust}} = \max\{X_1, \ldots, X_n\}$.

According formulas on the previous page, we have

$$F_{T_{\text{fragile}}}(x) = 1 - (1 - F(x))^n \quad \text{and} \quad F_{T_{\text{robust}}}(x) = F^n(x).$$

Example

Let the lifetime of each component be an exponential random variable with parameter λ. Find the distribution of the lifetime of the fragile and robust systems of n components.

Solution: Since $F(x) = 1 - e^{-\lambda x}$, we have

$$F_{T_{\text{fragile}}}(x) = 1 - e^{-n\lambda x} \quad \text{and} \quad F_{T_{\text{robust}}}(x) = (1 - e^{-\lambda x})^n.$$

Note that the lifetime of the fragile system is itself an exponential random variable with parameter $n\lambda$.

More on Multicomponent Systems (optional material)

Some systems are between fragile and robust: they require at least k components working, where $1 < k < n$. Let T be the lifetime of such a system. Its distribution function is $F_T(x) = \mathbb{P}(T < x)$. Note that the event $T < x$ occurs whenever, by the time x, less than k components survive (i.e., more than $n - k$ die). For each component, the probability of survival is $p = \mathbb{P}(X_i > x) = 1 - F(x)$, and the probability of failure (or dying) by the time x is $q = 1 - p = F(x)$. Now, we have n independent components, each can survive with probability p or die with probability q. The number of survivors, call it Y, is then a binomial random variable, $Y = b(n, p)$. Recall that the event $T < x$ occurs whenever $Y \le k - 1$, hence

$$F_T(x) = \mathbb{P}(T < x) = \mathbb{P}\big(b(n,p) \le k-1\big) = \sum_{i=0}^{k-1} C_{n,i} p^i q^{n-i}.$$

Remembering that $p = 1 - F(x)$ and $q = F(x)$ we can write

$$F_T(x) = \sum_{i=0}^{k-1} C_{n,i} \big[1 - F(x)\big]^i \big[F(x)\big]^{n-i}.$$

This formula is convenient if k is small, $k < n/2$. If $k > n/2$, it is easier use the "complement" formula

$$F_T(x) = 1 - \sum_{i=k}^{n} C_{n,i} \big[1 - F(x)\big]^i \big[F(x)\big]^{n-i}.$$

Example (optional)

Let a system consist of 6 components whose lifetime is a uniform random variable on the interval $(0, 20)$. Suppose the system requires 3 working components to be operational. Find the distribution of its lifetime.

Solution: We have $F(x) = x/20$ for $0 < x < 20$. Then

$$\begin{aligned} F_T(x) &= C_{6,0}(x/20)^6 + C_{6,1}(1 - x/20)(x/20)^5 + C_{6,2}(1 - x/20)^2(x/20)^4 \\ &= (x/20)^6 + 6(1 - x/20)(x/20)^5 + 15(1 - x/20)^2(x/20)^4. \end{aligned}$$

By differentiating, we find the density function

$$f_T(x) = F_T'(x) = 3(1 - x/20)^2(x/20)^3.$$

Remark on Density Function (optional)

Differentiating in the above example shows that all the terms but the last one remarkably cancel out. This is not coincidental. It is a general rule: differentiating the function $F_T(x)$ leads to the cancelation of all the terms but the last one and gives

$$f_T(x) = nC_{n-1,k-1}\big[1 - F(x)\big]^{k-1}\big[F(x)\big]^{n-k} f(x),$$

where $f(x) = F'(x)$ is the density of $F(x)$.

Order Statistics (optional)

Let $X_1, \dots, X_n$ be i.i.d. random variables. Their values, sorted in an increasing order, are called *order statistics* and denoted by

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}.$$

Note that $X_{(1)}$ is always the smallest of X_i's, $X_{(2)}$ is the second smallest, etc. Similarly, $X_{(n)}$ is the largest of X_i's, etc.

If $X_1, \dots, X_n$ are the lifetimes of the n components in a system that requires k working components, then the lifetime of the system is $T = X_{(n-k+1)}$. For the fragile system, $k = n$, and $T_{\text{fragile}} = X_{(1)}$, for the robust system, $k = 1$, and $T_{\text{robust}} = X_{(n)}$.

Chapter 10

Mean Value

Spinning Roulette

A roulette wheel has 18 black spots, 18 red spots and 2 green spots. You can bet \$1 on black or red and win \$1 if that color comes up or lose \$1 if not (a green spot is always a casino win, so all the gamblers lose). Whether you bet on black or on red, your chance of winning is 18/38. If you play 100 times, how much do you expect to win (or lose)?

Solution: It is fair to expect that you win 18 times in 38 plays. So, in 100 plays you expect to win $100 \times 18/38 \approx 47.37$ times and to lose $100 - 47.37 = 52.63$ times. Then your net expected gain is $47.37 - 52.63 = -5.26$, i.e. you expect to lose \$5.26 in 100 plays (of course, you should expect to lose in a casino, not to win!). Your expected loss per play is 5.26/100=0.0526, a little more than 5 cents.

Rolling Die

You roll a die 100 times and add up the numbers it shows. How much do you expect to get, in the end?

Solution: The die shows the numbers 1,2,3,4,5,6 with the same probability, 1/6. The average of these numbers is $(1 + \cdots + 6)/6 = 3.5$. In 100 rolls, you then expect to accumulate $100 \times 3.5 = 350$ total. Note that now you expect to get, approximately, 3.5 points per roll.

Concept of Mean Value

In the above examples, we computed the *expected* values of some random variables, trying to be as fair as possible. Of course, the actual values of those random variables may be different: in the roulette example, you can win as much as \$100 or lose as much as \$100, and in any case you gain or loss is a whole number of dollars, it can never be \$5.26. So, what is \$5.26? It is the most fair estimate of your loss, it is what you lose "on the average". This is what we call the *mean value*, or the *expected value* of the random variable.

Old Textbook Version of Mean Value

Another way to look at the mean value is this: if you observe values of a random variable X over and over, for a long time, and compute their average, then you should get the mean (expected) value.

More precisely, let $x_1, \ldots, x_n$ be the values of X observed (or obtained) empirically. Then their average $(x_1 + \cdots + x_n)/n$ should be approximately the mean value of X. Even more precisely, the (empirical) average $(x_1 + \cdots + x_n)/n$ *approaches* the (theoretical) mean value in the limit, as n increases, i.e., as $n \to \infty$.

While this rule does not work when n is small, it is quite precise for large n's (of order of thousands or millions). Then the empirical average is practically indistinguishable from the theoretical mean value.

The above description of the mean value was actually adopted in old textbooks (prior to 1930s) as an official definition of the mean value of a random variable. Now we have more elegant formulas, see below.

Mean Value for Discrete Random Variables

If X is a discrete random variable that takes values $x_1, x_2, \ldots$ with corresponding probabilities $p_1, p_2, \ldots$, then its *mean value* is

$$\mathbb{E}(X) = x_1 p_1 + x_2 p_2 + \cdots$$

Mean value is also called *expectation* of a random variable. This explains the symbol $\mathbb{E}$ in the formula.

Note: in the Rolling Die example on the previous page we had

$$\frac{1 + \cdots + 6}{6} = 1 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6}$$

so our calculations were consistent with the above rule.

Mean Value for Uniform Discrete Random Variables

Recall that a uniform discrete random variable X takes values $1, \ldots, n$, each with the same probability $1/n$ (page 29). Thus

$$\mathbb{E}(X) = \frac{1 + \cdots + n}{n} = \frac{n(n+1)}{2} \cdot \frac{1}{n} = \frac{n+1}{2}$$

Note: it is the average of the very first and the very last values (1 and n).

Mean Value for Geometric Random Variables

A geometric random variable X takes values $n = 1, 2, \ldots$ with probabilities $\mathbb{P}(X = n) = pq^{n-1}$ (page 28). Its mean value is

$$\mathbb{E}(X) = 1 \cdot p + 2 \cdot pq + 3 \cdot pq^2 + 4 \cdot pq^3 + \cdots$$

We use the following trick to compute it:

$$\begin{aligned} \mathbb{E}(X) = p + pq + pq^2 &+ pq^3 + \cdots \\ +pq + pq^2 &+ pq^3 + \cdots \\ +pq^2 &+ pq^3 + \cdots \\ &+ pq^3 + \cdots \end{aligned}$$

The first row consists on probabilities that sum to one. In other rows we need to factor out q, q^2, etc., to get the same sum as in the first row. Hence, we obtain

$$\mathbb{E}(X) = 1 + q + q^2 + q^3 + \cdots = \frac{1}{1-q} = \frac{1}{p}$$

Here we used the Calculus rule for geometric series, as we did in our formula (11) back on page 28.

Mean Value for Binomial Random Variables

Binomial random variable $X = b(n,p)$ takes values $k = 0, 1, \ldots, n$ with probabilities $\mathbb{P}(X = k) = C_{n,k}p^k q^{n-k}$ (page 27). Then

$$\mathbb{E}(X) = \sum_{k=0}^{n} k \, \frac{n!}{k!(n-k)!} \, p^k q^{n-k} = np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} \, p^{k-1} q^{n-k}$$

The last sum equals one, because it contains the probabilities of the random variable $b(n-1,p)$ (this can be seen more easily if we change variables $n' = n-1$ and $k' = k-1$). Hence, $\mathbb{E}(X) = np$.

Mean Value for Poisson Random Variables

Poisson random variable X=poisson(λ) takes values $k = 0, 1, 2, \ldots$ with probabilities $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ (page 32). Then

$$\begin{aligned} \mathbb{E}(X) &= \sum_{k=0}^{\infty} k \cdot \mathbb{P}(X = k) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \end{aligned}$$

The last sum equals one, because it contains the probabilities of the same random variable X (this can be seen more easily if we change variables $k' = k-1$). Hence, $\mathbb{E}(X) = \lambda$.

Matching Averages

On page 31, we already remarked that $\lambda = np$ had an intuitively clear meaning of being the *average* number of successes. Now we see that np, is, indeed, the average (mean value) of $b(n,p)$ and λ is, indeed, the average (mean value) of poisson(λ). Thus we can say that our approximation of a binomial random variable by a Poisson random variable in Chapter 4 is based on "matching" their mean values: $\lambda = np$. A very natural principle!

Mean Value for Continuous Random Variables

If X is a continuous random variable with density function $f(x)$, then its *mean value* is

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

If there is a minimum and/or a maximum value of X, then $-\infty$ can be replaced with the minimum and ∞ with the maximum, thus simplifying the integration. (This is another instance of our Min/Max rule on page 39.)

Mean Value for Uniform Random Variables

If X is uniform $U(a, b)$, then $f(x) = 1/(b-a)$ for $a < x < b$ (page 42). Then

$$\mathbb{E}(X) = \int_a^b \frac{x}{b-a}\, dx = \frac{x^2}{2(b-a)}\bigg|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

Note that $\mathbb{E}(X)$ is exactly the midpoint of the interval (a, b), which makes a perfect sense: all the points of the interval are "equally likely", so the midpoint is the most fair expected value.

Mean Value for Exponential Random Variables

Let X be exponential(λ), then its density is $f(x) = \lambda e^{-\lambda x}$ for $0 < x < \infty$ (page 44). Then integration by parts gives

$$\mathbb{E}(X) = \int_0^{\infty} x\lambda e^{-\lambda x}\, dx = -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}\, dx$$

The first term is zero, because $e^{-\infty} = 0$. The last integral can be written as

$$\int_0^{\infty} e^{-\lambda x}\, dx = \lambda^{-1} \int_0^{\infty} \lambda e^{-\lambda x}\, dx$$

Now the new integral above equals one, because it is the integral of the density function $f(x) = \lambda e^{-\lambda x}$ (by the normalization rule on page 38 the integral of any density function must be equal to one). Hence,

$$\mathbb{E}(X) = 1/\lambda$$

Example from page 70 (continued)

We have seen in an example on page 70 that the lifetime of a fragile system is exponential($n\lambda$) provided each of its n components has exponential(λ) lifetime. Now we see that the mean lifetime of each component is $\mathbb{E}(T_{\text{component}}) = \frac{1}{\lambda}$ and the mean lifetime of the system is $\mathbb{E}(T_{\text{fragile}}) = \frac{1}{n\lambda}$. Thus, the lifespan of the system is, on average, n times shorter than that of each component! No wonder we call it fragile.

Mean Value for Standard Normal Random Variable

Let $Z = \mathcal{N}(0, 1)$ be a standard normal random variable. Then

$$\mathbb{E}(Z) = \int_{-\infty}^{\infty} x f_Z(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{x^2}{2}}\, dx.$$

The density function here is even, $f_Z(x) = f_Z(-x)$, so the product $x f_Z(x)$ is odd. By the obvious symmetry, the integral must be zero, which it is. So, $\mathbb{E}(Z) = 0$.

Mean Value for Cauchy Random Variable

Cauchy random variable X has density function

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

for all $-\infty < x < \infty$ and the distribution function

$$F_X(x) = \frac{1}{\pi} \tan^{-1} x + \frac{1}{2}.$$

Graduate students have seen it in one of the homework problems in Chapter 7 (a drunk with a flashlight). Cauchy r.v. is very special. Its density $f_X(x)$ is also even, just like $f_Z(x)$ above, but its mean value

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \frac{x}{\pi(1 + x^2)}\, dx$$

is ... not zero! This integral diverges, so it cannot be computed. It has no numerical value!

Non-existence of Mean Value

What does it mean that the mean value does not exist, in practical terms? Recall: the mean value $\mathbb{E}(X)$ is the limit of empirical averages of observed values $x_1, \ldots, x_n$ of the random variable X, as $n \to \infty$ (page 74).

In particular, if we have a sequence $z_1, \ldots, z_n$ of observed values of the standard normal random variable $\mathcal{N}(0,1)$, then their average $(z_1 + \cdots + z_n)/n$ will, indeed, nicely converge to zero (which is $\mathbb{E}(Z)$), as $n \to \infty$.

On the contrary, if we have a sequence $x_1, \ldots, x_n$ of observed values of Cauchy random variable, then their average $(x_1 + \cdots + x_n)/n$ will **not** converge to anything. That average will oscillate wildly, as n grows, going up and down "like crazy" and reaching arbitrary large values (both positive and negative). One has to do a computer experiment to observe this spectacular process!..

Rules for Mean Value

Since the mean value $\mathbb{E}(X)$ is given by an integral, it has properties similar to those of integrals.

Rule 1. If $Y = aX$, where a is a constant, then $\mathbb{E}(Y) = a\,\mathbb{E}(X)$. That is, a constant can be 'factored out'. For example, $\mathbb{E}(2X) = 2\,\mathbb{E}(X)$, $\mathbb{E}(-X) = -\mathbb{E}(X)$, etc.

Rule 2. If $Y = X + b$, where b is a constant, then $\mathbb{E}(Y) = \mathbb{E}(X) + b$. For example, $\mathbb{E}(X - 2) = \mathbb{E}(X) - 2$.

Rule 3. If $Y = X_1 + X_2$, then $\mathbb{E}(Y) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$. Also, if $Y = X_1 - X_2$, then $\mathbb{E}(Y) = \mathbb{E}(X_1) - \mathbb{E}(X_2)$.

Rule 4. Suppose X is a constant, i.e. takes just one value, c, with probability one, i.e., $\mathbb{P}(X = c) = 1$. Then $\mathbb{E}(X) = c$.

Here is an example of how these rules can be used:

$$\mathbb{E}(2X - 4Y + 7) = 2\,\mathbb{E}(X) - 4\,\mathbb{E}(Y) + 7.$$

Mean Value for Normal Random Variables

Recall that an arbitrary normal random variable $Y = \mathcal{N}(\mu, \sigma^2)$ is related to a standard normal by $Y = \mu + \sigma Z$ (page 57). Then by Rules 1-4 we have

$$\mathbb{E}(Y) = \mu + \sigma\,\mathbb{E}(Z) = \mu + 0 \cdot \sigma = \mu$$

So the first parameter μ of any normal $\mathcal{N}(\mu, \sigma^2)$ represents its mean value.

Bernoulli Random Variable

Recall that a Bernoulli trial is a simple experiment with two possible outcomes: a success (labeled S) and a failure (labeled F); see page 26. Success occurs with probability p and failure with probability $q = 1 - p$.

Let us mark success by 1 and failure by 0. Then we get a random variable X that takes two values: 1 (with probability p) and 0 (with probability $q = 1 - p$). This is called *Bernoulli random variable*. Its mean value is

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot q = p$$

Mean Value for Binomial Random Variables (alternatively)

Recall that a binomial random variable is the number of successes in n independent Bernoulli trials (page 27). With each trial we associate a Bernoulli random variable (as above), so now we have n independent Bernoulli random variables $X_1, \ldots, X_n$. For example, if $n = 3$ and the outcomes of the trials are SFS, then $X_1 = 1$, $X_2 = 0$, $X_3 = 1$.

A crucial observation: adding $X_1 + \cdots + X_n$ gives exactly the number of successes in n Bernoulli trials! Therefore,

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

where X is the binomial random variable, $b(n, p)$, and X_i are independent Bernoulli random variables. Now, the summation Rule 3 on page 79 yields

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = \underbrace{p + \cdots + p}_{n} = np$$

Mean Value of Function of Random Variables

Let X be a random variable, and $y = g(x)$ a function. Then $Y = g(X)$ is another random variable, as in Chapter 7. Here we provide rules to compute the mean value $\mathbb{E}(Y)$.

If X is discrete and takes values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$, then the corresponding values of Y are $g(x_1), g(x_2), \ldots$. Hence

$$\mathbb{E}(Y) = g(x_1)p_1 + g(x_2)p_2 + \cdots$$

If X is continuous with density function $f_X(x)$, then

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

Moments of Random Variables

In particular, if $g(x) = x^k$, then $Y = X^k$. The mean value $\mathbb{E}(Y) = \mathbb{E}(X^k)$ is called the k-th *moment* of the random variable X. The term *moment* has its origin in the study of mechanics.

Example

Compute the k-th moment of the Bernoulli random variable X.

Solution: Since X only takes values 0 and 1, we always have $X^k = X$, for any k. Therefore $\mathbb{E}(X^k) = \mathbb{E}(X) = p$ for any $k \geq 1$.

Example

Compute the k-th moment of the uniform random variable $X = U(0, 1)$.

Solution: Recall that $f(x) = 1$ for $0 < x < 1$ (page 43). Then

$$\mathbb{E}(X^k) = \int_0^1 x^k f(x)\, dx = \int_0^1 x^k\, dx = \frac{1}{k+1} \tag{19}$$

Special Rule for Independent Random Variables

If X and Y are independent random variables, then

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

This rule works for independent random variables only, it usually fails for dependent random variables.

Chapter 11

Variance

Motivation

We have seen that the mean value $\mathbb{E}(X)$ of a random variable X gives the most fair expectation of X. Is this enough to describe X or predict it in practice?

Suppose you are going to stay in Seattle, WA next March and wonder what the temperature there might be. The climatological data (from weather.com) show that the average temperature in Montana in March is 47° F. This is exactly like knowing the mean value of a random variable. Is this enough for you? You realize that the actual temperature might fluctuate around 47°. If typical fluctuations are small, then 47° can be a pretty accurate prediction. Or, on the contrary, a typical weather pattern may be such that intervals of hot weather (60° to 70° F) follow intervals of cold weather (10° to 20° F), just giving 47° F on the average. In the latter case the average value of 42 tells you practically nothing of what you should really expect.

It is then necessary to supply the mean value of 47° with the range of typical fluctuations ("spread"). For example, 47 ± 3 would say that the weather is stable and the temperature between 44 and 50 degrees can be expected. Or, on the contrary, 47 ± 25 would tell you that the weather is very unstable and you should expect anything from 22° F (very cold) to 72° F (very warm). The conclusion is that the range of typical fluctuations around the mean value is practically just as important as the mean value itself.

Measuring Spread

The difference between the actual value of X and its mean value is $X - \mathbb{E}(X)$. Should we just find the average of this difference? Let us try this:

$$\mathbb{E}[X - \mathbb{E}(X)] = \text{ (by the rules on page 79) } = \mathbb{E}(X) - \mathbb{E}(X) = 0$$

It gives us nothing! The reason is clear: positive differences ($X - \mathbb{E}(X) > 0$) and negative differences ($X - \mathbb{E}(X) < 0$) cancel out in the end.

Variance

The variance of a random variable X is defined to be

$$\mathsf{Var}(X) = \mathbb{E}\big[X - \mathbb{E}(X)\big]^2$$

Alternatively, one can use a "shortcut formula"

$$\mathsf{Var}(X) = \mathbb{E}\big(X^2\big) - \big[\mathbb{E}(X)\big]^2$$

To see that these two formulas are equivalent, let us expand the square in the first one and use the rules on page 79:

$$\begin{aligned}\mathbb{E}\big[X - \mathbb{E}(X)\big]^2 &= \mathbb{E}\big(X^2 - 2X \cdot \mathbb{E}(X) - [\mathbb{E}(X)]^2\big) \\ &= \mathbb{E}\big(X^2\big) - 2\big[\mathbb{E}(X)\big]^2 + \big[\mathbb{E}(X)\big]^2 \qquad (20)\end{aligned}$$

and so we get the second formula for the variance. The second formula is often more convenient in practical calculations.

Examples

(a) Let X take values 0 and 1 with probability 1/2 each. Note that $X^2 = X$. Then $\mathsf{Var}(X) = \mathbb{E}(X) - [\mathbb{E}(X)]^2 = 1/2 - 1/4 = 1/4$.

(b) Let $X = U(0, 1)$. Then, according to formula (19) on page 81

$$\mathsf{Var}(X) = \mathbb{E}\big(X^2\big) - \big[\mathbb{E}(X)\big]^2 = 1/3 - 1/4 = 1/12$$

(c) Let X be the number shown by a die. Then, some tedious calculations (we omit them) show that $\mathsf{Var}(X) \approx 2.92$.

Note: In these examples one can easily find all possible deviations of X from its mean value $\mathbb{E}(X)$, and then find the average one. In the example (a), it is 1/2, in (b) it is 1/4, in (c) it is 1.5. Why are these numbers different from the values of $\mathsf{Var}(X)$ found above? The main reason is that $\mathsf{Var}(X)$ measures *squared* deviations, rather than deviations. So, we need to take the square root of $\mathsf{Var}(X)$, to describe average deviations of X.

Standard Deviation

The *standard deviation* of a random variable X is defined to be

$$\sigma_X = \sqrt{\mathsf{Var}(X)}.$$

Examples (continued)

The standard deviations in the previous examples are:

(a) $\sigma_X = \sqrt{1/4} = 1/2$

(b) $\sigma_X = \sqrt{1/12} \approx 0.29$

(c) $\sigma_X = \sqrt{2.92} \approx 1.71$

Note: Still, only in (a) the standard deviation matches the average deviation computed directly. In (b) and (c) the standard deviation is slightly higher than the average deviation. Yes, this is true: squaring the actual deviations to compute the variance and then taking square root of the variance gives slightly distorted (overestimated) value of typical deviations. But, on the other hand, there are many advantages of working with the standard deviation as defined above, rather than with precisely computed average deviation. In any case, it is traditional in probability theory to work with the standard deviation, so we have little choice..:-)

Rules for Variance and Standard Deviation

Rule 1. Suppose X is a constant, i.e. takes just one value, c, with probability one, i.e., $\mathbb{P}(X = c) = 1$. Then $\mathsf{Var}(X) = 0$ and $\sigma_X = 0$.

Rule 2. If $Y = aX$, where a is a constant, then $\mathsf{Var}(Y) = a^2\,\mathsf{Var}(X)$. (That is, a constant must be squared before it can be factored out). For example,

$$\mathsf{Var}(2X) = 4\,\mathsf{Var}(X), \quad \mathsf{Var}(-5X) = 25\,\mathsf{Var}(X), \quad \mathsf{Var}(-X) = \mathsf{Var}(X).$$

In this case, also, $\sigma_Y = |a|\sigma_X$.

Rule 3. If $Y = X + b$, where b is a constant, then $\mathsf{Var}(Y) = \mathsf{Var}\,(X)$ and $\sigma_Y = \sigma_X$. For example, $\mathsf{Var}(X - 2) = \mathsf{Var}(X)$.

Rule 4. If $Y = X_1 + X_2$, and X_1 and X_2 are *independent*, then $\mathsf{Var}(Y) = \mathsf{Var}(X_1) + \mathsf{Var}(X_2)$. Also, $\sigma_Y = \sqrt{\sigma_{X_1}^2 + \sigma_{X_2}^2}$.

Comments

In Rule 1, a constant value cannot possibly vary, this is why $\mathsf{Var}(X) = 0$.

In Rule 3, adding a constant means translating (moving) all the values of X by a fixed distance on the real line. In this case the mean value $\mathbb{E}(X)$ is moved by the same distance, so all the deviations of X from its mean value will not change, this is why $\mathsf{Var}(X)$ and σ_X remain unchanged.

In Rule 4, does it remind you the Pythagorean theorem? There is, indeed, a deep connection with it, but we will not explore it.

Rule 4 can be derived by the following calculation:

$$\begin{aligned} \mathsf{Var}(X_1 + X_2) &= \mathbb{E}[(X_1 + X_2)^2] - [\mathbb{E}(X_1 + X_2)]^2 \\ &= \mathbb{E}(X_1^2) + 2\mathbb{E}(X_1X_2) + \mathbb{E}(X_2^2) \\ &\quad -[\mathbb{E}(X_1)]^2 - 2\mathbb{E}(X_1) \cdot \mathbb{E}(X_2) - [\mathbb{E}(X_2)]^2 \qquad (21) \end{aligned}$$

Now, by the special rule on page 81, we have $\mathbb{E}(X_1X_2) = \mathbb{E}(X_1) \cdot \mathbb{E}(X_2)$, so the terms $2\mathbb{E}(X_1X_2)$ and $2\mathbb{E}(X_1) \cdot \mathbb{E}(X_2)$ cancel out. The remaining terms can be easily grouped to make $\mathsf{Var}(X_1) + \mathsf{Var}(X_2)$.

Tricky Question

Let X and Y be independent random variables. Is it true that

$$\mathsf{Var}(X - Y) = \mathsf{Var}(X) - \mathsf{Var}(Y)?$$

Answer. No. This is an incorrect "application" of Rule 4. A correct application would be

$$\mathsf{Var}[X + (-Y)] = \mathsf{Var}(X) + \mathsf{Var}(-Y) = \mathsf{Var}(X) + \mathsf{Var}(Y),$$

where the last equation is due to Rule 2.

We note also that $\mathsf{Var}(X) \geq 0$ and $\sigma_X \geq 0$. Moreover, $\mathsf{Var}(X) = 0$ and $\sigma_X = 0$ only if X is a constant, as in Rule 1.

Example

Can there be a random variable with

$$\mathbb{E}(X) = 4 \qquad \text{and} \qquad \mathbb{E}(X^2) = 13\,?$$

Solution: No, because such a random variable would have $\mathsf{Var}(X) = 13 - 4^2 = -3$, which is impossible because we know that $\mathsf{Var}(X) \geq 0$.

Example

Let X and Y be independent, $\mathbb{E}(X) = 5$ and $\mathbb{E}(Y) = -3$, as well as $\sigma_X = 2$ and $\sigma_Y = 3$. Compute the mean value and the standard deviation of $Z = 3X - 2Y - 2$.

Solution: Using the rules on page 79 gives

$$\mathbb{E}(Z) = 3 \cdot \mathbb{E}(X) - 2 \cdot \mathbb{E}(Y) - 2 = 15 + 6 - 2 = 19.$$

Using the rules on page 86 gives

$$\mathsf{Var}(Z) = 3^2\,\mathsf{Var}(X) + 2^2\,\mathsf{Var}(Y) = 9 \cdot 4 + 4 \cdot 9 = 72,$$

hence $\sigma_Z = \sqrt{72} = 6\sqrt{2}$.

Variance of Bernoulli Random Variable

Recall that Bernoulli random variable X takes values 1 and 0 with probabilities p and q, respectively (page 80). Also note that $X^2 = X$. Then $\mathbb{E}(X^2) = \mathbb{E}(X) = p$, so

$$\mathsf{Var}(X) = \mathbb{E}\big(X^2\big) - \big[\mathbb{E}(X)\big]^2 = p - p^2 = p(1-p) = pq$$

Variance of Binomial Random Variable

Recall that a binomial random variable $X = b(n,p)$ is the sum $X = X_1 + \cdots + X_n$ of n independent Bernoulli random variables (page 80). Hence, by Rule 4 on page 86

$$\mathsf{Var}(X) = \mathsf{Var}(X_1) + \cdots + \mathsf{Var}(X_n) = \underbrace{pq + \cdots + pq}_{n} = npq.$$

Meaning of σ_X

The standard deviation σ_X is the most typical (average) deviation of the random variable X from its mean value $\mathbb{E}(X)$. Hence, one can expect X typically take values $\mathbb{E}(X) \pm \sigma_X$, in a way mentioned on page 83. We will express this by a "formula"

$$X \approx \mathbb{E}(X) \pm \sigma_X. \tag{22}$$

Of course, σ_X is the average deviation, while actual deviations may be smaller or larger. Deviations up to $2\sigma_X$ should be considered as still quite likely, while those over $3\sigma_X$ are already quite unlikely (see page 59).

One can also have a visual image in mind: typical values of a random variable X make a cluster (a dense cloud) of points on the real line. Now $\mathbb{E}(X)$ is the center of that cluster and σ_X is, approximately, one quarter of its size.

Binomial Random Variable for Large n

Let X be a binomial random variable $b(n,p)$. We know already that $\mathbb{E}(X) = np$ and $\sigma_X = \sqrt{npq}$. Hence, by the previous pattern, we can represent X by

$$X \approx np \pm \sqrt{npq}$$

These are typical, most expected values of $X = b(n,p)$.

For an illustration, let $p = q = 1/2$ (think of tossing of a coin, and X being the number of Heads in n tosses), then

(a) for $n = 100$, we have $X \approx 50 \pm 5$;
(b) for $n = 1000$, we have $X \approx 500 \pm 16$;
(c) for $n = 10,000$, we have $X \approx 5,000 \pm 50$;

This shows that the expected values of a binomial random variable $X = b(n,1/2)$ are all quite close to $n/2$. Even though the typical deviations do grow with n (as 5, 16, and 50 above), but they grow much more slowly than n and $\mathbb{E}(X) = n/2$ do. So, relative to $\mathbb{E}(X)$, the deviations become less and less visible.

Relative Frequency

The contrast between the changes of $\mathbb{E}(X)$ and σ_X, as n grows, becomes even more pronounced if we consider $\bar{X}_n = X/n$, the relative frequency of successes. By the rules for mean values and variances, $\mathbb{E}(\bar{X}_n) = \mathbb{E}(X)/n = p$ and $\mathsf{Var}(\bar{X}_n) = \mathsf{Var}(X)/n^2 = pq/n$. Hence,

$$\bar{X}_n \approx p \pm \sqrt{pq}/\sqrt{n}.$$

For an illustration, again let $p = q = 1/2$. Then

$$\bar{X}_n \approx \frac{1}{2} \pm \frac{1}{2\sqrt{n}}.$$

We see that as n increases, the typical deviations decrease and converge to zero. Hence, $\bar{X}_n$ is concentrating (grouping, clustering) more and more tightly near its mean value $1/2$. So, as $n \to \infty$, we expect the values of $\bar{X}_n$ to be closer and closer to $1/2$, that is to *converge* to $1/2$. This is, indeed, the case, and we will get back to this issue in Chapter 14.

The rest of Chapter 11 is optional. We just mention two more, quite special numerical characteristics of random variables.

Skewness (optional material)

The skewness of a random variable X is

$$\beta_1 = \frac{\mathbb{E}\big[X - \mathbb{E}(X)\big]^3}{\sigma_X^3}.$$

It characterizes the degree of *asymmetry* of the density function of X about the mean value $\mathbb{E}(X)$.

For example, $\beta_1 = 0$ for any normal random variable $\mathcal{N}(\mu, \sigma^2)$, because its density is perfectly symmetric about its mean value μ. The same is true for any uniform random variable $U(a, b)$. But this is not true for exponential random variable, as its density is quite skewed.

Kurtosis (optional material)

The kurtosis of a random variable X is

$$\beta_2 = \frac{\mathbb{E}\big[X - \mathbb{E}(X)\big]^4}{\sigma_X^4}.$$

This one characterizes the heaviness of the *tails* of the density $f(x)$ of X, i.e. its behavior of $f(x)$ far away from the mean value $\mathbb{E}(X)$. More precisely, if $f(x)$ has heavy (thick) tails far from the mean value $\mathbb{E}(X)$, then the kurtosis is large. We note that $\beta_2 \geq 1$ for all random variables.

Any uniform random variable $U(a, b)$ have practically no tails (its density drops to zero beyond the interval (a, b)), and it has $\beta_2 = 1.8$.

Normal random variables have tails, but those are thin, they decrease to zero and practically vanish very rapidly away from $\mathbb{E}(X)$. For all normal random variables, $\beta_2 = 3$.

Chapter 12

Moment Generating Function

General (Abstract) Formula

Let X be a random variable, and consider the function $g(x) = e^{tx}$ of x (here t is an additional variable, its role will be clarified shortly). Then $Y = g(X) = e^{tX}$ is another random variable. Its mean value of $\mathbb{E}(Y) = \mathbb{E}(e^{tX})$ will depend on t, so it will be a function of t. It is called the *moment generating function* (m.g.f., for short) of the random variable X:

$$\mathbb{M}_X(t) = \mathbb{E}(e^{tX}).$$

The variable t becomes the argument of this function.

Practical Formulas

According to rules on page 81, if X is a discrete random variable and takes values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$, then

$$\mathbb{M}_X(t) = \mathbb{E}(e^{tX}) = e^{tx_1}p_1 + e^{tx_2}p_2 + \cdots$$

If X is a continuous random variable with density function $f_X(x)$, then

$$\mathbb{M}_X(t) = \mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx.$$

Examples

(a) Is $e^{-t}/3 + 2e^{5t}/3$ a moment generating function?

(b) What about $e^{t}/2 + e^{-2t}/3$?

Solution: (a) Yes, this formula fits the pattern of the first practical formula on the previous page, with $x_1 = -1$, $x_2 = 5$ and $p_1 = 1/3$, $p_2 = 2/3$.

(b) No, because the sum of the coefficients, $1/2 + 1/3$, is not equal to one (and the probabilities must sum to one!).

Generating Moments

What is the use of the m.g.f. $\mathbb{M}_X(t)$? Let us differentiate it and substitute $t = 0$. It is easier done with discrete random variables:

$$\mathbb{M}'_X(t) = e^{tx_1}x_1p_1 + e^{tx_2}x_2p_2 + \cdots$$

and the substitution $t = 0$ eliminates all the exponential factors, leaving us with only $x_1p_1 + x_2p_2 + \cdots$, which is $\mathbb{E}(X)$. So we arrive at

$$\mathbb{M}'_X(0) = \mathbb{E}(X)$$

Differentiating once more gives

$$\mathbb{M}''_X(t) = e^{tx_1}x_1^2p_1 + e^{tx_2}x_2^2p_2 + \cdots$$

and the substitution $t = 0$ gives $x_1^2p_1 + x_2^2p_2 + \cdots$, which is $\mathbb{E}(X^2)$, so

$$\mathbb{M}''_X(0) = \mathbb{E}(X^2)$$

which is the second moment of X. In the same way we get

$$\mathbb{M}^{(k)}(0) = \mathbb{E}(X^k).$$

for any $k \geq 1$. Hence, to compute the kth moment of X we can differentiate the function $\mathbb{M}_X(t)$ k times and then substitute $t = 0$. This is its main use of $\mathbb{M}_X(t)$ – to generate the moments of X.

We will use the moment generating functions to compute the variance for the exponential, normal and Poisson random variables.

M.g.f. for Exponential Random Variable

Let X be exponential(λ). Its density is $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, so

$$\mathbb{M}_X(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx = \int_0^\infty \lambda e^{-(\lambda - t)x}\, dx = \frac{\lambda}{\lambda - t}.$$

Taking derivative and substituting $t = 0$ gives

$$\mathbb{E}(X) = \mathbb{M}'_X(0) = \frac{\lambda}{(\lambda - t)^2}\Big|_{t=0} = \frac{1}{\lambda}$$

(which we know already from page 77), and

$$\mathbb{E}(X^2) = \mathbb{M}''_X(0) = \frac{2\lambda}{(\lambda - t)^3}\Big|_{t=0} = \frac{2}{\lambda^2}$$

Therefore

$$\mathsf{Var}(X) = \frac{2}{\lambda^2} - \Big(\frac{1}{\lambda}\Big)^2 = \frac{1}{\lambda^2} \quad \text{and} \quad \sigma_X = \frac{1}{\lambda}$$

Unpredictability for Exponential Random Variable

In the spirit of our general formula (22) on page 88, we can represent exponential random variable as

$$X \approx \mathbb{E}(X) \pm \sigma_X = \frac{1}{\lambda} \pm \frac{1}{\lambda}$$

Note a striking fact: typical deviations are about the same as the mean value! This tells us that the exponential random variable is completely unpredictable, in the sense described below.

We mentioned on page 45 that the distances between randomly deployed patrol cars on a long highway are exponential random variables. Suppose the average distance between them is 20 miles. Then the standard deviation is also 20 miles, so the typical range is 20 ± 20, i.e., from 0 to 40...

A naïve driver may pass a trooper and think: "Ok, it will be about 20 miles to the next one, so I can go over the speed limit unnoticed and get away with it". Such a driver would very wrong – the next trooper may very

well be quite close, as even zero distance between the troopers is within the typical range of 20 ± 20.

M.g.f. for Standard Normal Random Variable

If Z is a standard normal random variable, then

$$\mathbb{M}_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} e^{t^2/2}\, dx.$$

The term $e^{t^2/2}$ can be factored out (it does not contain x, the variable of integration). The remaining integral would contain the density function of a normal random variable $\mathcal{N}(t, 1)$; see page 57. Hence, the remaining integral will be equal to one, by the normalization rule on page 38. Hence

$$\mathbb{M}_Z(t) = e^{t^2/2}.$$

Variance of Normal Random Variable

and using the chain rule gives

$$\mathbb{M}'_Z(t) = t\, e^{t^2/2}$$

and so $\mathbb{E}(Z) = \mathbb{M}'_Z(0) = 0$, which we know already (page 78).

Differentiating $\mathbb{M}_Z(t)$ once again gives

$$\mathbb{M}''_X(t) = (1 + t^2)\, e^{t^2/2}$$

and so $\mathbb{E}(Z^2) = \mathbb{M}''_Z(0) = 1$. Therefore,

$$\mathsf{Var}(Z) = \mathbb{E}(Z^2) - [\mathbb{E}(Z)]^2 = 1 - 0^2 = 1$$

If X is an arbitrary normal $\mathcal{N}(\mu, \sigma^2)$, then $X = \mu + \sigma Z$ (page 57) and by rules on page 86

$$\mathsf{Var}(X) = \sigma^2\, \mathsf{Var}(Z) = \sigma^2$$

and then $\sigma_X = \sqrt{\mathsf{Var}(X)} = \sqrt{\sigma^2} = \sigma$.

Notational remark. Now we see that the second parameter of a normal random variable $X = \mathcal{N}(\mu, \sigma^2)$ is its variance $\sigma^2 = \sigma_X^2$. It is now clear why it is denoted by σ^2, it so conveniently matches the more general notation σ_X^2. Moreover, in many textbooks even the mean value of any random variable X is denoted by μ_X, again to match the first parameter μ of the normal random variable $\mathcal{N}(\mu, \sigma^2)$.

M.g.f. for Normal Random Variable

If X is a normal random variable $\mathcal{N}(\mu, \sigma^2)$, then $X = \mu + \sigma Z$. Now

$$\mathbb{M}_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}(e^{\mu t + \sigma t Z}) = \mathbb{E}(e^{\mu t} e^{\sigma t Z})$$

The term $e^{\mu t}$ does not contain X, so it can be factored out by Rule 1 on page 79. Thus

$$\mathbb{M}_X(t) = e^{\mu t}\, \mathbb{E}(e^{(\sigma t)Z}) = e^{\mu t}\, \mathbb{M}_Z(\sigma t) = e^{\mu t} e^{(\sigma t)^2/2} = e^{\mu t + \sigma^2 t^2/2}. \tag{23}$$

M.g.f. for Geometric Random Variable

A geometric random variable X takes values $n = 1, 2, \ldots$ with probabilities $\mathbb{P}(X = n) = pq^{n-1}$; see page 28. Then

$$\mathbb{M}_X(t) = \sum_{n=1}^{\infty} pq^{n-1} e^{tn} = pe^t \sum_{n=1}^{\infty} (qe^t)^{n-1}.$$

This is a geometric series, so by a general Calculus formula on page 28

$$\mathbb{M}_X(t) = \frac{pe^t}{1 - qe^t}.$$

Variance of Geometric Random Variable

By differentiating the above function (we omit tedious details) we get

$$\mathsf{Var}(X) = \frac{q}{p^2}$$

M.g.f. for Poisson Random Variable

A Poisson random variable X takes values $k = 0, 1, 2, \ldots$ with probabilities $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$; see page 32. Then

$$\begin{aligned} \mathbb{M}_X(t) &= \sum_{k=0}^{\infty} e^{tk} \cdot \mathbb{P}(X = k) = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} e^{-\lambda} = \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} e^{-\lambda e^t} e^{\lambda e^t - \lambda} \end{aligned}$$

The term $e^{\lambda e^t - \lambda}$ can be factored out (it does not contain the variable k). Then the remaining sum simply equals one, because it adds the probabilities of the poisson(λe^t) random variable. Hence,

$$\mathbb{M}_X(t) = e^{\lambda e^t - \lambda} = e^{\lambda(e^t - 1)},$$

a "double exponential" function.

Variance of Poisson Random Variable

Differentiating the above $\mathbb{M}_X(t)$ gives

$$\mathbb{M}'_X(t) = \lambda\, e^t\, e^{\lambda(e^t - 1)}$$

and so $\mathbb{E}(X) = \mathbb{M}'_X(0) = \lambda$, which we know already (page 76).

With a little extra work, we can differentiate it once again:

$$\mathbb{M}''_X(t) = (\lambda^2 + \lambda)\, e^t\, e^{\lambda(e^t - 1)}$$

and so $\mathbb{E}(X^2) = \mathbb{M}''_X(0) = \lambda^2 + \lambda$. Therefore

$$\mathsf{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

and then $\sigma_X = \sqrt{\lambda}$.

Variance of Uniform Random Variable

Let X be uniform $U(a,b)$. Its density is $f(x) = \frac{1}{b-a}$; see page 42. Its mean value is $\mathbb{E}(X) = \frac{a+b}{2}$; see 77. Now the variance of X can be found by direct integration:

$$\begin{aligned} \mathsf{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \int_a^b \frac{x^2}{b-a}\,dx - \frac{(a+b)^2}{4} \\ &= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{b^2+a^2-2ab}{12} = \frac{(b-a)^2}{12} \end{aligned}$$

Summary

The above results are summarized in a chart below:

	$\mathbb{E}(X)$	$\mathsf{Var}(X)$	σ_X	$\mathbb{M}_X(t)$
binomial(n,p)	np	npq	$\sqrt{npq}$	$(pe^t+q)^n$
geometric(p)	$1/p$	q/p^2	$\sqrt{q}/p$	$\frac{pe^t}{1-qe^t}$
poisson(λ)	λ	λ	$\sqrt{\lambda}$	$e^{\lambda(e^t-1)}$
uniform $U(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{b-a}{\sqrt{12}}$	
exponential(λ)	$1/\lambda$	$1/\lambda^2$	$1/\lambda$	$\frac{\lambda}{\lambda-t}$
st.normal $\mathcal{N}(0,1)$	0	1	1	$e^{t^2/2}$
normal $\mathcal{N}(\mu,\sigma^2)$	μ	σ^2	σ	$e^{\mu t+\sigma^2 t^2/2}$

Special rule for Moment Generating Functions

If X and Y are independent random variables, then

$$\mathbb{M}_{X+Y}(t) = \mathbb{M}_X(t)\mathbb{M}_Y(t)$$

Indeed, we have

$$\mathbb{M}_{X+Y}(t) = \mathbb{E}\left(e^{t(X+Y)}\right) = \mathbb{E}\left(e^{tX}e^{tY}\right) = \mathbb{E}(e^{tX})\cdot\mathbb{E}(e^{tY}) = \mathbb{M}_X(t)\mathbb{M}_Y(t).$$

In the middle of the above line, we used another special rule for independent random variables; see page 81.

Stable Distributions

The special rule on the previous page can be used to find the distribution of $X + Y$, if X and Y are independent random variables. For example, if $X = \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y = \mathcal{N}(\mu_2, \sigma_2^2)$ are two independent *normal* random variables, then $X + Y$ has m.g.f.

$$\mathbb{M}_{X+Y}(t) = e^{\mu_1 t + \sigma_1^2 t^2/2} \cdot e^{\mu_1 t + \sigma_1^2 t^2/2} = e^{(\mu_1+\mu_2)t + (\sigma_1^2+\sigma_2^2)t^2/2}.$$

By general formula (23) on page 95, this is the m.g.f. of a normal random variable, $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Hence, the sum of two independent normal random variables is also normal, and we get a rule:

$$\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Similarly, if X is poisson(λ_1) and Y is poisson(λ_2), then

$$\mathbb{M}_{X+Y}(t) = e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$$

hence the sum of two independent Poisson random variables is also Poisson:

$$\text{poisson}(\lambda_1) + \text{poisson}(\lambda_2) = \text{poisson}(\lambda_1 + \lambda_2).$$

Lastly, for binomial random variables: if X is $b(n_1, p)$ and Y is $b(n_2, p)$, then

$$\mathbb{M}_{X+Y}(t) = (pe^t + q)^{n_1} (pe^t + q)^{n_2} = (pe^t + q)^{n_1+n_2}$$

hence the sum of two independent binomials (with the same p) is also binomial:

$$b(n_1, p) + b(n_2, p) = b(n_1 + n_2, p).$$

The last conclusion is intuitively obvious, though, because $b(n_1, p) + b(n_2, p)$ is the total number of successes in two series of trials, one of lengths n_1 and the other of length n_2. So we have $n_1 + n_2$ trials total, with the same probability of success p in each. Now it is clear that total number of successes is $b(n_1 + n_2, p)$.

If the sum of two independent random variables of the same type is a random variable that same type, we call that type *stable*. Hence, binomial, Poisson, and normal types of random variables are stable. The others (uniform, geometric, exponential) are not.

Manipulations with Normal Random Variables

Combining the rule for normal random variables given on page 58 and the rule for two independent normals from the previous page gives the following:

Let $X = \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y = \mathcal{N}(\mu_2, \sigma_2^2)$ be two independent normal random variables and a, b two constants. Then $aX + bY$ is a normal random variable:

$$aX + bY = \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

Examples

Suppose $X = \mathcal{N}(-1, 3)$ and $Y = \mathcal{N}(2, 1)$ are independent normals. Compute the probabilities of the following events:

(a) $\mathbb{P}(X + 2Y > 2)$;

(b) $\mathbb{P}(X > Y)$;

(c) $\mathbb{P}(2X < Y - 2)$.

Solution:

(a) $X + 2Y$ is $\mathcal{N}(-1 + 2 \cdot 2, 3 + 2^2 \cdot 1) = \mathcal{N}(3, 7)$, hence

$$\mathbb{P}(X + 2Y > 2) = 1 - \Phi\left(\frac{2 - 3}{\sqrt{7}}\right) = 0.6480.$$

(b) $X - Y$ is $\mathcal{N}(-1 - 2, 3 + 1) = \mathcal{N}(-3, 4)$, hence

$$\mathbb{P}(X > Y) = \mathbb{P}(X - Y > 0) = 1 - \Phi\left(\frac{0 - (-3)}{\sqrt{4}}\right) = 0.0668.$$

(c) $2X - Y$ is $\mathcal{N}\big(2(-1) - 2, 2^2 \cdot 3 + 1\big) = \mathcal{N}(-4, 13)$, hence

$$\mathbb{P}(2X < Y - 2) = \mathbb{P}(2X - Y < -2) = \Phi\left(\frac{(-2) - (-4)}{\sqrt{13}}\right) = 0.7088.$$

Chapter 13

Covariance and Correlation

Covariance

On page 86 we derived a formula for the variance $\mathsf{Var}(X+Y)$ and obtained

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\,\mathbb{E}(XY) - 2\,\mathbb{E}(X)\cdot\mathbb{E}(Y).$$

If X and Y are *independent*, then by the special rule on page 81 we have $\mathbb{E}(XY) = \mathbb{E}(X)\cdot\mathbb{E}(Y)$, and then the last two terms cancel out. Here we are going to deal with *dependent* random variables. Then the quantity

$$\mathsf{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\cdot\mathbb{E}(Y) \tag{24}$$

may not be zero. It is called the *covariance* between X and Y. Alternatively, one can put

$$\mathsf{Cov}(X,Y) = \mathbb{E}\big[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big]. \tag{25}$$

To convert (24) to (25), one can do calculations similar to (20) on page 84.

Variance of $X+Y$

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\,\mathsf{Cov}(X,Y).$$

This rule is valid for *any* random variables X and Y, independent or not.

A curious note: the above equation resembles a simple algebraic formula $(a+b)^2 = a^2 + b^2 + 2ab$. This may help to remember it.

Covariance and Dependence

For independent random variables X and Y we have $\mathsf{Cov}(X, Y) = 0$. The covariance is often regarded as the *measure of dependence* between random variables. The larger the covariance, the stronger the dependence between the random variables.

Note, however, that sometimes dependent random variables have zero covariance (such an "odd" example is given below).

Rules for Covariance

(a) For any random variable X

$$\mathsf{Cov}(X, X) = \mathsf{Var}(X).$$

(b) The covariance is symmetric:

$$\mathsf{Cov}(X, Y) = \mathsf{Cov}(Y, X).$$

(c) The covariance is linear in both arguments:

$$\mathsf{Cov}(a_1 X_1 + a_2 X_2, Y) = a_1 \,\mathsf{Cov}(X_1, Y) + a_2 \,\mathsf{Cov}(X_2, Y)$$

$$\mathsf{Cov}(X, b_1 Y_1 + b_2 Y_2) = b_1 \,\mathsf{Cov}(X, Y_1) + b_2 \,\mathsf{Cov}(X, Y_2).$$

(d) If X and Y are independent, then $\mathsf{Cov}(X, Y) = 0$ (but not vice versa!).

(e) If X is a constant, i.e., it takes just one value, with probability one, then $\mathsf{Cov}(X, Y) = 0$ for any random variable Y.

One can verify these rules easily by using the rules for means (page 79).

Example

Let X be uniform $U(0,1)$. Find $\mathsf{Cov}(X, X^2)$.

Solution: Here we use (24) from the previous page and formula (19) from page 81:

$$\mathsf{Cov}(X, X^2) = \mathbb{E}(X^3) - \mathbb{E}(X) \cdot \mathbb{E}(X^2) = \frac{1}{4} - \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{12}.$$

Example

Let X and Y be two independent random variables, both uniform $U(0,1)$. Find $\mathsf{Cov}(X + 2Y, X^2 - Y)$.

Solution: Here we use the rules for covariance from page 101:

$$\begin{aligned} \mathsf{Cov}(X + 2Y, X^2 - Y) &= \mathsf{Cov}(X, X^2) + 2\,\mathsf{Cov}(Y, X^2) - \mathsf{Cov}(X, Y) - 2\,\mathsf{Cov}(Y, Y) \\ &= \frac{1}{12} + 0 + 0 - 2 \times \frac{1}{12} = -\frac{1}{12}. \end{aligned}$$

Here the first $1/12$ comes from the previous example, and the second $1/12$ comes from $\mathsf{Cov}(Y, Y) = \mathsf{Var}(Y) = 1/12$; see page 97.

Odd Example

A random variable X takes three values $-2, 0$ and 2, with probability $1/3$ each. Let $Y = X^2$. Compute $\mathsf{Cov}(X, Y)$.

Solution: It is easy to see that $\mathbb{E}(XY) = \mathbb{E}(X^3) = 0$, and also $\mathbb{E}(X) = 0$ and $\mathbb{E}(Y) = 8/3$ (the value of $\mathbb{E}(Y)$ is not important, though). Then $\mathsf{Cov}(X, Y) = 0$.

Note: in the above example X and Y are obviously dependent (knowing X one can compute Y precisely). But, for some reason, X and Y happen to have zero covariance. This "mystery" is cleared on the next page.

The Sign of Covariance

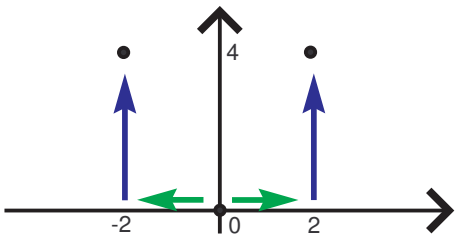
Suppose $\mathsf{Cov}(X,Y) > 0$. Then the formula (25) on page 100 tells us that the product $(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))$ tends to be positive, i.e., the terms $(X - \mathbb{E}(X))$ and $(Y - \mathbb{E}(Y))$ tend to be of the same sign (both positive or both negative). This means that if X happens to be above its mean value (fluctuates upward), then Y is also likely to be above its mean value. If X is below its mean value, then the same probably happens to Y. In other words, X and Y tend to fluctuate in accord (*in unison*).

On the contrary, if $\mathsf{Cov}(X,Y) < 0$, then the terms $(X - \mathbb{E}(X))$ and $(Y - \mathbb{E}(Y))$ tend to have opposite signs (one positive and the other negative). In other words, X and Y tend to fluctuate in the opposite directions: when X goes up Y goes down, and vice versa; thus X and Y tend to fluctuate "in discord".

	$X - \mathbb{E}(X)$	$Y - \mathbb{E}(Y)$
$\mathsf{Cov}(X,Y) > 0$	$+$ $-$	$+$ $-$
$\mathsf{Cov}(X,Y) < 0$	$+$ $-$	$-$ $+$

Odd Example (continued)

In the Odd Example on the previous page, we have $\mathsf{Cov}(X,Y) = 0$. Look closely at this example and you see that whether X happens to be above its mean value 0 (i.e., $X = 2$) or below it (i.e., $X = -2$), we have $Y = 4$. So, changing X either way (up or down) from its mean value sends Y in one direction – upward. Hence fluctuations of X and Y are equally likely to occur in the same direction and in the opposite directions. This is exactly the reason why the covariance between X and Y turns zero.



If $\mathsf{Cov}(X,Y) \neq 0$, this tells us that X and Y are dependent. But how strong is the dependence? The numerical value of the covariance $\mathsf{Cov}(X,Y)$ does not tell us much – it depends on the units in which X and Y are measured.

The following modification of covariance is better – it does not depend on the units (it is *dimensionless*). It estimates the strength of dependence quite accurately.

Correlation

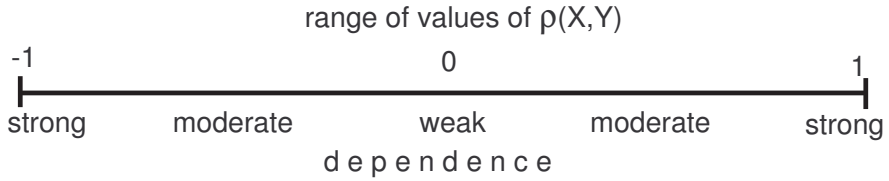
$$\rho(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

This is the *correlation*, or *correlation coefficient*, between X and Y. It always takes values between -1 and 1:

$$-1 \leq \rho(X,Y) \leq 1$$

Interpretation of Correlation

- If $\rho(X,Y) = 0$, then X and Y are *practically* independent (one often says that X and Y are *uncorrelated*).
- If $\rho(X,Y) \approx 0$, then X and Y are *weakly* dependent, i.e. knowing X very little can be said of Y (and vice versa).
- If $\rho(X,Y) \approx \pm 1$, then X and Y are *strongly* dependent, i.e. knowing X the value of Y can be determined quite accurately, if not precisely (and vice versa).



Example

Let X be uniform $U(0,1)$. Compute $\rho(X, X^2)$.

Solution: In the first example on page 102 we found that $\mathsf{Cov}(X,Y) = 1/12$. We also know that $\mathsf{Var}(X) = 1/12$; see page 97. And by the general formula (19) on page 81

$$\mathsf{Var}(X^2) = \mathbb{E}(X^4) - [\mathbb{E}(X^2)]^2 = 1/5 - (1/3)^2 = 4/45$$

Therefore

$$\rho(X, X^2) = \frac{\mathsf{Cov}(X, X^2)}{\sigma_X \cdot \sigma_{X^2}} = \frac{1/12}{\sqrt{1/12} \cdot \sqrt{4/45}} \approx 0.968.$$

This is close to 1.0, so it shows a very strong dependence between X and X^2. Indeed, they are obviously dependent: knowing X one can compute $X^2 = X \cdot X$, and knowing X^2 one can compute $X = \sqrt{X^2}$.

Extreme cases: $\rho(X,Y) = \pm 1$

One may wonder why the correlation between X and X^2 in the last example falls short of its maximum value, 1. As it turns out, $\rho(X,Y) = \pm 1$ only if X and Y are related by a linear formula, i.e., $Y = aX + b$ with some constants a and b. More precisely, $\rho(X,Y) = 1$ if $a > 0$ and $\rho(X,Y) = -1$ if $a < 0$.

For instance, if X is the forecast for tomorrow's temperature in Fahrenheit and Y is the same in Celcius, then $X = 1.8 \cdot Y + 32$, which is a linear relation. Therefore $\rho(X,Y) = 1$ (because the coefficient 1.8 is positive).

Chapter 14

Law of Large Numbers

Motivation: Relative Frequency

We resume our discussion started on page 89. Recall that we dealt with a binomial random variable $X = b(n, p)$, and then considered $\bar{X}_n = X/n$, the relative frequency of successes in n trials. Its mean value was $\mathbb{E}(\bar{X}_n) = p$, and its variance was $\mathsf{Var}(\bar{X}_n) = pq/n$. We concluded that $\bar{X}_n$ tended to concentrat tightly near its mean value p, as n grows. Here we investigate this phenomenon further. It is one of the central facts in probability theory.

We need to estimate precisely by how much a random variable can deviate from its mean value.

Markov Inequality

Let $X \geq 0$ be a random variable that only takes non-negative values, and $t > 0$ a real number. Then

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

This inequality estimates the probability that X takes large values.

Brief explanation: Let Y be a new random variable defined by

$$Y = \begin{cases} t & \text{if } X \geq t \\ 0 & \text{if } 0 \leq X < t \end{cases}$$

Note that Y takes two values (t and 0), so it is a discrete random variable, and

$$\mathbb{E}(Y) = t \cdot \mathbb{P}(Y = t) + 0 \cdot \mathbb{P}(Y = 0) = t \cdot \mathbb{P}(X \geq t)$$

Note that $X \geq Y$ in all cases, so

$$\mathbb{E}(X) \geq \mathbb{E}(Y) = t\,\mathbb{P}(X \geq t)$$

It remains to divide the last formula by t.

Example

Let X be nonnegative and $\mathbb{E}(X) = \mu$. What can we cay about the probability $\mathbb{P}(X \geq 10\mu)$?

Solution: By Markov inequality,

$$\mathbb{P}(X \geq 10\mu) \leq \frac{\mu}{10\mu} = \frac{1}{10}.$$

Hence, there is not much chance (at most 10%) that a random variable is as large as ten times its mean value.

Chebyshev's Inequality

Let X be a random variable, and $y > 0$ a real number. Then

$$\mathbb{P}\big(|X - \mathbb{E}(X)| \geq y\big) \leq \frac{\mathsf{Var}(X)}{y^2}.$$

This inequality estimates the probability that X deviates by more than y from its mean value $\mathbb{E}(X)$.

Brief explanation: Let Y be a new random variable defined by

$$Y = \begin{cases} y^2 & \text{if } |X - \mathbb{E}(X)| \geq y \\ 0 & \text{otherwise} \end{cases}$$

Note that Y takes two values (y^2 and 0), so it is a discrete random variable, and

$$\mathbb{E}(Y) = y^2 \cdot \mathbb{P}(Y = y^2) + 0 \cdot \mathbb{P}(Y = 0) = y^2 \cdot \mathbb{P}(|X - \mathbb{E}(X)| \geq y)$$

Note that $|X - \mathbb{E}(X)|^2 \geq Y$ in all cases, so

$$\mathsf{Var}(X) = \mathbb{E}(|X - \mathbb{E}(X)|^2) \geq \mathbb{E}(Y) = y^2 \cdot \mathbb{P}(|X - \mathbb{E}(X)| \geq y)$$

It remains to divide the last formula by y^2.

Example

Let $X = \mathcal{N}(\mu, \sigma^2)$. We can estimate the probability $\mathbb{P}\big(|X - \mu| \geq 3\sigma\big)$ by Chebyshev's inequality:

$$\mathbb{P}\big(|X - \mu| \geq 3\sigma\big) \leq \frac{\sigma^2}{(3\sigma)^2} = \frac{1}{9}$$

On the other hand, the exact probability is

$$\mathbb{P}\big(|X - \mu| \geq 3\sigma\big) = 2(1 - \Phi(3)) = 0.0028$$

from the table for the function Φ. Since $0.0028 < 1/9$, the estimate is correct.

"Overkill"

In the above example, the estimate $1/9 = 0.1111$ is rather crude, it is much higher than the actual value 0.0028. It is like an ad of an auto insurance company promising that your monthly premium would never exceed $10,000. Such a statement may be, technically, correct but practically useless, if not ridiculous, as normally monthly premiums are nowhere near $10,000.

Back to Chebyshev's inequality, indeed, for most random variables it is an "overkill" - it grossly overestimates $\mathbb{P}\big(|X - \mathbb{E}(X)| \geq y\big)$. On the other hand, Chebyshev's inequality is universal, it applies to *any* random variable. And, it cannot be improved without sacrificing universality, because for every $y > 0$ one can find a random variable X (although quite "ugly") for which $\mathbb{P}\big(|X - \mathbb{E}(X)| \geq y\big) = \mathsf{Var}(X)/y^2$, i.e. Chebyshev's inequality turns into an equality. We will see that below.

Turning Chebyshev's Inequality into Equality

The logic of the explanation on page 107 suggests an idea how to construct an "ugly" random variable X for which $\mathbb{P}\big(|X - \mathbb{E}(X)| \geq y\big) = \mathsf{Var}(X)/y^2$. Such a random variable must satisfy the condition $Y = |X - \mathbb{E}(X)|^2$. In other words, X can only take three values: $\mathbb{E}(X)$ and $\mathbb{E}(X) \pm y$.

Let $a = \mathbb{E}(X)$ and $y > 0$ be given. Then X must be a discrete random variable taking three values: a, $a - y$, and $a + y$. One needs to assign probabilities to those values properly so that $\mathbb{E}(X) = a$ and $\mathbb{P}\big(|X - a| \geq y\big) = \mathsf{Var}(X)/y^2$. Further details are left as an exercise.

Example

Suppose that a random variable X has mean value $\mathbb{E}(X) = 10$ and variance $\mathsf{Var}(X) = 9$. By using Chebyshev's inequality, estimate the probability $\mathbb{P}(X \geq 25)$.

Solution: The event $\{X \geq 25\}$ can be represented as $\{X \geq 10 + 15)$, or $\{X - 10 \geq 15\}$. Hence it is a part of the larger event $\{|X - 10| \geq 15\}$. Then by Chebyshev's inequality

$$\mathbb{P}(X \geq 25) \leq \mathbb{P}\big(|X - 10| \geq 15\big) \leq \frac{9}{15^2} = \frac{1}{25}.$$

Back to Relative Frequency

Again, as on page 89, let $X = b(n, p)$ be the number of successes in n Bernoulli trials, and $\bar{X}_n = X/n$ the relative frequency of successes. We have $\mathbb{E}(\bar{X}_n) = p$ and $\mathsf{Var}(\bar{X}_n) = pq/n$. Let $y > 0$ be any number. By Chebyshev's inequality

$$\mathbb{P}\big(|\bar{X}_n - p| \geq y\big) \leq \frac{pq}{y^2 n}.$$

As n grows, the right hand side decreases to zero. Hence, the probability that $\bar{X}_n$ deviates from p by more than y is getting smaller and vanishes as $n \to \infty$. This is true for all $y > 0$. Hence, all deviations of $\bar{X}_n$ from p vanish as $n \to \infty$. The random variable $\bar{X}_n$ indeed converges to p as $n \to \infty$, as we guessed earlier, on page 89.

We observed on page 80 that a binomial random variable $b(n, p)$ is the sum of n independent Bernoulli random variables: $X = X_1 + \cdots + X_n$. Then the relative frequency of successes $\bar{X}_n$ can be written as

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

i.e., $\bar{X}_n$ is the average of n Bernoulli random variables $X_1, \ldots, X_n$. The above fact then says that the average of $X_1, \ldots, X_n$ converges to the mean value $\mathbb{E}(X_i) = p$ as $n \to \infty$. In other words, the empirical average of n (experimentally) observed values of the Bernoulli random variable converges to its (theoretical) mean. Hence the experiment must agree with the theory. In this form, the fact can be extended to more general random variables; see next.

Law of Large Numbers

Let $X_1, X_2, \ldots$ be independent identically distributed (i.i.d) random variables. Let $\mu = \mathbb{E}(X_i)$ be the common mean value of all X_i's, and $\sigma^2 = \mathsf{Var}(X_i)$ the common variance of all X_i's. For each n let

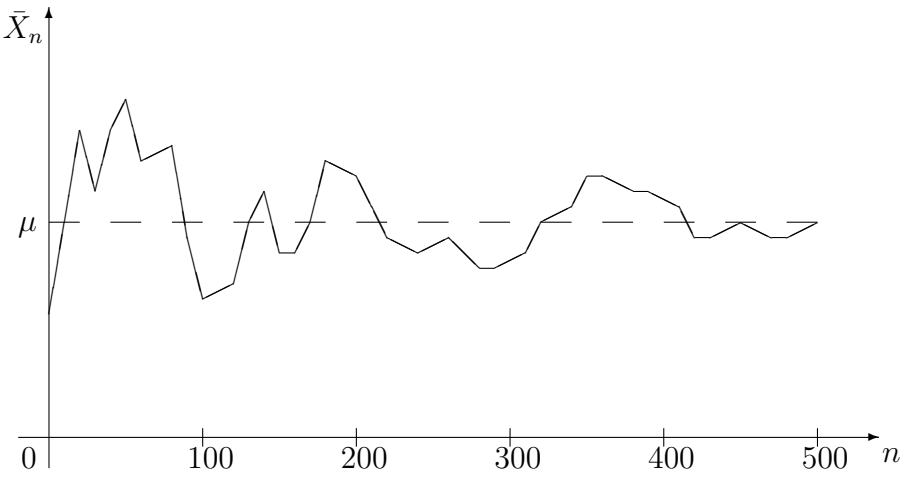
$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{26}$$

denote the average of the first n values. Then the random variable $\bar{X}_n$ converges to μ as $n \to \infty$. Precisely, for every real number $y > 0$ we have

$$\mathbb{P}\big(|\bar{X}_n - \mu| \geq y\big) \leq \frac{\sigma^2}{y^2 n} \to 0$$

as $n \to \infty$, i.e., all the deviations from μ vanish.

Indeed, the above inequality is simply Chebyshev's inequality, because $\mathsf{Var}(\bar{X}_n) = [\mathsf{Var}(X_1) + \cdots + \mathsf{Var}(X_n)]/n^2 = \sigma^2/n$.



The convergence of $\bar{X}_n$ to μ

Time average. One gets a sequence of i.i.d. random variables every time a random experiment is repeated under the same conditions and experimental values $X_1, \ldots, X_n$ are observed. Here n plays the role of time, and the empirical average (26) is called the *time average*. Then the Law of Large Numbers says that the time average approaches, as time goes on, the mean value $\mu = \mathbb{E}(X_i)$.

Monte-Carlo Integration (optional)

The law of large numbers can be used to compute integrals of functions. For example, suppose we want to compute a double integral

$$\iint_R f(x,y)\,dx\,dy \tag{27}$$

over a region R (for simplicity, let R be a part of the unit square $0 \le x, y \le 1$). Such integrals are often hard to compute, because either the function f or the region R, or both, are quite complicated. Standard methods for numerical integration may be hard to implement or become inefficient. There is, however, a straightforward computer algorithm based on the Law of Large Numbers.

The computer program generates pairs of uniformly distributed random variables (x_1, y_1), $(x_2, y_2), \ldots$. Each pair (x_i, y_i) is generated by calling an RNG (page 43) twice (once for x_i and once more for y_i). Pairs (x_i, y_i) that are not in the region R, are ignored. For each pair (x_i, y_i) that is in R, one computes the value $f_i = f(x_i, y_i)$. Then the average value

$$\frac{f_1 + f_2 + \cdots + f_n}{n} \tag{28}$$

(where n is the total number of generated pairs (x_i, y_i), including those which are not in the region R) approximates the integral (27). The larger n the better approximation. Easy, isn't in?

Since this algorithm is based on random numbers, i.e., resembles playing a roulette, it is called *Monte-Carlo integration* (Monte-Carlo is a world famous gambling center in Monaco, southern France).

Note that our example of computation of the number π with the help of random numbers on page 66 is exactly a case of Monte-Carlo integration. In that example, the region R was a quarter of the unit circle, and the function $f(x, y) = 1$ (so that the integral (27) actually coincided with the area of R).

A drawback of the Monte-Carlo integration is a slow convergence of the average (28) to the integral (27). One really needs to compute a lot of f_i's in (28) to obtain an accurate value of (27). We will see that in the next Chapter.

Chapter 15

Central Limit Theorem

Binomial Random Variable for Large n

Let X be a binomial random variable $b(n,p)$ with large n. Its probability function is

$$\mathbb{P}(X=k) = C_{n,k}p^k(1-p)^{n-k} \qquad \text{for } \ k=0,1,\dots,n,$$

(see page 27). This formula for $\mathbb{P}(X=k)$ is precise but very impractical for large n. We found in Chapter 4 a simple approximative formula (Poisson law) that works for large n and small p. But what if p is not small?

Remarkably, the binomial random variable $X = b(n,p)$ can be well approximated by a normal random variable, $Y = \mathcal{N}(\mu,\sigma^2)$. The calculations that lead to this approximation are quite complicated, and we skip them. Instead, we focus on the proper choice of μ and σ^2, the parameters of Y. Recall that $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \mathsf{Var}(Y)$. Now if we want the normal random variable $Y = \mathcal{N}(\mu,\sigma^2)$ to match the binomial random variable X, then we want their mean values to be equal, $\mathbb{E}(X) = \mathbb{E}(Y)$, and their variances to be equal, too: $\mathsf{Var}(X) = \mathsf{Var}(Y)$. This gives

$$\mu = \mathbb{E}(Y) = \mathbb{E}(X) = np \qquad \text{and} \qquad \sigma^2 = \mathsf{Var}(Y) = \mathsf{Var}(X) = npq.$$

De Moivre-Laplace[3] Theorem

For large n, the binomial random variable $X = b(n,p)$ is approximated by a normal $Y = \mathcal{N}(\mu,\sigma^2)$ with $\mu = np$ and $\sigma^2 = npq$.

How large n should be for the normal approximation to be used? A common practical rule of thumb is to apply normal approximation when $n \geq 30$. It may be good even for smaller n, like $n \sim 20$ or $n \sim 15$...

[3]Named after two French mathematicians: Abraham de Moivre (1667–1754) and Pierre-Simon Laplace (1749-1827).

Example

Toss a coin 100 times and let X be the number of Heads. Find the probability $\mathbb{P}(X \le 50)$.

Solution: First, we follow a naïve approach that leads to a wrong answer, then we will fix it. By normal approximation, $X \approx Y = \mathcal{N}(50, 25)$, because $\mu = 100 \cdot 0.5 = 50$ and $\sigma^2 = 100 \cdot 0.5 \cdot 0.5 = 25$. Now we proceed as

$$\mathbb{P}(X \le 50) \approx \mathbb{P}(Y \le 50) = \Phi\left(\frac{50-50}{\sqrt{25}}\right) = \Phi(0) = 0.5.$$

Wait a minute. Is this right? Not quite. Indeed, if this was right, then $\mathbb{P}(X \ge 51) = 1 - \mathbb{P}(X \le 50) = 0.5$. On the other hand, by normal approximation again

$$\mathbb{P}(X \ge 51) \approx \mathbb{P}(Y \ge 51) = 1 - \Phi\left(\frac{51-50}{\sqrt{25}}\right) = 1 - \Phi(0.2) = 0.4207.$$

Of course, $0.5 \ne 0.4207$, there is an almost 8% difference! Something is wrong.

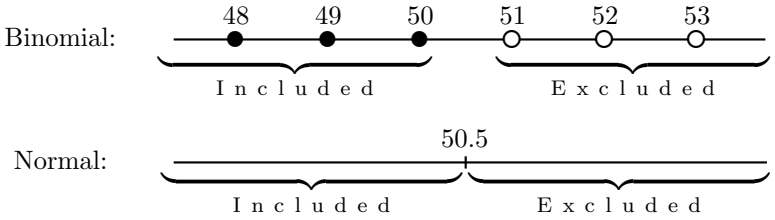
One can see now what is wrong. The random variable $X = b(n, p)$ is discrete, all its values are integers (whole numbers). So, the events $X \le 50$ and $X \ge 51$ are complementary to each other. The open interval in between, $50 < X < 51$, is irrelevant, its probability is zero. On the contrary, $Y = \mathcal{N}(\mu, \sigma^2)$ is a continuous random variable, and the probability $\mathbb{P}(50 < Y < 51)$ is positive, and not small. It is exactly this probability that we overlooked.

Note that the interval $(50, 51)$ is a "border" interval, a gap between the event $X \le 50$ and its complement $X \ge 51$. To take a proper care of this border interval, we divide it in half and include one half into the event $X \le 50$ and the other half into the complement $X \ge 51$. In other words, the event and its complement "split up" the border interval.

Now we present the correct solution. The proper range for Y is $Y \le 50.5$, hence

$$\mathbb{P}(X \le 50) \approx \mathbb{P}(Y \le 50.5) = \Phi\left(\frac{50.5-50}{\sqrt{25}}\right) = \Phi(0.1) = 0.5398.$$

Note: the exact probability $\mathbb{P}(X \le 50)$, by the binomial formula, is 0.539795. So the normal approximation gives all 4 digits correct.



Correction for Continuity (or "Histogram Correction")

When applying De Moivre-Laplace theorem, divide the border interval(s) in half and include one half into the region for Y.

Example

Toss a coin 100 times and let X be the number of Heads. Find the probability $\mathbb{P}(40 \le X \le 55)$.

Solution: As in the previous example, X is $b(100, 0.5)$ and its normal approximation is $X \approx Y = \mathcal{N}(50, 25)$. Now we apply correction for continuity. The event in question is $40 \le X \le 55$. The complement is $X \le 39$ and $X \ge 56$. There are two border intervals: $(39, 40)$ and $(55, 56)$. Then the proper range for Y is $39.5 \le Y \le 55.5$, and

$$\begin{aligned} \mathbb{P}(40 \le X \le 55) &\approx \mathbb{P}(39.5 \le Y \le 55.5) \\ &= \Phi\left(\frac{55.5-50}{\sqrt{25}}\right) - \Phi\left(\frac{39.5-50}{\sqrt{25}}\right) \\ &= \Phi(1.1) - \Phi(-2.1) = 0.8643 - 0.0179 = 0.8464. \end{aligned}$$

(The exact probability, by the binomial formula, is 0.846772.)

Note: binomial probabilities can be computed by some advanced calculators. Or you can use the on-line calculator on the instructor's web page.

Example

Toss a coin 100 times and let X be the number of Heads. Find the probability $\mathbb{P}(X = 50)$.

Solution: As before, X is $b(100, 0.5)$ and $X \approx Y = \mathcal{N}(50, 25)$. Now we apply correction for continuity. The event in question is $X = 50$, or in terms of inequalities, $50 \le X \le 50$. The complement is the union of two intervals: $X \le 49$ and $X \ge 51$. There are two border intervals: $(49, 50)$ and $(50, 51)$. So the proper range for Y is $49.5 \le Y \le 50.5$, and

$$\begin{aligned} \mathbb{P}(50 \le X \le 50) &\approx \mathbb{P}(49.5 \le Y \le 50.5) \\ &= \Phi\left(\frac{50.5-50}{\sqrt{25}}\right) - \Phi\left(\frac{49.5-50}{\sqrt{25}}\right) \\ &= \Phi(0.1) - \Phi(-0.1) = 0.5398 - 0.4602 = 0.0796. \end{aligned}$$

This finally answers the question posed on page 5!

Example

A student knows answers to 75% of questions in a course. A test is made up of some 12 questions. What is the probability that the student answers at least 10 correctly?

Solution: Let X be the number of test questions the student answers correctly. Then X is $b(12, 0.75)$. Its normal approximation is $X \approx Y = \mathcal{N}(9, 9/4)$. Now we apply correction for continuity. The event in question is $X \geq 10$. The complement is $X \leq 9$. The border interval is $(9, 10)$, so the proper range for Y is $Y \geq 9.5$, and

$$\begin{aligned}\mathbb{P}(X \geq 10) \approx \mathbb{P}(Y \geq 9.5) &= 1 - \Phi\Big(\frac{9.5 - 9}{\sqrt{9/4}}\Big) \\ &= 1 - \Phi(0.33) = 1 - 0.6293 = 0.3707.\end{aligned}$$

Extra note: The exact probability, by the binomial formula, is 0.3907. We used normal approximation, even though n was quite small ($n = 12$). Our approximate answer 0.3707 is still pretty good.

Example (with a twist)

A student knows answers to 75% of questions in a course. The professor asks the student questions until 20 correct answers are given. What is the probability that at least 25 questions will be necessary?

Solution: Note that the number of trials (questions) is not specified here. So, we need to describe our event differently. We begin with a logical observation: if at least 25 questions are necessary, then 24 are not enough. This means that after the 24th question, the student still has not given 20 correct answers.

Now our event can be described in more familiar terms. Let X be the number of correct answers given to the first 24 questions. Our event is $X < 20$, i.e. $X \leq 19$. Now we are ready to solve the problem. First, X is $b(24, 0.75)$. Its normal approximation is $X \approx Y = \mathcal{N}(18, 9/2)$. Now we apply correction for continuity. The event in question is $X \leq 19$. The complement is $X \geq 20$, so the range for Y is $Y \leq 19.5$, and

$$\mathbb{P}(X \leq 19) \approx \mathbb{P}(Y \leq 19.5) = \Phi\left(\frac{19.5 - 18}{\sqrt{9/2}}\right) = \Phi(0.71) = 0.7611.$$

(The exact probability, by the binomial formula, is 0.75335.)

Generalizing De Moivre-Laplace

We observed on page 80 that a binomial random variable $b(n,p)$ is the sum of n independent Bernoulli random variables: $X = X_1 + \cdots + X_n$. Therefore, $\mathbb{E}(X) = n\,\mathbb{E}(X_1)$ and $\mathsf{Var}(X) = n\,\mathsf{Var}(X_1)$. Now De Moivre-Laplace theorem can be stated as follows: $X = X_1 + \cdots + X_n$ is approximately a normal $Y = \mathcal{N}(\mu, \sigma^2)$ with $\mu = n\,\mathbb{E}(X_1)$ and $\sigma^2 = n\,\mathsf{Var}(X_1)$. In this form the theorem can be extended to more general random variables:

Let $X_1, X_2, \ldots, X_n$ be independent identically distributed (i.i.d) random variables. Let $\mu_X = \mathbb{E}(X_i)$ be the common mean value of all X_i's, and $\sigma_X^2 = \mathsf{Var}(X_i)$ the common variance of all X_i's. For each n let $S_n = X_1 + \cdots + X_n$.

Central Limit Theorem (for S_n)

For large n, the sum S_n is approximately normal:

$$S_n \approx \mathcal{N}(\mu, \sigma^2)$$

with

$$\mu = \mathbb{E}(S_n) = n\mu_X, \qquad \text{and} \qquad \sigma^2 = \mathsf{Var}(S_n) = n\sigma_X^2$$

Now let $\bar{X}_n = S_n/n = (X_1 + \cdots + X_n)/n$ be the average of X_i's.

Central Limit Theorem (for $\bar{X}_n$)

For large n, the variable $\bar{X}_n$ is approximately normal:

$$\bar{X}_n \approx \mathcal{N}(\mu, \sigma^2)$$

with

$$\mu = \mathbb{E}(\bar{X}_n) = \mu_X, \qquad \text{and} \qquad \sigma^2 = \mathsf{Var}(\bar{X}_n) = \sigma_X^2/n$$

The term *central* in the name of the theorem is due solely to its central role and ultimate importance in probability theory.

Example

Suppose we roll a die 50 times. What is the probability that the sum of the numbers obtained lies between 150 and 190 (inclusive)?

Solution: The sum of the numbers is $S_{50} = X_1 + \cdots + X_{50}$ where $\mathbb{E}(X_i) = 3.5$ (page 73) and $\mathsf{Var}(X_i) = 2.92$ (page 84). By Central Limit Theorem we have $S_{50} \approx Y = \mathcal{N}(175, 146)$. Since the random variable S_n is discrete, we need to apply correction for continuity, thus we get the range for Y as $149.5 < Y < 190.5$. Hence

$$\begin{aligned}
\mathbb{P}(150 \le X \le 190) &\approx \mathbb{P}(149.5 \le Y \le 190.5) \\
&= \Phi\left(\frac{190.5 - 175}{\sqrt{146}}\right) - \Phi\left(\frac{149.5 - 175}{\sqrt{146}}\right) \\
&= \Phi(1.28) - \Phi(-2.11) = 0.8997 - 0.0174 = 0.8823.
\end{aligned}$$

Example

Suppose the weight of a certain brand of bolt has a mean of 1 gram and a standard deviation of 0.13 grams. Estimate the probability that 100 of these bolts will weigh more than 102 grams.

Solution: By Central Limit Theorem the total weight W is approximately normal $Y = \mathcal{N}(100, 100 \times 0.13^2) = \mathcal{N}(100, 1.69)$. Since W is (obviously) continuous, we do not apply correction for continuity. Hence

$$\begin{aligned}
\mathbb{P}(W > 102) \approx \mathbb{P}(Y > 102) &= 1 - \Phi\left(\frac{102 - 100}{\sqrt{1.69}}\right) \\
&= 1 - \Phi(1.54) = 1 - 0.9382 = 0.0618.
\end{aligned}$$

Note that we do not know the distribution of the weights of individual bolts, that distribution is irrelevant! All we need to know is the mean weight and its standard deviation.

Universality of Central Limit Theorem

The classical CLT deals with sums of i.i.d. random variables, but its modern versions also handle sums of variables that have *different* distributions and are *weakly dependent*. Basically, any random variable that is the sum of many (almost) independent small components is approximately normal. Or, put it differently, any experimental measurement that is affected by a combination of many small random factors should be approximately normal.

This explains why it is customary to assume that practical data have normal distribution, like the height of an adult person, the weight of a fish in a pond, etc. It is not too much of an exaggeration to say: *everything random in nature has normal distribution, unless some specific constraints affect it and make it non-normal.*

Example

A basketball player makes 80% of his free throws on the average. During the season he makes 1000 free throws in official games. Let $\bar{X}$ be the frequency of his successes. Estimate $\mathbb{P}(|\bar{X} - 0.8| < 0.01)$.

Solution: The total number of successes in 1000 throws is $X = b(1000, 0.8)$, and $\bar{X} = X/1000$. Hence, $\mathbb{E}(\bar{X}) = 0.8$ and $\mathsf{Var}(\bar{X}) = 0.16/1000 = 0.00016$. By Section 15.12, $X \approx Y = \mathcal{N}(0.8, 0.00016)$. Then

$$\begin{aligned}\mathbb{P}(|X - 0.8| < 0.01) &\approx \mathbb{P}(|Y - 0.8| < 0.01)\\ &= \Phi\left(\frac{0.81 - 0.8}{\sqrt{0.00016}}\right) - \Phi\left(\frac{0.79 - 0.8}{\sqrt{0.00016}}\right)\\ &= \Phi(0.79) - \Phi(-0.79) \approx 0.7852 - 0.2148 = 0.5704.\end{aligned}$$

Example

Suppose the lifetime of a light bulb is an exponential random variable with mean 5 hours. A housekeeper wants to buy a set of light bulbs with total lifetime at least 100 hours. What is the probability that 22 bulbs will not be enough?

Solution: The total lifetime of 22 bulbs is $S_{22} = X_1 + \cdots + X_{22}$, where each X_i is exponential(1/5) (because $\lambda = 1/\mathbb{E}(X_i) = 1/5$). Now we have $\mathsf{Var}(X_i) = 1/\lambda^2 = 25$ (page 93). By using normal approximation

$$S_{22} \approx Y = \mathcal{N}(22 \cdot 5, 22 \cdot 25) = \mathcal{N}(110, 550)$$

Note that S_{22} is (obviously) continuous, so we do not use correction for continuity. Hence,

$$\mathbb{P}(S_{22} \leq 100) \approx \mathbb{P}(Y \leq 100) = \Phi\left(\frac{100 - 110}{\sqrt{550}}\right) = \Phi(-0.43) = 0.3336.$$

Extra remark: The average lifetime of 20 bulbs is 100 hours already, and the housekeeper has two extra bulbs (with total average lifetime 10 hours!), just in case. Still, these 22 bulbs may not be enough with a rather large probability of 33%... How come?

The reason why this strange fact takes place is the unpredictability of exponential random variables we noted on page 93. Indeed, some of those light bulbs can burn down very quickly, and this is not at all unusual (page 93). Devices with exponential lifetime are very unpredictable and unreliable!

Normal Approximation to Poisson

Let X be poisson(λ) with a large λ. Since Poisson random variable is stable (page 98), we can think that X is the sum of n independent Poisson random variables, each with a smaller parameter λ/n. More precisely, $X = X_1 + \cdots + X_n$ where $X_i =$poisson(λ/n). Hence, by Central Limit Theorem we have $X \approx \mathcal{N}(\mu, \sigma^2)$ with $\mu = \mathbb{E}(X) = \lambda$ and $\sigma^2 = \mathsf{Var}(X) = \lambda$, i.e.,

$$\text{poisson}(\lambda) \approx \mathcal{N}(\lambda, \lambda)$$

for large λ. When you use this rule, do not forget to apply correction for continuity – Poisson random variable is discrete!

Example

Suppose accidents on a 500 miles stretch of a highway occur at a rate of one per 20 miles. What is the chance that there is at most 26 accidents on that stretch?

Solution: The total number of accidents X is a Poisson random variable (as we noted on page 45) with average $500/20 = 25$. By Central Limit Theorem $X \approx Y = \mathcal{N}(25, 25)$. Applying correction for continuity gives

$$\begin{aligned}\mathbb{P}(X \le 26) \approx \mathbb{P}(Y < 26.5) &= \Phi\left(\frac{26.5 - 25}{\sqrt{25}}\right) \\ &= 1 - \Phi(0.3) = 0.6179.\end{aligned}$$

(The exact probability, by Poisson formula, is 0.629386, so our approximation is quite good, despite $\lambda = 25$ being no so very large.)

Example

Let S_n be the sum of n independent uniform $U(0, 1)$ random variables. Approximate S_n by a normal.

Solution: Recall that $\mathbb{E}(X) = 1/2$ and $\mathsf{Var}(X) = 1/12$ (page 97). Hence, $\mathbb{E}(S_n) = n/2$ and $\mathsf{Var}(S_n) = n/12$. We thus get

$$S_n \approx \mathcal{N}(n/2, n/12).$$

When you use this rule, do not apply correction for continuity (uniform random variable is continuous).

Generating Normal Random Variable by Computer

The above example suggests a simple and popular method of generating a standard normal random variable $Z = \mathcal{N}(0,1)$ by computer. It is convenient to pick $n = 12$, then $S_{12} \approx \mathcal{N}(6,1)$, so that $S_{12} - 6 \approx \mathcal{N}(0,1)$. A simple computer code calls a standard random number generator 12 times, adds the resulting 12 numbers, subtracts six from the sum, and that's it!

If you need to generate any normal variable $Y = \mathcal{N}(\mu, \sigma^2)$, then generate Z as above and compute $Y = \mu + \sigma Z$.

Below is a short computer code (in two popular languages, FORTRAN and C) that does the job. RAND is the call of a random number generator producing a uniform random value on $(0,1)$.

```
FORTRAN:     Z=-6.0                    C:     Z=-6.0;
             DO 1 I=1,12                      for (i=0;i<12;i++)
           1 Z=Z+RAND                           Z=Z+RAND;
             Y=MU+SIGMA*Z                     Y=mu+sigma*Z;
```

Extra note: the above method for generating normal random variables is simple but not precise. There is a more sophisticated method that produces precise normal distributions, but it is beyond the scope of this course.

Last Example

One plays a game repeatedly, each time either winning \$1 with probability p or losing \$1 with probability $q = 1 - p$. Let S_n be the total gain (or loss) after playing n rounds. Approximate S_n by a normal random variable.

Solution: We can represent $S_n = X_1 + \cdots + X_n$, where $X_i = \pm 1$ are gains/losses in individual rounds. Note that each X_i is a random variable that takes two values: $+1$ with probability p and -1 with probability $q = 1 - p$. Then we have

$$\mathbb{E}(X_i) = 1 \cdot p + (-1) \cdot q = p - q = 2p - 1$$

and

$$\mathsf{Var}(X_i) = \mathbb{E}(X_i^2) - [\mathbb{E}(X_i)]^2 = 1 - (2p-1)^2 = 4p - 4p^2 = 4pq.$$

Hence, $\mathbb{E}(S_n) = (p-q)n$ and $\mathsf{Var}(S_n) = 4pqn$. Thus our normal approximation is

$$S_n \approx \mathcal{N}\big((p-q)n, 4pqn\big).$$

This example serves as a bridge to the next chapter of the course.

Chapter 16

Random Walks (Gambler's Ruin)

Gambler's Capital

A gambler plays repeatedly, each time winning \$1 with probability p or losing \$1 with probability $q = 1 - p$. He starts with x dollars, and after n games possesses S_n dollars. Approximate S_n by a normal random variable.

Solution: The only difference here from Last Example of Chapter 15 is the given initial capital of x dollars. Hence, after n games the gambler has

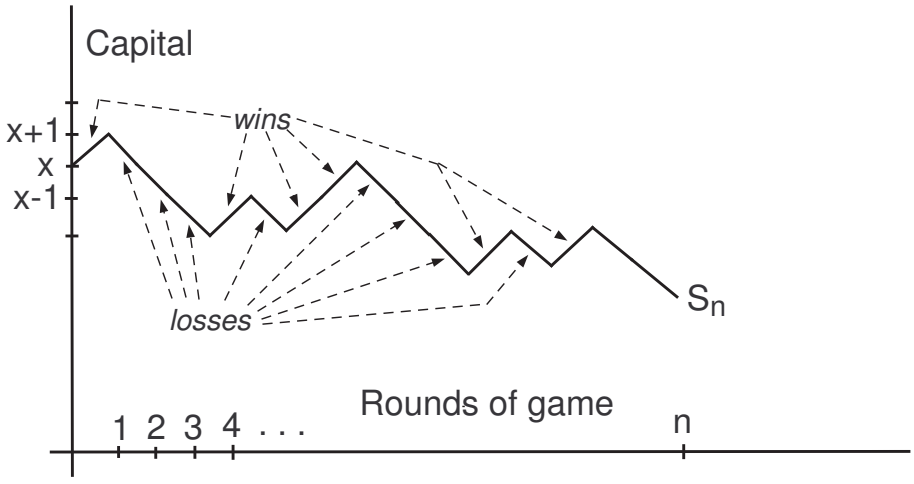
$$S_n = x + X_1 + \cdots + X_n \tag{29}$$

dollars. Therefore

$$\mathbb{E}(S_n) = x + (p - q)n \qquad \text{and} \qquad \mathsf{Var}(S_n) = 4pqn$$

and

$$S_n \approx \mathcal{N}\big(x + (p - q)n, 4pqn\big). \tag{30}$$



Fair Game (Symmetric Case $p = q$)

It is interesting to see how S_n evolves in the distant future, i.e. asymptotically, as $n \to \infty$. We first examine the "fair game" situation, where $p = q = 1/2$. It is a symmetric process, where S_n is equally likely to go either way, up or down. The expression (30) on page 121 takes form

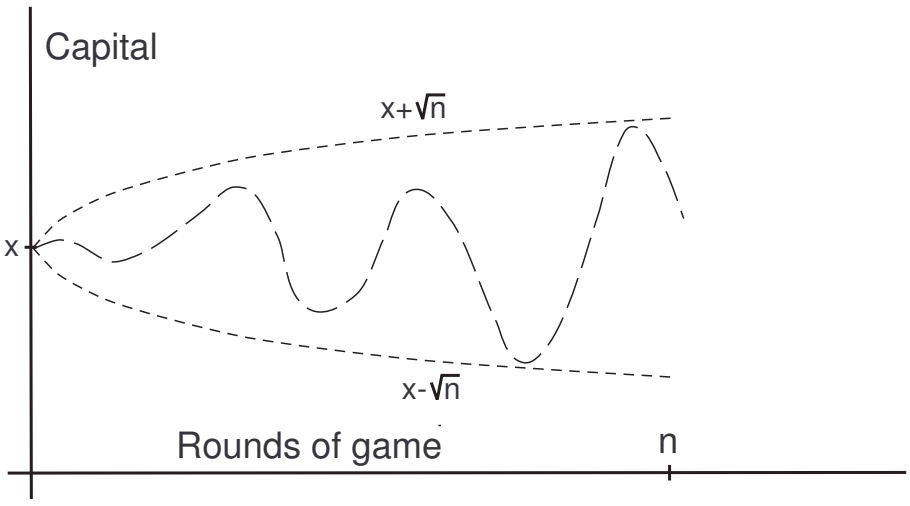
$$S_n \approx \mathcal{N}(x, n). \tag{31}$$

Hence, S_n is approximately normal with a constant mean value ($= x$) and a growing standard deviation $\sigma = \sqrt{n}$. This means that:

(a) on the average, S_n remains unchanged (the gambler does not win or lose)

(b) the typical values of S_n are $x \pm \sqrt{n}$, i.e. total gains or losses grow as $\sqrt{n}$.

The typical values of S_n are getting farther and farther away from x (equally likely in both directions) as n grows. Loosely speaking, S_n "drifts away" from x, then comes back and drifts in the opposite direction, comes back again, etc., each time its journeys away from x are going farther and getting longer in time. This type of process is called "diffusion" in physics. The evolution of S_n resembles the diffusion of a molecule in a dense gas.



Unfair Game (Asymmetric Case $p \neq q$)

We now examine the "unfair game" situation, when $p \neq q$. The evolution of S_n is described by (30) on page 121. The cases $p > q$ and $p < q$ are very similar to each other, we will only look at the case $p > q$, i.e. where the gambler is more likely to win than lose. (Not realistic, eh? Well, think of the casino owner then – in each game played by the customers, the casino is more likely to win than lose.)

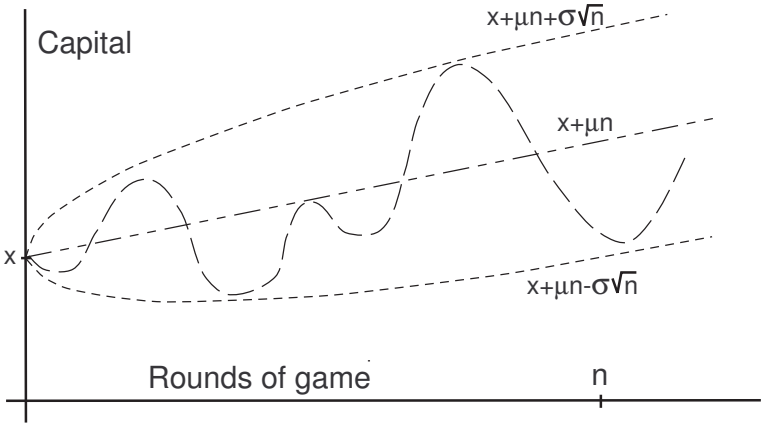
Put $\mu = p - q$ (mean gain per game) and $\sigma = \sqrt{4pq}$. Then (30) reads

$$S_n \approx \mathcal{N}(x + \mu n, \sigma^2 n).$$

Now S_n is approximately normal with a *growing* mean value $x + \mu n$ and a growing standard deviation $\sigma\sqrt{n}$. That is,

$$S_n \approx x + \mu n \pm \sigma\sqrt{n}$$

Thus S_n grows on the average, and its typical deviations from the average keep growing, too. Typically, S_n goes up, but not monotonically or steadily, it fluctuates up and down, as it grows. As time goes on, its average value grows, but the fluctuations, too, are getting larger and more violent. Pretty much like the stock market...



From the point of view of the casino owners, these are two competing processes: one (growth on the average) is good, it makes them richer and their business stronger. The other (growing fluctuations) is bad, it brings risk: a random downturn may erase their earnings or even ruin them. A vital question is then: Which process is stronger:

(i) *the steady growth of the average* or (ii) *growing random fluctuations?*

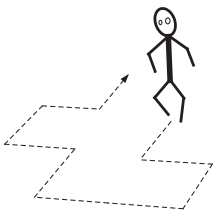
The ultimate answer is: the steady growth is always stronger. The mean value grows as $x+\mu n$, i.e. *linearly* in n. The fluctuations (typical deviations) grow as $\sigma\sqrt{n}$, i.e. as $n^{1/2}$, which is much slower than n, because $n^{1/2} \ll n$ for large n.

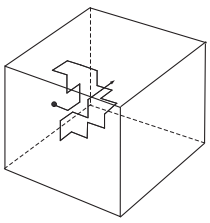
Employing the language of physics, S_n behaves as a molecule in a dense gas that is slowly blown by a light wind in a certain direction with a constant speed $\mu > 0$. Then the molecule slowly drifts with the gas and at the same time wonders about randomly between other molecules. Thus the drift is combined with diffusion.

Random Walks

In probability theory, another pictorial description of the above process is more customary. A drunk person walks on a street, making each step randomly, either forward with probability p or backward with probability q. If he starts at a point x, then his position after n (random) steps will be S_n.

This is a random walk on a line. One can make a random walk on a 2D surface (xy plane), where a drunk makes a step forward or backward or sideways (either left or right) with, say, the same probability 1/4. A more advanced (but difficult to realize in practice) model is a random walk in 3D space: a drunk person (or let us just say, a moving particle) makes steps forward or backward, left or right, up or down, with the same probability 1/6. This model describes, quite accurately, the motion of a molecule in a real 3D gas.





Restricted Random Walks

In practice, random walks are often restricted. A gambler cannot afford to have his capital S_n drop below zero - when he loses his initial capital (i.e. when $S_n = 0$), then his gambling is over for him. Most gamblers (at least the wiser ones) set an upper limit, too, – they decide in advance to quit when they reach a certain level $S_n = b$ (the "goal of the day").

Hence, some random walks have to stop as they reach certain limits: a lower bound = a, or an upper bound = b, or both. (In practice often the lower bound a is zero, but we do not require that.) Such random walks are said to be *restricted*. If just one bound is set (lower or upper), then we have a one-sided restricted random walk.

Parameters of Restricted Random Walks

For a restricted random walk, the normal approximation (30) does not fully apply. In fact, if the random walk has two restrictions, $a \le S_n \le b$, then the normal approximation tells us that sooner or later S_n will hit one of the bounds, it cannot stay in the interval (a, b) forever. Hence, a two-sided restricted random walk necessarily stops. It stops at some (random) time $T \ge 1$ and its value S_T is final, it will never change again. Such a walk can be then characterized by three parameters: the probabilities $P_a = \mathbb{P}(S_T = a)$ and $P_b = \mathbb{P}(S_T = b)$ of stopping at a and b, respectively, and the mean lifetime $\mathbb{E}(T)$. We note that since the random walk stops either at a or at b, we have the relation

$$P_a + P_b = 1.$$

If the random walk has just one restriction, say, $S_n \ge a$, then it may either stop at a at a random time T with probability P_a, or evolve indefinitely (without coming down to a). Again, the mean lifetime $\mathbb{E}(T)$ is a relevant parameter.

Wald's Identity

According to formula (29) on page 121, at the stopping time $n = T$ we have

$$S_T = x + X_1 + \cdots + X_T.$$

Recall also that $\mathbb{E}(X) = p - q = \mu$. The following equation then looks natural (but we omit the argument):

$$\mathbb{E}(S_T) = x + \mu \cdot \mathbb{E}(T). \tag{32}$$

It is known as *Wald's identity*. In the case of two restrictions, S_T only takes values a and b, hence $\mathbb{E}(S_T) = aP_a + bP_b$, so Wald's identity reads

$$aP_a + bP_b = x + \mu \cdot \mathbb{E}(T). \tag{33}$$

This is a very helpful relation.

Two Sided Restrictions: Symmetric Case

In the case $p = q = 1/2$ we have $\mu = 0$, so the Wald's identity (33) reads

$$aP_a + bP_b = x.$$

We also have $P_a + P_b = 1$, so solving these two equations for P_a and P_b gives

$$P_a = \frac{b-x}{b-a} \qquad \text{and} \qquad P_b = \frac{x-a}{b-a}.$$

The mean lifetime $\mathbb{E}(T)$ can be found from another Wald's identity:

$$\mathsf{Var}(S_T) = \sigma^2\, \mathbb{E}(T).$$

From this, it is easy to find (we omit details) that

$$\mathbb{E}(T) = (x-a)(b-x).$$

Example

You play with a friend by flipping a coin. If it comes up Heads, you win \$1, otherwise you lose \$1. You start with \$10 and plan to stop when your capital is \$50. What is the probability that you will reach your goal? What is the mean number of times you will play?

Solution: We assume the coin is fair, so that $p = q = 1/2$. We have $x = 10$, $a = 0$ (obviously) and $b = 50$. Then the probability of winning is

$$P_{50} = \frac{10-0}{50-0} = 0.2.$$

So, your chances to hit \$50 are not so high, just 20%. The mean number of games is $\mathbb{E}(T) = (10-0)(50-10) = 400$. Quite a long affair, too!

Suppose you want to increase your chances to win and decide to bet \$5 instead of \$1 each time. Does it help? Now it is convenient to treat \$5 as a unit (one step). Then, in these new units, $x = 2$, $a = 0$ and $b = 10$. So we have $P_{\text{win}} = (2-0)/(10-0) = 0.2$. The same as before! You can easily check that even if you bet \$10 each time, nothing will change, still $P_{\text{win}} = 0.2$.

Two Sided Restrictions: Asymmetric Case

In the case $p \neq q$ we have $\mu \neq 0$, so the Wald's identity (33) gives

$$\mathbb{E}(T) = \frac{aP_a + bP_b - x}{p - q}. \tag{34}$$

It remains to find P_a and P_b, for which special formulas exist:

$$P_a = \frac{(q/p)^{x-a} - (q/p)^{b-a}}{1 - (q/p)^{b-a}} \qquad \text{and} \qquad P_b = \frac{1 - (q/p)^{x-a}}{1 - (q/p)^{b-a}}.$$

(in practice, it is easier to compute P_b as above and then $P_a = 1 - P_b$.)

Example

You play roulette in a casino. A roulette wheel has 18 red spots and 18 black spots and 2 green spots. You can bet \$1 on red or black. If it comes up green, the casino wins either way. So, your chances to win are 18/38=9/19. You start with \$10 and plan to stop when your capital is \$50. What is the probability that you win? What is the mean number of times you will play?

Solution: We have $p = 9/19$ and $q = 1 - p = 10/19$. The game is asymmetric, i.e. unfair, but it seems to be unfair just slightly... Well, let us compute the chances. As before, we have $x = 10$, $a = 0$ and $b = 50$. Then the probability of winning is

$$P_{50} = \frac{1 - (10/9)^{10}}{1 - (10/9)^{50}} = 0.0097$$

and the mean number of games is

$$\mathbb{E}(T) = \frac{0 \cdot (1 - 0.0097) + 50 \cdot 0.0097 - 10}{-1/19} = 180.8.$$

Now compare these results to those in the previous example. The chances to win drop from 20% to below 1%! Though the game seems to be just *slightly* unfair, in the end it turned out to be an almost certain ruin of the gambler... The good news is that it will be over much sooner, after just 180 rounds instead of 400...

Suppose you want, as in the previous example, to increase your chances to win and decide to bet \$5 instead of \$1 each time. Does it help now? Again, we treat \$5 as a unit (one step). Then, in these new units, $x = 2$, $a = 0$ and $b = 10$. So we have

$$P_{\text{win}} = \frac{1 - (10/9)^2}{1 - (10/9)^{10}} = 0.1256.$$

Notice a dramatic improvement to over 12% from under 1%. Better yet, you can bet \$10 each time and get

$$P_{\text{win}} = \frac{1-(10/9)}{1-(10/9)^5} = 0.1602.$$

Wow! This is 16%, almost as high as 20% in the *fair* game of the previous example. The moral of this story is this: if you have to risk in an unfavorable situation, when the odds are against you, risk "big". The longer you play (trying to achieve your goal in small steps), the more odds against you accumulate, and you will almost certainly lose.

One Sided Restrictions: Symmetric Case

We now consider a one-sided restriction $S_n \geq a$ (a lower bound is set, but no upper bound). The other case, when only an upper bound is set, is symmetric and left as an exercise.

In the case $p = q = 1/2$ (a symmetric walk) we use the formulas for P_a and $\mathbb{E}(T)$ on page 126 and take the limit as $b \to \infty$ (the logic is: moving the upper bound b to infinity will effectively eliminate it and give us a walk with one-sided restriction). We obtain

$$P_a = 1 \qquad \text{and} \qquad \mathbb{E}(T) = \infty$$

This means that in a fair game with one restriction, the random walk hits the lower bound a and stops, sooner or later. This is consistent, by the way, with the diffusive character of the random walk observed on page 122: the deviations from the mean value x become longer and longer and go both ways, up and down. Hence, no matter where the bound is set, it will be hit eventually terminating the walk.

A surprise comes with the formula $\mathbb{E}(T) = \infty$. This means that in practice, it takes arbitrary long (one can say, "indefinitely long") to stop a symmetric random walk with one restriction. If you are lucky, the walk will drift to the bound and hit it. But it may well drift in the opposite direction and stay there very, very long time.

One Sided Restrictions: Asymmetric Case

Here we examine a one-sided restriction $S_n \geq a$, but $p \neq q$ (an asymmetric walk). We use the formulas for P_a and $\mathbb{E}(T)$ on page 127 and take the limit as $b \to \infty$. There are two distinct cases here.

Assuming $p < q$ we obtain

$$P_a = 1 \qquad \text{and} \qquad \mathbb{E}(T) = \frac{x-a}{q-p}$$

The first result comes at no surprise: if the chances to lose (step down) are higher than the chances to win (step up), then sooner or later the lower bound $S_n = a$ will be hit (this was so even in the symmetric case $p = q$). But now it will not take indefinitely long time: the average lifetime is finite.

Assuming $p > q$ we obtain

$$P_a = (q/p)^{x-a} \qquad \text{and} \qquad \mathbb{E}(T) = \infty$$

Now the random walk does NOT have to hit the lower bound! With a positive probability, $1 - P_a$, it may stay above it and live forever. The average lifetime is, obviously, infinite. This is consistent with the "drift+diffusion" model of the random walk given on page 123: the drift upward is stronger than the diffusion, so it takes the values S_n up to infinity eventually. If the walk escapes the deadly encounter with the bound a during the early period (when it may drift dangerously close to a), it will live forever.

Example

A person plans to open a casino with just one roulette and allow the customers to bet \$1 each round. The prospective owner wants to minimize risk and deposit an initial capital x, so that his chances to go broke (hit zero) will be less that 0.01%. How much money does he need to deposit before he opens the business?

Solution: The casino owner wins when the customer loses, i.e. with probability $p = 10/19$, see 16.10. Hence, $q = 1 - p = 9/19$. So we have $P_0 = (q/p)^x = 0.9^x$. We need $P_0 < 0.0001$. Equating $0.9^x = 0.0001$ we get $x = 87.4$. Therefore, an initial capital of \$88 will suffice.

Notice how small an initial capital is required to secure an almost guaranteed success when the odds are in your favor! Even when the game looks "almost" fair to the other party (their chances in each game are $9/19 = 47.4\%$, just slightly below 50%), the bias in favor of the casino owner accumulates from game to game and practically denies the customers any chance in the end.

The rest of Chapter 16 is optional material.

Hit Probabilities in Symmetric Random Walk

Suppose a symmetric random walk (with $p = q = 0.5$) starts at x and has two-sided restrictions $a \leq S_n \leq b$. We want to find the probability $P(x, y)$ that it hits another point $y \neq x$ before it stops.

Suppose that $y > x$. Then we note that the random walk S_n only has two options: hit a and stop before reaching y, and hit y (of course without hitting a earlier). These are the same options as for a random walk with restrictions at a and y (instead of b). By the equations on page 126 we have

$$P(x, y) = \frac{x - a}{y - a}$$

Suppose that $y < x$. In a similar way, the random walk S_n only has two options: hit b and stop before reaching y or hit y (of course without hitting b earlier). These are the same options as for a random walk with restrictions at b and y (instead of a). By the equations on page 126 we have

$$P(x, y) = \frac{b - x}{b - y}$$

Example

You play with a friend by flipping a coin. If it comes up Heads, you win \$1, otherwise you lose \$1. You start with \$10 and plan to stop when your capital is \$50. What is the probability that you ever hit \$40?

Solution. We have

$$P(10, 40) = \frac{10 - 0}{40 - 0} = 0.25$$

Return Probabilities in Symmetric Random Walk

Suppose a symmetric random walk starts at x and has two-sided restrictions $a \le S_n \le b$. We want to find the probability $P(x,x)$ that the walk ever returns to x (before stopping at either a or b).

After starting at x, the walk jumps either to $x-1$ or to $x+1$, with the same probability $1/2$. Now we can use the formulas $P(x-1,x)$ and $P(x+1,x)$ developed in 16.14 to find the probability to hit x again:

$$\begin{aligned} P(x,x) &= \frac{1}{2}\cdot P(x-1,x) + \frac{1}{2}\cdot P(x+1,x) \\ &= \frac{1}{2}\cdot\frac{x-a-1}{x-a} + \frac{1}{2}\cdot\frac{b-x-1}{b-x} = 1 - \frac{b-a}{2(b-x)(x-a)} \end{aligned}$$

Example (continued)

In the previous example, what is the probability that you ever have exactly \$10 again?

Solution. We have

$$P(10,10) = 1 - \frac{50-0}{2(50-10)(10-0)} = \frac{15}{16}$$

Number of Returns in Symmetric Random Walk

Now we want to find the mean number $G(x,x)$ of returns to x (before the walk stops at either a or b).

After starting at x, the walk can return to x with probability $P(x,x)$. If it does return to x, it will evolve again starting from x, as if nothing happened before. Hence, again the probability of return to x is $P(x,x)$. So, considering successive returns to x, we see that after each return the walk can return again with probability $P(x,x)$ or stop (die) with probability $1-P(x,x)$. It is therefore a sequence of trials till the first success – the trials are returns and the "success" is the termination (death) of the random walk before another return occurs. One can conclude that the number of returns plus one is a geometric random variable. Therefore, its mean value is

$$G(x,x) = \frac{1}{1-P(x,x)} - 1 = \frac{P(x,x)}{1-P(x,x)}$$

Example (continued)

In the previous example, what is the mean number of times that you get back to exactly \$10 before the game ends either way?

Solution. We have

$$G(10,10) = \frac{15/16}{1-15/16} = 15$$

Note: it was not a good idea to set such a high goal (\$50) in the first place: this was a fair game where you started with just a ten. Your chances to win the entire match were 20% (page 126). But now we see that before you lose, you will come back to \$10 as many as 15 times (on the average), so you will have enough time to reconsider your goal...

Returns in Symmetric Random Walk Without Restrictions

Suppose a symmetric random walk starts at x with no restrictions on either side. We want to find the probability $P(x,x)$ of ever coming back to x and the mean number $G(x,x)$ of returns.

We simply take the limit as $a \to -\infty$ and $b \to \infty$ in the expressions obtained above. We get

$$P(x,x) = 1 \qquad \text{and} \qquad G(x,x) = \infty$$

This means that the random walk starting at x will come back with probability one, and it will do so infinitely many times. For this reason, the random walk is said to be *recurrent*.

In fact, the recurrence is consistent with the diffusive character of the random walk observed on page 122: the deviations from x must always go both ways, up and down. In order to go from above x to below x or vice versa the walk has to cross the point x.

Returns in Symmetric Random Walk in 2D and 3D

On page 124 we described random walks in plane (2D) and space (3D). In each case the walk has the same probability to jump in each available direction (1/4 on the plane and 1/6 in the space), so there is a complete symmetry. The diffusive character of the walk consists of growing deviations from x in all possible directions, as time goes on.

But now, unlike the 1D walk, the 2D and 3D walks do *not* have to cross x in order to change the direction of deviation: they can come back close to x, go around x, and then evolve in another direction... So, it is not quite clear whether the symmetric random walk in 2D or 3D is recurrent or not. You can make your best guess.

The answer is that the 2D walk (on the plane) is still *recurrent*, the walk comes back exactly to the starting point x with probability one, and it does so infinitely many times.

But the 3D walk (in the space) is *not recurrent* anymore. In 3D, the probability to come back to the starting point x is less than one (it is about 35%) and the average number of returns is not infinite (it is actually quite small – just 0.5 returns, on the average).

This brings up a philosophical question: why do we live in a 3D world? Is there any substantial difference between the 3D world and the 2D world, except the obvious lack of one dimension in the latter? The probability theory gives one substantial but not obvious difference: the non-recurrence of 3D random walks. Maybe this has something to do with the physics of gases and fluids...

Chapter 17

Poisson Process (optional)

Reminder

We introduced Poisson Process on page 45. Here we study it more deeply and formally.

Recall that *Poisson process* is a sequence of random points on a line such that the locations of those points is random and even their number in any given interval is random. The average number of those points per unit length is denoted by $\lambda > 0$ and is called the density or *rate*; it is a numerical parameter of the whole process.

Number of Points within Intervals

Let (a, b) be a given interval (segment) on the line. The number of random points of the process in this segment $N_{(a,b)}$ is a Poisson random variable with parameter $\lambda(b - a)$:

$$N_{(a,b)} = \text{poisson}(\lambda(b - a))$$

Also, if (a, b) and (c, d) are two disjoint (non-overlapping) intervals, then $N_{(a,b)}$ and $N_{(c,d)}$ are independent random variables.

Waiting Times (inter-arrival times)

Intervals between successive points in a Poisson process are called *waiting times* or *inter-arrival times*. If $0 < P_1 < P_2 < \cdots$ are the successive points, then $W_1 = P_1$, $W_2 = P_2 - P_1, \ldots$ are waiting times. The term is motivated by the applications where calls or customers arrive at random times, and between successive arrivals the business "waits".

Each W_k is an exponential random variable with parameter λ. The parameter does not depend on k, so all waiting times have the same exponential distribution, i.e. each W_k is exponential(λ). Also, the waiting times $W_1, W_2, \ldots$ are independent.

Meaning of λ

Let $(a, a+T)$ be a long interval in the Poisson process. The number of points on this interval is a Poisson random variable with parameter λT. Its mean value is λT. So, we expect, on the average, λT points on the interval $(a, a+T)$. If we have $N \approx \lambda T$ points on the given interval, they partition it into $\approx \lambda T$ subintervals (waiting times). Hence, the average length of a waiting time is expected to be $T/(\lambda T) = 1/\lambda$. Is it correct? Yes, the waiting times between successive intervals are exponential random variables with parameter λ. The mean value of such a random variable is exactly $1/\lambda$. So, all our estimates are consistent.

Example

On a long stretch of a highway, accidents occur at a rate of one per 20 miles. You drive a car on this highway and pass two accidents in a row. What are chances that no more accidents occur within the next 40 miles?

Solution. The waiting times are independent of each other, so it does not matter how many accidents you have passed. The probability that the interval to the next accident ("the waiting time") is longer than 40 miles is

$$\mathbb{P}(W > 40) = 1 - F_W(40) = 1 - (1 - e^{-\lambda \cdot 40}) = e^{-2}$$

since $\lambda = 1/20$ (one accident per 20 miles).

Note: if you enter the highway at any point, the distance from your entry point to the nearest accident would be the same, an exponential random variable with parameter $\lambda = 1/20$.

Example (continued)

What are the chances that on a given stretch of 10 miles of the highway more than one accident occur?

Solution. The number of accidents N on that stretch of the highway is a Poisson random variable with parameter $10/20 = 0.5$. Hence,

$$\mathbb{P}(N > 1) = 1 - \mathbb{P}(N = 0) - \mathbb{P}(N = 1) = 1 - e^{-0.5} - 0.5e^{-0.5} = 0.09$$

Paradox

Suppose that, as in the previous example, accidents on an east-west highway occur at a rate of one per 20 miles. Hence, the intervals between accidents are exponential random variables and the average interval is 20 miles.

Now suppose you enter the highway at some point P. The distance from your entry point to the next accident to the east, call it W_e, is a "waiting time", so it has an exponential distribution with the mean value 20. At the same time, the distance from your entry point to the next accident to the west, call it W_w, is also a "waiting time", so it has an exponential distribution with the mean value 20, too. Hence, their sum $W_e + W_w$ has the mean value $20 + 20 = 40$.

One the other hand, the sum $W_e + W_w$ is exactly one interval between two successive accidents! Hence, it is a "waiting time" itself. So, its mean value must be 20, rather than 40. What is wrong?

Paradox Solved

Actually, $W_e + W_w$ has mean value 40, rather than 20. Why? Because of the way we select that interval. Each interval between successive accidents has mean value 20, indeed, but some intervals are smaller and some other intervals are larger. When you enter the highway at some point, you are more likely to hit a larger interval between successive accidents than a smaller interval. This is quite clear. Therefore, this is not a random interval, it is an interval selected with some preference, the choice is "biased" toward longer intervals.

Making a preferred selection puts additional restrictions on the probability distribution and changes the mean value (and everything else).

Poisson Process in Plane and Space

Above we discussed Poisson processes on a line. One can consider Poisson processes on a plane or in space. If random points occur on a plane, with a given average density of λ (per unit area), then we have a Poisson process on the plane. Examples: accidents (or fires or crimes) that occur in a big city, mushrooms that grow in a forest, etc.

If random points occur in space, with a given average density of λ (per unit volume), then we have a Poisson process in space. Examples: molecules in a gas or a fluid, explosions in the air during fireworks, etc.

Number of Points in Given Region

Let R be a given region on the plane (or in space) where a Poisson process with the average density λ occurs. Denote by N_R the number of random points of the process in this region. Then N_R is a Poisson random variable with parameter $\lambda|R|$. Here $|R|$ is the area of R on the plane (or the volume of R in space). So we have this:

$$N_R = \text{ poisson}(\lambda|R|)$$

Also, if R_1 and R_2 are two disjoint (non-overlapping) regions, then N_{R_1} and N_{R_2} are independent random variables.

Example

Let R_1 and R_2 be two *overlapping* regions. Then the random variables N_{R_1} and N_{R_2} are dependent. Find their covariance $\mathsf{Cov}(N_{R_1}, N_{R_2})$.

Solution. Denote by $D_0 = R_1 \cap R_2$ the common part of R_1 and R_2. Let also $D_1 = R_1 \setminus D_0$ and $D_2 = R_2 \setminus D_0$. Now all the three regions D_0, D_1, and D_2 are disjoint. So, the random variables N_{D_0}, N_{D_1}, and N_{D_2} are independent. Also, obviously, $N_{R_1} = N_{D_0} + N_{D_1}$ and $N_{R_2} = N_{D_0} + N_{D_2}$. Therefore,

$$\begin{aligned} \mathsf{Cov}(N_{R_1}, N_{R_2}) &= \mathsf{Cov}(N_{D_0} + N_{D_1}, N_{D_0} + N_{D_2}) \\ &= \mathsf{Cov}(N_{D_0}, N_{D_0}) + 0 + 0 + 0 = \mathsf{Var}(N_{D_0}) = \lambda|D_0| \end{aligned}$$

In the last step we used the fact that the variance of a Poisson random variable with parameter λ was equal to λ.

Example

Suppose we have a Poisson process in space with average density λ. Pick a point O in space (call it the origin) and let X be the distance from O to the nearest point in the process. Find the distribution function of the random variable X.

Solution. We have

$$F_X(x) = \mathbb{P}(X \leq x) = 1 - \mathbb{P}(X > x)$$

The condition $X > x$ means that the nearest point of the Poisson process is farther than x (units of length) from the origin O. That is, the ball of radius x, call it B_x, contains no points of the process. Hence,

$$F_X(x) = 1 - \mathbb{P}(N_{B_x} = 0) = 1 - e^{-\lambda |B_x|} = 1 - e^{-\frac{4}{3}\lambda \pi x^3}$$

In the last step we used the fact from elementary geometry that the volume of a ball of radius x was $\frac{4}{3}\pi x^3$.

Chapter 18

Additional Reading for Graduate Students

Note: this material is for MA 585 students. Read it before you attempt graduate homework exercises.

Stirling's Formula

Recall that if X is a binomial random variable, $b(n, p)$, then its probabilities are given by

$$\mathbb{P}(X = m) = \frac{n!}{m!\ (n-m)!}\ p^m\, q^{n-m}, \qquad \text{for } \ m = 0, 1, \ldots, n$$

where $q = 1 - p$. The factorials here are extremely hard to compute in practice, when m and n are large. The Stirling's formula helps a lot:

$$n! \approx \sqrt{2\pi n}\ \left(\frac{n}{e}\right)^n.$$

Extra note: A little more precise version of Stirling's formula is

$$n! = \sqrt{2\pi n}\ \left(\frac{n}{e}\right)^n e^{\varepsilon_n},$$

where

$$\frac{1}{12n+1} < \varepsilon_n < \frac{1}{12n},$$

but we will not use this.

Let us see how Stirling's formula helps to compute the binomial probabilities. Denote $y = m/n$. Then the binomial probabilities become

$$\mathbb{P}(X = m) \approx \frac{1}{\sqrt{2\pi y(1-y)n}}\ \left(\frac{p^y(1-p)^{1-y}}{y^y(1-y)^{1-y}}\right)^n.$$

Note: you cannot just use this formula for your homework, you have to derive it from Stirling's formula.

Continuity of Probabilities

In theoretical analysis of probabilities, one often has to deal with an infinite collection (or sequence) of events $A_1, A_2, \ldots$, and take a limit as $n \to \infty$.

We say that the sequence of A_n's is *increasing* if $A_1 \subset A_2 \subset \cdots$, i.e. each event is larger than the previous one (A_n contains A_{n-1}). For an increasing sequence of events, their probabilities $\mathbb{P}(A_n)$ grow with n and approach the probability of their union, i.e.

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\big(\cup_{n=1}^{\infty} A_n\big).$$

Similarly, we say that the sequence of A_n's is *decreasing* if $A_1 \supset A_2 \supset \cdots$, i.e. each event is smaller than the previous one (A_n is contained in A_{n-1}). For an decreasing sequence of events, their probabilities $\mathbb{P}(A_n)$ get smaller with n and approach the probability of their intersection, i.e.

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\big(\cap_{n=1}^{\infty} A_n\big).$$

This two laws are called the *continuity of probabilities*.

As an example, let X be a uniform random variable $U(0,1)$. Let us find the probability $\mathbb{P}(X \text{ is rational})$. This seems to be a complicated question. (One naïve answer: since X is either rational or irrational, then the probability of each of these two events is 0.5. This answer is incorrect.)

Let $A_1 = \{X = 1/2\}$. Obviously, $\mathbb{P}(A_1) = 0$, because X is a continuous random variable, hence it takes each value with probability zero. Let

$$A_2 = \{X = 1/3 \text{ or } 1/2 \text{ or } 2/3\}$$

You can easily show that $\mathbb{P}(A_2) = 0$ as well (do that!). Now let

$$A_3 = \{X = 1/4 \text{ or } 1/3 \text{ or } 1/2 \text{ or } 2/3 \text{ or } 3/4\}$$

Again, you can easily show that $\mathbb{P}(A_3) = 0$. Note the pattern and guess what A_4 should be, and then A_n for any $n \geq 1$. Argue that $\mathbb{P}(A_n) = 0$ for *every* $n \geq 1$. Lastly, note that $A_1 \subset A_2 \subset A_3 \cdots$ and their union $\cup_{n=1}^{\infty} A_n$ is exactly the event $\{X \text{ is rational}\}$. Finally, find the limit

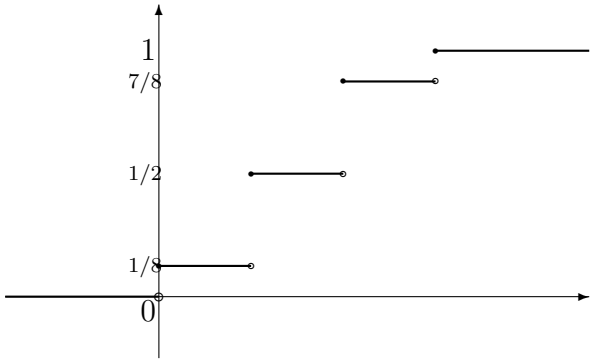
$$\mathbb{P}(X \text{ is rational}) = \mathbb{P}\big(\cup_{n=1}^{\infty} A_n\big) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

Distribution Function for Discrete Random Variables

Here we describe distribution functions for discrete random variables. For example, let X be the number of Heads in three tosses of a coin (page 25).

We compute its distribution function $F(x)$ as follows. Say, $F(-1) = 0$, because the event $X \leq -1$ is impossible. Next, $F(0) = 1/8$, because the event $X \leq 0$ occurs only when $X = 0$. Then, $F(0.6) = 1/8$, as the event $X \leq 0.6$ occurs only if $X = 0$. Then, $F(1) = 1/2$ because the event $X \leq 1$ occurs if $X = 0$ or $X = 1$, i.e. two values of X ($X = 0$ and $X = 1$) are covered by this event, etc.

The resulting graph of $F(x)$ is shown below.



Distribution function of X.

The graph is a "staircase" with "steps" going up from left to right. The steps are straight horizontal segments. Left endpoints are included (shown by solid circles), right endpoints are excluded (shown by hollow circles). The function has discontinuities ("jumps") at all values that are actually taken by X, i.e. at 0, 1, 2, 3. The height of each jump is the corresponding probability, i.e. the jump at $X = x$ equals $\mathbb{P}(X = x)$.

These are general rules for distribution functions of discrete random variables. Note that such variables have no density functions in the sense of Chapter 5 (because $F(x)$ cannot be differentiated at discontinuity points).

Transformations of Pairs of Random Variables

Suppose X and Y are two random variables with a joint density function $f_{X,Y}(x,y)$. Let D denote the domain of possible values of these random variables, i.e. the domain (in the xy-plane) where $f_{X,Y}(x,y) > 0$. Next, let $u(x,y)$ and $v(x,y)$ be two functions, each with two arguments, x and y. Define two new random variables by

$$U = u(X,Y) \qquad \text{and} \qquad V = v(X,Y).$$

We want to find the joint density $f_{U,V}(u,v)$ of the pair, U and V.

Generally, this is difficult. But it is relatively easy under one condition: the functions u, v are one-to-one in the following sense: for any pair of numbers U, V the system of equations

$$U = u(X,Y) \qquad \text{and} \qquad V = v(X,Y)$$

has *at most one* solution in the domain D; i.e. there is at most one pair of numbers $(X,Y) \in D$ that satisfies both equations.

Now the density $f_{U,V}(u,v)$ of U and V is given by

$$f_{U,V}(u,v) = \frac{f_{X,Y}(x,y)}{J(x,y)}.$$

Here x, y is the unique pair in D that satisfies the two equations

$$u = u(x,y) \qquad \text{and} \qquad v = v(x,y)$$

and $J(x,y)$ is the so-called Jacobian factor. To compute J, you need to calculate the 2×2 matrix of partial derivatives

$$M = \begin{bmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial v/\partial y \end{bmatrix},$$

then J would be the absolute value of its determinant:

$$J(x,y) = \big|\det M\big|.$$

This rule can be explained as follows. The density of a pair of random variables can be thought of as the "probability per unit area", and the Jacobian J is the factor by which the area changes when the x, y plane is transformed onto the u, v plane.

How do we check that the functions u, v are one-to-one? It is enough to check that J cannot be zero within the domain D, i.e., $J(x, y) \neq 0$ for all $(x, y) \in D$.

Example

Suppose X and Y are independent uniform random variables, both $U(0, 1)$. Define two new random variables $U = X + Y$ and $V = \frac{X}{X+Y}$. Find a formula for the joint density function $f_{U,V}(u, v)$.

Solution: The joint density function $f_{X,Y}(x, y)$ is

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) = 1 \cdot 1 = 1$$

for $0 < x < 1$ and $0 < y < 1$, i.e., in the unit square $D = \{0 < x, y < 1\}$.

Next, our functions are $u = x + y$ and $v = \frac{x}{x+y}$. The matrix of partial derivatives is

$$M = \begin{bmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial v/\partial y \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \frac{x}{(x+y)^2} & -\frac{y}{(x+y)^2} \end{bmatrix}$$

Its determinant is

$$\det M = -\frac{x+y}{(x+y)^2} = -\frac{1}{x+y} = -\frac{1}{u}$$

Since $x, y > 0$ in D, we have $\det M < 0$ in D, therefore

$$J = \left|\det M\right| = \frac{1}{x+y} = \frac{1}{u}$$

Note that $J \neq 0$ in D, so our functions are one-to-one, as required.

Finally, the formula for the joint density $f_{U,V}(u, v)$ is

$$f_{U,V}(u, v) = \frac{f_{X,Y}(x, y)}{J(x, y)} = \frac{1}{1/(x+y)} = x + y = u.$$

Independent Random Variables

Two random variables X and Y are independent if for any $a < b$ and $c < d$ we have

$$\mathbb{P}(a < X \le b \text{ and } c < Y \le d) = \mathbb{P}(a < X \le b) \cdot \mathbb{P}(c < Y \le d),$$

which precisely means that X and Y take their values independently. The above equation can be simplified if X and Y are both discrete or both continuous. If X and Y are discrete, then their independence means that for any possible value a of X and for any possible value b of Y we have

$$\mathbb{P}(X = a \text{ and } Y = b) = \mathbb{P}(X = a) \cdot \mathbb{P}(Y = b).$$

This is what we used in example on page 62. If X and Y are continuous, then their independence means that their joint density function $f_{X,Y}(x, y)$ satisfies

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$$

for all $-\infty < x, y < \infty$. This is a necessary and sufficient condition for independence.

I.I.D. Random Variables

Let $X_1, \ldots, X_n$ be independent identically distributed (i.i.d.) random variables. Since they all have the same distribution and the same relation to each other (mutual independence), they play equal roles in any combination you make with them.

For example, let $U = X_1 + X_2^2$ be a new random variable, it has a certain distribution. If we interchange X_1 and X_2 and define another random variable $V = X_2 + X_1^2$, it would have the same distribution as U. They will have the same mean values, $\mathbb{E}(U) = \mathbb{E}(V)$, the same variance $\mathsf{Var}(U) = \mathsf{Var}(V)$, etc.

Another example: the random variables $U = \frac{X_1}{X_1+X_2}$ and $V = \frac{X_2}{X_1+X_2}$ have the same distribution, the same mean value, the same variance, etc.