

1 Combinatorics

1.1 Example. Five cards are labelled 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 right order: 1,2,3,4,5?

Answer: The number of ways is $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5! = 120$. The chance of the (only) right lineup is 1:120. If one plays a game betting \$1 and winning x dollars should the right lineup occur, then the game is fair if $x = 120$.

1.2 Rule 1 for Permutations. The number of ways to line up n objects is

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

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

1.3 Example. In a race of ten horses, you bet that horse A arrives first and horse B arrives second. What is the chance that your bet wins, assuming all the horses are equally likely to come first and second?

Answer: Only the first and second horse to arrive will matter. The number of ways to pick the winner and the runner-up from 10 horses is $10 \cdot 9 = 90$. The chance that your choice of horses (i.e., A first and B second) wins is then 1:90. If you bet \$1, then you would expect to win \$90, otherwise the game will be unfair.

1.4 Rule 2 for Permutations. The number of ways to select and order (line up) m objects from a pool of n objects is

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

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

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

Answer: The number of possible choices of a pair of cards is $10 \times 9 / 2 = 45$. Here we divide by two because each pair can be ordered in two ways, then we can argue as in Example 1.3.

1.6 Rule for Combinations. The number of ways to select m objects (without ordering) from a pool of n objects is

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

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

1.7 Remarks. It is standard to assume $0! = 1$. Note that

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

Note the symmetry rule:

$$\binom{n}{m} = \binom{n}{n-m} \tag{1.1}$$

The above number is just the number of ways to partition a pool of n objects into two groups, one of m objects and the other of $n-m$ objects.

1.8 Partitions. What is the number of ways to partition a pool of n objects into two groups (of arbitrary size)?

First solution: Since the size of the first group may take values $m = 0, 1, \dots, n$, and for each m we have the formula (1.1), then the total number 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 make the partition.

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

$$\sum_{m=0}^n \binom{n}{m} = 2^n \tag{1.2}$$

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

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

Note that (1.2) is a particular case of Newton's formula, obtained by the substitution $a = b = 1$.

1.10* Remark. Note that another substitution, $a = 1$ and $b = -1$, gives one more interesting formula:

$$\sum_{m=0}^n (-1)^m \binom{n}{m} = 0$$

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

$$\begin{array}{cccccccc}
 & & & & & & & 1 \\
 & & & & & & 1 & 1 \\
 & & & & & 1 & 2 & 1 \\
 & & & 1 & 3 & 3 & 1 \\
 & & 1 & 4 & 6 & 4 & 1 \\
 & 1 & 5 & 10 & 10 & 5 & 1 \\
 1 & 6 & 15 & 20 & 15 & 6 & 1 \\
 1 & 7 & 21 & 35 & 35 & 21 & 7 & 1
 \end{array}$$

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

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

Solution: The results of 100 tosses can be recorded by a string of H's and T's, for example HTTHHTHHH...T, of length 100. How many such strings do we have? It is 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? This is the number of ways to pick 50 positions out of 100 available (say, we pick 50 positions for H's, filling the rest by T's). This number is $C_{100,50}$. Hence, the chance to observe 50 Heads is

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

More generally, if one tosses a coin n times, then the chance to observe exactly m Heads is

$$\frac{1}{2^n} \binom{n}{m} \tag{1.3}$$

The numbers like above are very hard to compute for large m and n . Even a computer can fail due to an overflow or an underflow in the process of calculation. One of the goals in probability theory is to find effective ways to compute the above numbers approximately.

1.13 Question. Guess what the number $\frac{1}{2^{100}} \binom{100}{50}$ is, approximately. One naive idea: 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 is 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 ideas are well 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 to observe 50 Heads is 1:10, or 10%. The exact answer is 7.96%, we will arrive at it in Section 15.

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

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

Solution: Each die has six faces and shows a number from 1 to 6. 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 there 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 $4/36=1/9$.

1.16 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 take 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

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

1.17 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 within the committee. So, the total number is $k C_{n,k}$.

1.18* Committee of variable size. 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 people to his/her committee. The chairman will select a committee from the remaining $n - 1$ people, thus partitioning them into two groups – the committee itself and the rest of the group. By 1.8, 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 an interesting relation:

$$\sum_{k=1}^n k C_{n,k} = n 2^{n-1} \tag{1.4}$$

2 Probability Space

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

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

Solution: We know from 1.12 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:

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

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

2.2 Stubborn coin flipper. A stubborn person tosses a coin until it lands on Head. What are possible outcomes? What is the chance that three or more tosses will be necessary?

Solution: The coin may land on Head at once, or else it may land on Tail once or several times before it lands on Head. The experiment ends when Head appears. The possible outcomes are:

$$\text{H, TH, TTH, TTTH, } \dots, \underbrace{\text{T} \dots \text{T}}_{n-1} \text{H, } \dots$$

The probability of any string of T's and H's of length n is $1/2^n$. So, the above outcomes have the corresponding probabilities

$$1/2, 1/4, 1/8, 1/16, \dots, 1/2^n, \dots$$

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

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

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

$$\frac{1}{8} + \frac{1}{16} + \dots + \frac{1}{2^n} + \dots = \frac{1}{4}$$

2.3 Archery. One shoots at a target that 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. The outcome of this experiment is the point where the arrow lands. All the points on the surface of the target 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 on the target:

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

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

We summarize the above three examples. 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.

2.4 Probability space. The set of all possible outcomes of a random experiment is called a *probability space*, we denote it by Ω . Its elements, or points, are called outcomes, they are denoted by ω . The result of the random experiment is always one point ω of Ω .

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 . If ω happens to be outside of A , the event A does not occur. Each event has a probability, which is a number between 0 and 1. The probability of an event A is denoted by $P(A)$.

2.5 Rules for events and probabilities.

(a) The entire Ω is called a *certain* event. It always occurs because it contains

every possible outcome ω . So, its probability is one: $P(\Omega) = 1$.

(b) There is a special notation \emptyset for the event that never occurs. It contains no outcomes. Its probability is zero, $P(\emptyset) = 0$. The event \emptyset is said to be *impossible*. It is also called an empty set.

(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 doesn't, and vice versa. The probability of A^c is given by $P(A^c) = 1 - P(A)$.

(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). Then we have $P(A) \leq P(B)$.

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

2.6 Venn's diagrams. The above diagrams illustrate the rules of 2.5. The big rectangle always represents the probability space Ω . The disks inside the rectangle represent events A , B , etc.

2.7* De Morgan's laws. The following De Morgan's laws may be useful:

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

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

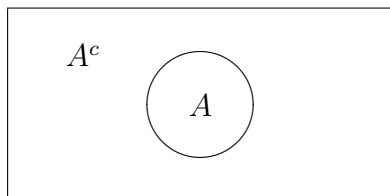
2.8* Distributive laws. The following distributive laws may be useful:

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

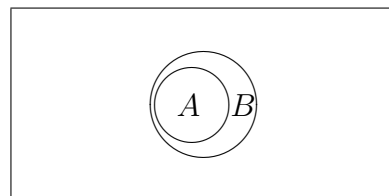
To verify these, draw three overlapping circles representing A, B, C and shadow the related areas.

2.9 Summation rule. For two events, A and B , we have

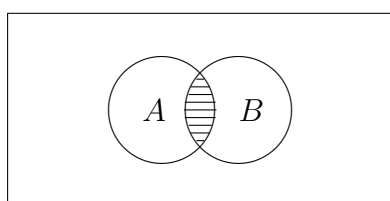
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



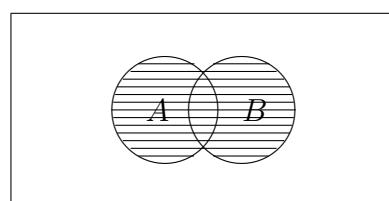
$$P(A^c) = 1 - P(A)$$



$$A \subset B; P(A) \leq P(B)$$



$$A \cap B, AB; A \text{ and } B$$



$$A \cup B; A \text{ or } B$$

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

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

Similarly, for four events A_1, \dots, A_4 we have

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

This type of formulas are called inclusion-exclusion formulas.

2.10 Example. 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	×	×	×	×	×	×

Clearly, the chance is $11/36$. Alternatively, let $A = \{\text{The first die shows 6}\}$ and $B = \{\text{The second die shows 6}\}$. Then

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

2.11 Example 2.2 continued. Note that we can solve the problem 2.2 by the rule 2.5(c) as follows:

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

Another interesting remark about Example 2.2: it is possible (at least theoretically) that the coin always lands on Tails and the stubborn flipper will never stop. So, there is one outcome that we have overlooked: TTT... (infinitely many T's). This outcome has probability zero, though, so it can be ignored without any harm.

2.12 Example. Let $P(A) = 0.5$, $P(B) = 0.3$ and $P(A \cap B) = 0.1$. Find $P(A^c \cap B^c)$ and $P(A^c \cup B^c)$. Answers: 0.3 and 0.9, respectively. (Draw a Venn's diagram to see this.)

2.13 Example. Let $P(A) = 0.8$, $P(B) = 0.7$. Find the maximum and minimum possible values for $P(A \cap B)$.

Solution: The maximum is 0.7, it occurs when $B \subset A$. To find the minimum, note that

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 1.5 - P(A \cup B)$$

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

2.14 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 a coincidence is actually quite high.

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

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

Hence, the chance of a coincidence is $1 - 0.2937 = 0.7063$. (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, maybe 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 $C_{30,2} = 435$. Each pair has a common birthday with probability $1/365$.

3 Conditional Probability and Independence

3.1 Example. A friend tosses a coin three times. You accidentally notice that the first time the coin shows Head. What is the chance that the friend observes 2 Heads?

Solution. In Example 2.1, we found all eight possible outcomes. Now, with the additional information at our disposal, we can exclude the outcomes starting with a T. Only four possible outcomes remain: HHH, HHT, HTH, HTT. Two of them contain exactly two Heads. So, the chance is $2/4=1/2$.

Note that here we have two events: $A = \{2 \text{ Heads are observed}\}$ and $B = \{\text{First toss is Heads}\}$. We know that $P(A) = 3/8$, see Example 2.1. 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 .

3.2 Conditional probability. The conditional probability of an event A , given an event B , is

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

3.3 Multiplication rule. The above formula can be rewritten as

$$P(A \cap B) = P(B) \cdot P(A/B)$$

Due to the symmetry, one can rewrite this as

$$P(A \cap B) = P(A) \cdot P(B/A)$$

3.4 Example. 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, $P(A) = 13/52 = 1/4$. 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 $P(B/A) = 12/51$. Hence,

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

3.5 Extended multiplication rule. If A_1, A_2, \dots, A_n are events, then

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2/A_1) \cdot P(A_3/A_1 \cap A_2) \cdots P(A_n/A_1 \cap \dots \cap A_{n-1})$$

3.6 Birthday problem revisited. The problem 2.14 now can be solved by using the extended multiplication rule:

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

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.

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

3.8 Law of total probability. Let $\{B_1, \dots, B_n\}$ be a partition of Ω , and A an event. Then

$$P(A) = P(B_1) \cdot P(A/B_1) + \cdots + P(B_n) \cdot P(A/B_n)$$

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

3.9 Example. Alex goes to school by bus or train, whichever comes first. He notice that the bus comes first with probability 30% and the train with probability 70%. When Alex takes train, he arrives late to school with probability 5%. When he takes bus, he is late to school with probability 20%. What is the probability that he is 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,

$$P(A) = P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2) = 0.3 \times 0.2 + 0.7 \times 0.05 = 0.095$$

3.10 Two-stage experiments. Amanda rolls a die and then flips a coin the number of times shown on the die. What is the chance she observes two Heads?

Solution: In the first stage, the die shows one of the six numbers $1, \dots, 6$. These are six events, which we denote by B_1, \dots, 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. Applying the law of total probability gives

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

Note that we used the formula (1.3) from Example 1.12 to find the probability of observing 2 Heads in n tosses for $n = 2, \dots, 6$.

3.11 Example. 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

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

Note that $P(A) = 1/6$, so that

$$P(A/B) = P(A)$$

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.

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

$$P(A/B) = P(A)$$

By using 3.2, we can rewrite this equation as

$$P(A \cap B) = P(A)P(B)$$

and also as

$$P(B/A) = P(B)$$

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

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

3.13 Example. Flip two coins. Let $A = \{\text{First coin shows Head}\}$ and $B = \{\text{Both coins show the same face}\}$. Are A and B independent?

Solution. One easily finds that $P(A) = 1/2$, $P(B) = 1/2$ and $P(A \cap B) = 1/4$. Then we check that $1/2 \times 1/2 = 1/4$. Yes, they are independent.

Note: Sometimes the independence is obvious, like in 3.11 (because there is no way the result of the first roll can affect the second). Sometimes the independence is harder to recognize, as it is in 3.13 above. One can explain the independence in 3.13 noting that the second coin may or may not show the same face as the first with probability $1/2$, no matter what face the first coin shows.

3.14* Remark. If two events A, B are independent, then A^c, B^c are also independent, i.e. $P(A^c \cap B^c) = P(A^c)P(B^c)$. Moreover, A and B^c are independent, i.e. $P(A \cap B^c) = P(A)P(B^c)$. Similarly, A^c and B are independent, i.e. $P(A^c \cap B) = P(A^c)P(B)$.

3.15 Independence of three events. Three events A, B, C are said to be mutually (or jointly) independent if

- (a) any two of them are independent in the sense of 3.12, and
- (b) the following holds:

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

Neither one of the conditions (a) and (b) alone is enough for joint independence. One needs to check both (a) and (b) to verify the joint independence of A, B, C .

Example 3.13 continued. Let $C = \{\text{Second coin shows Head}\}$. Are A, B, C jointly independent?

Solution: We have seen in 3.13 that A and B are independent. Similarly, B and C are independent. Obviously, A and C are independent. So, the requirement (a) in 3.15 holds. On the other hand, $P(A \cap B \cap C) = 1/4$, and $1/2 \times 1/2 \times 1/2 \neq 1/4$, so the requirement (b) in 3.15 fails. The events A, B, C are not jointly independent.

3.16 Example. Flip three coins. Let $A = \{\text{First coin shows Heads}\}$, $B = \{\text{Second coin shows Tails}\}$, $C = \{\text{At least two coins show Heads}\}$. Are A, B, C independent?

Solution: One easily finds that $P(A) = P(B) = P(C) = 1/2$ (for $P(C)$, recall Example 2.1). Also, $P(A \cap B \cap C) = 1/8$, so the requirement (b) in 3.15 holds. But A and C are dependent, so the requirement (a) fails.

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

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

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

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

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

So, the entire system will function with probability 99.9%. Note a very high reliability!

An additional note: it is more difficult to find the probability that 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 work. Therefore, the probability that two components fail is

$$\begin{aligned} P(\text{two fail}) &= P(1, 2 \text{ fail}) + P(1, 3 \text{ fail}) + P(2, 3 \text{ fail}) \\ &= 0.1 \cdot 0.2 \cdot 0.95 + 0.1 \cdot 0.8 \cdot 0.05 + 0.9 \cdot 0.2 \cdot 0.05 = 0.032 \end{aligned}$$

3.19* Remark. This generalizes Remark 3.14. If A_1, \dots, A_n are independent, one can replace any number of these events by their complements (e.g., A_1 by A_1^c , etc.) and the new collection of events will be also independent.

3.20 Bayes formula. Recall the law of total probability in 3.8 and suppose we need to compute $P(B_i/A)$ for some $i = 1, \dots, n$. By using the formulas in 3.2-3.3 we get

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

Now we replace the denominator $P(A)$ by its expansion given by the law of total probability and obtain

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

This is called *Bayes formula*. Note that the numerator here is one of the terms that appear in the denominator

3.21 Interpretation. We first recall how we interpreted the law of total probability in 3.8. An event A can occur under different conditions (hypotheses) B_1, \dots, 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 3.8.

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 took place when the event A occurred. This is what the Bayes formula does.

3.22 Example. Three factories, F_1 , F_2 , and F_3 , produce computer chips. The factory F_1 produces 50% of all the chips on 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 on the market. After a 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 $P(A/B_1) = 0.01$, $P(A/B_2) = 0.02$, $P(A/B_3) = 0.03$. Now, by the Bayes formula, we have

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

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

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

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

3.23 Example. Statistics show that 3% of men smoke but only 1% of women do. During a non-smoking flight on an airplane, 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, $P(M) = 0.4$ and $P(F) = 0.6$. Let S denote the event that a passenger is a smoker. Then $P(S/M) = 0.03$ and $P(S/W) = 0.01$. By the Bayes formula we have

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

3.24 Example. 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 $P(B_1) = P(B_2) = P(B_3) = 1/3$. Now it is simple to find $P(A/B_1) = 1/2$, $P(A/B_2) = 1/2$ and $P(A/B_3) = 3/8$. Then

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

3.25 A 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 would assume that it is exactly the same as in Example 3.24, i.e. $4/11$. Why? Remember, boys and girls appear equally likely. In 3.24, 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, it does make a difference if you 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 $P(A/B_1) = 1/2$, $P(A/B_2) = 1/4$ and $P(A/B_3) = 1/8$. Then

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

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...

4 Discrete Random Variables

4.1 Counting Heads. Consider again Example 2.1 where a coin is tossed three times. 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 four distinct values, as opposed to 8 distinct outcomes in the experiment. The number of values is less 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).

4.2 Success/Failure Trials. Let us generalize Example 4.1. 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 format of Example 2.1:

	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 Example 4.1 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 $P(\text{HHT}) = ppq = p^2q$, $P(\text{THT}) = qpq = pq^2$, etc. This way we can find the probabilities that X takes values 0, 1, 2, 3. We summarize them in the table

below:

values	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 $q^3 + 3pq^2 + 3p^2q + p^3 = (q + p)^3 = 1^3 = 1$, so the probabilities sum up to one, as they should.

4.3 Bernoulli trials. Any simple trial with only two possible outcomes is called a *Bernoulli 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 (as explained in 1.12 and 4.2). 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 $p^k q^{n-k}$.

4.4 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 $0, 1, \dots, n$. 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 $n = 3$, then $X(\text{SSS}) = 3$, $X(\text{FSF}) = 1$, $X(\text{FFF}) = 0$, etc.

4.5 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 in Example 4.1).

4.6 Probabilities of a binomial random variable. Back to Example 4.4, we call X there 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 $p^k q^{n-k}$, cf. 4.3. There are $C_{n,k}$ of such strings, cf. 1.12. Hence,

$$P(X = k) = \binom{n}{k} p^k q^{n-k} \quad 0 \leq k \leq n \quad (4.1)$$

This formula is good for all $k = 0, 1, \dots, n$. 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 1.9. This is why X is called the *binomial* random variable.

Note that the probability $P(X = k)$ in (4.1) 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 denote $X = \text{binomial}(n, p)$ or shortly $X = b(n, p)$.

4.7 Geometric random variable. Generalizing the stubborn flipper example 2.2, consider independent Bernoulli trials that are performed until a success occurs. The outcomes in this experiment are

$$S, \text{ FS}, \text{ FFS}, \text{ FFFS}, \dots, \underbrace{\text{F} \dots \text{F}}_{n-1} S, \dots$$

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

Hence,

$$P(X = n) = pq^{n-1} \quad n \geq 1 \quad (4.2)$$

Note that the sum of these probabilities is

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

so the probabilities sum to one, as they should. The probability $P(X = n)$ in (4.2) involves the only parameter p . We denote X by $X = \text{geometric}(p)$ or shortly $X = g(p)$.

4.8* Problem. For a geometric random variable X , compute $P(X > k)$.

Solution. We have

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

4.9 Discrete random variables. Generalizing 4.4 and 4.7, we say that a random variable X may take some values x_1, x_2, \dots with corresponding probabilities p_1, p_2, \dots . (The list of values may be finite as in 4.4 or infinite as in 4.7.) It is sometimes convenient to put them all in a table:

X	x_1	x_2	x_3	\cdots
P	p_1	p_2	p_3	\cdots

An important requirement here is that the probabilities sum to one:

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

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

4.10 Probability density function. The function that assigns the probability p_k to the value x_k is called *probability density function* or just *probability function*. For example, the formulas (4.1) and (4.2) give probability density functions for binomial and geometric random variables, respectively.

4.11 Uniform (discrete) random variable. Let $n \geq 1$. A very simple random variable takes values $1, 2, \dots, n$ with equal probability, $1/n$. Its probability density function is

$$P(X = k) = 1/n \quad 1 \leq k \leq n$$

We call X a *uniform* (discrete) random variable. An example of such a variable is the number shown when a die is rolled, cf. 1.15, there we had $n = 6$ and $P(X = k) = 1/6$.

4.12 Remark. It is important to show the range of the variable in the formula for the density function, such as $0 \leq k \leq n$ in (4.1), $n \geq 1$ in (4.2), etc. It is assumed that the probability density is zero for all other values of k . For example, if $X = b(n, p)$, then $P(X = -1) = 0$, $P(X = n + 5) = 0$, $P(X = 1.4) = 0$, etc.

4.13 Special binomials: n large, p small. In some practical situations we have a binomial random variable with a very large n and a very small p . Here are some 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 each individual customer calls now is usually very small.

(b) The number of customers arriving at a supermarket or a car shop today. Again, we have a large number of potential customers and a small probability that every particular customer will arrive today.

(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 test on 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.

4.14 Poisson approximation to binomial. Let $X = b(n, p)$ be a binomial random variable with a small p and large n , as described in 4.13. The exact formula (4.1) for $P(X = k)$ 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 $P(X = k)$ by a formula that only involves “reasonable” numbers. We assume that n is huge, p is tiny, and k is reasonably small: $k = 0, 1, 2, \dots$

First, we note that

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

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

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

There is a useful formula in calculus:

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

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

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

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 $X = b(n, p)$ is a binomial random variable with large n and small p , one can compute the probability $P(X = k)$ for small k by

$$P(X = k) \approx \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{with} \quad \lambda = np \quad (4.3)$$

This formula is called *Poisson approximation to binomials*.

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

4.15 Poisson random variable. Motivated by (4.3), we introduce a Poisson random variable X that takes values $0, 1, 2, \dots$ with probabilities

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad k \geq 0$$

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 (4.3) works very well in practical applications.

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

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

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

with

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

Hence,

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

All these numbers are easily computable with a calculator.

Note that the Poisson approximation (4.3) only requires the average number of successes $\lambda = np$, the values of n and p separately are not involved and need not be known.

4.17 Example: 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

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

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

5 Continuous Random Variables

Here we consider random variables that take any values on the real line or in an interval.

5.1 Example: Archery. Continuing Example 2.3, 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). Thus, $0 \leq X \leq 30$.

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

5.3 Remarks. 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, in the way we did for discrete random variables in Section 4. This is a new type of random variables, that we will call continuous. Another distinction: for any possible value x , the probability $P(X = x)$ is zero, in both Examples 5.1 and 5.2. (In 5.1, the value $X = x$ is taken on a circle of radius x , which is just a curve on the target surface, and according to 2.3 the probability of the event $X = x$ is zero, because the area of any curve is zero.) Since we have $P(X = x) = 0$ for every individual number 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 $P(a < X < b)$ for arbitrary $a < b$. Such probabilities are usually positive, and they describe the random variable X in a meaningful and complete way.

For a given random variable X , the probability $P(a < X < b)$ 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:

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

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

5.4 Distribution function. Given a random variable X , the *distribution function* $F_X(x)$ is defined by

$$F_X(x) = P(X \leq x)$$

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

5.5 Archery example, continued. Compute the distribution function for the random variable X in Example 5.1.

Solution. The random variable X takes values $0 \leq X \leq 30$. Hence, if $x < 0$, then $X \leq x$ is an impossible event and $P(X \leq x) = 0$. If $x > 30$, then $X \leq x$ is always true (it is a certain event), hence $P(X \leq x) = 1$. If $0 \leq x \leq 30$, then the event $\{X \leq x\}$ occurs if the hit point lies in the inner disk of radius x , then by the rule of Example 2.3, we have

$$P(X \leq x) = \frac{\pi x^2}{\pi 30^2} = \frac{x^2}{900}$$

Finally, we have

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x^2/900 & \text{if } 0 \leq x \leq 30 \\ 1 & \text{if } x > 30 \end{cases} \quad (5.1)$$

5.6 Properties of the distribution function. It is clear that $F_X(x)$ always has the following properties:

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

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

5.7 Remark. 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.

5.8 Computation of probabilities. For any interval (a, b) we have

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

Also,

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

5.9 Archery example, continued. By the above formulas, we can compute

$$P(1 < X < 3) = F(3) - F(1) = (9 - 1)/900 = 8/900$$

$$P(X > 20) = 1 - F(20) = 1 - 20^2/900 = 1 - 4/9 = 5/9$$

$$P(10 < X < 40) = F(40) - F(10) = 1 - 10^2/900 = 1 - 1/9 = 8/9$$

Note that $F(40) = 1$, not $40^2/900$, see (5.1).

5.10 Continuous random variables. A random variable X is said to be *continuous* if for any real number x we have $P(X = x) = 0$. In this case the distribution function $F_X(x)$ is continuous.

Note: if X is continuous, then in all the equations of 5.8 we have $P(X = a) = P(X = b) = 0$. Thus, it does not matter whether one puts strict inequalities ($<$) or non-strict ones (\leq) in those equations. In particular, we have

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

We will only use distribution function for continuous random variables (except for one example, 5.22, which is provided “for fun” rather than for serious business). So, we will be rather casual in some formulas like $P(a < X < b)$ or $P(a \leq X \leq b)$ 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.

5.11 Probability density function. The expression $F(b) - F(a)$ in 5.8 and 5.10 reminds us of the fundamental theorem of calculus:

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

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

$$P(a < X < b) = \int_a^b f(x) dx$$

Also,

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

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

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

That is, F is an antiderivative of f .

5.12 Properties of the density function. The properties of $F(x)$ stated in 5.6 can be easily rewritten in terms of $f(x)$:

- $f(x) \geq 0$.
- If $X \leq X_{\max}$, then $f(x) = 0$ for all $x \geq X_{\max}$. Also, if $X \geq X_{\min}$, then $f(x) = 0$ for all $x \leq X_{\min}$.
- The total integral of $f(x)$ equals one:

$$\int_{-\infty}^{\infty} f(x) dx = 1 \tag{5.2}$$

The last property is called the normalization rule.

5.13 Min/max rules. It is important to emphasize that if X has a maximum value, X_{\max} , then $F(x) = 1$ and $f(x) = 0$ for all $x \geq X_{\max}$. Also, 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 $f(x)$ are trivial outside the interval $[X_{\min}, X_{\max}]$, on which the random variable takes all its values. It is therefore common to only give the values of $F(x)$ and/or $f(x)$ on the essential interval $[X_{\min}, X_{\max}]$ leaving out their values beyond that interval (to save space).

5.14 Archery example, continued. In particular, in the archery example 5.5 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 density function in that problem is $f(x) = F'(x) = x/450$ for $0 < x < 30$. Again, we understand 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!

5.15 Examples. Which of the following functions are distribution functions? For those who 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 satisfy the *power law*.

5.16 Example. 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 (5.2):

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

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

Example 5.16 continued. Compute $P(X > 2)$, $P(2 < X < 3)$, $P(X > 0)$, $P(X \geq 4)$, $P(X = 2)$ and $P(X > 2/X < 3)$.

Solution: First, 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,

$$P(X > 2) = 1 - F(2) = 1 - 3/15 = 4/5$$

$$P(2 < X < 3) = F(3) - F(2) = 8/15 - 3/15 = 1/3$$

$$P(X > 0) = 1 - F(0) = 1 - 0 = 1$$

$$P(X \geq 4) = 1 - F(4) = 1 - 15/15 = 0$$

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

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

5.17 Probability density function and area. The probability $P(a < X < b) = \int_a^b f(x) dx$ equals the area under the graph of the density function. Since the total probability of all possible values equals one, the area under the entire graph of the density function $y = f(x)$ equals one (this is exactly the normalization rule (5.2)).

5.18 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

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

So,

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

for small d . Hence, the density equals the ratio of the probability that X takes value in a small interval over the length of that interval. In other words, the density is the “probability per unit length”. (Compare this to the classical mechanical interpretation of the mass density, which is nothing but the mass per unit length!)

Note: the higher $f(x)$ on some interval, the more likely the values in that interval are taken by the random variable. In the archery example 5.1, the

density $f(x) = x/450$ (see 5.14) 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.

5.19 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. $f(x) = c$, some constant. To find c , we can use the normalization rule (5.2) that gives $\int_a^b f(x) dx = 1$ and work as in 5.16, then we get $c = 1/(b - a)$. Hence,

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

for $a < x < b$ (and these function take their trivial values outside the interval (a, b) , see 5.13). 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*.

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) . One can also think of a hit point when one shoots at the interval (a, b) randomly. This is similar to the archery example 2.3. According to the same principle as in 2.3, the probability to hit any small interval (u, v) inside (a, b) is proportional to its length, i.e. $P(u < X < v) = (v - u)/(b - a)$. This completely agrees with the above formulas for $f(x)$ and $F(x)$.

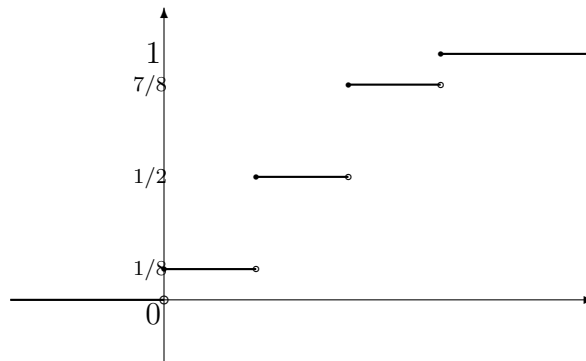
5.20 U(0,1). One uniform random variable, $X = U(0, 1)$, plays an exclusive role in probability theory. This will be clear in 5.21 and later on. We only note here that its density and distribution functions are given by very simple formulas:

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

5.21 Random number generators (RNG). The uniform random variable $U(0, 1)$ plays a special role in computer programming. Many computer software have a built-in random number generator, that upon request produces a number x between 0 and 1, which is supposed to be completely random. Each call to the RNG returns a new random number between 0 and 1. Hence,

the computer RNG simulates the uniform random variable $U(0, 1)$. The computer code that calls an RNG, usually looks like $X = RND$, $X = RAND$, $X = RANF$, etc.

5.22 Distribution function of discrete random variables. Let X be the number of Heads in three tosses of a coin. This is a discrete random variable completely described by Example 4.1. Its distribution function $F(x)$ can be found now by the rule 5.4. For example, $F(-1) = 0$ (the event $X \leq -1$ is clearly impossible), $F(0) = 1/8$ (the event $X \leq 0$ occurs only when $X = 0$), $F(0.6) = 1/8$ (again, the event $X \leq 0.6$ occurs only if $X = 0$), $F(1) = 1/2$ (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 in 5.22

Note: The graph is a “ladder” with “steps” going up. 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 $P(X = x)$. These are general rules for distribution functions of discrete random variables. Note that such variables have no density functions in the sense of 5.11 (because $F(x)$ cannot be differentiated at discontinuity points).

6 Exponential Random Variables

6.1 Time to failure. In Example 5.2, we discussed the lifetime of a TV or a car. Let us change it and denote by T the time to failure of a TV or a car (a failure requires service, but is not necessarily the end of life of the product). In many cases, T has the following distribution function, approximately:

$$F_T(x) = 1 - e^{-\lambda x} \quad x > 0$$

Note that T only takes positive values, so $F_T(x) = 0$ for $x < 0$. Here $\lambda > 0$ is a parameter (a technical characteristic of the product, the meaning of λ will be discussed below).

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

$$F_X(x) = 1 - e^{-\lambda x} \quad x > 0$$

By differentiating, we obtain the density function

$$f(x) = \lambda e^{-\lambda x} \quad x > 0$$

The constant $\lambda > 0$ is the parameter of the exponential random variable. We denote this variable by $X = \text{exponential}(\lambda)$.

6.3 Memoryless property (no aging). Let X be an exponential random variable. Consider two events, $A = \{X > a\}$ and $B = \{X > a + x\}$, for some $a > 0$ and $x > 0$. Note: A means that the time to failure exceeds a (the product functions fine a units of time) and B means that the product functions without failures $a + x$ units of time. Let us compute the conditional probability $P(B/A)$. To state the question differently: given that the product has worked without failures a units of time, what is the probability that it will work x more units of time without failures?

Solution: Recall that $P(B/A) = P(B \cap A)/P(A)$. Note that the event $B = \{X > a + x\}$ implies $A = \{X > a\}$, i.e. $B \subset A$, so $B \cap A = B$. Hence,

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

Compare this to the probability $P(X > x) = 1 - F(x) = e^{-\lambda x}$. They coincide!

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

6.4 Remark. 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.

6.5 Radioactive decay. Real cars and TV’s certainly deteriorate in time (get older and have “memory” of past failures and repairs), so exponential random variable can only describe their times to failure approximately. There is, however, a natural phenomenon that is characterized by an ideal memoryless/no aging property. 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”. Then the decay time (or the lifetime of the atom) is an exponential random variable.

The process of decay can be illustrated as follows. Suppose a piece of radioactive material contains N radioactive 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 $P(X < t) = 1 - e^{-\lambda t}$, so it will survive with probability $p = 1 - P(X < t) = e^{-\lambda t}$. Hence, approximately $(1 - p)N$ atoms disintegrate during the first interval, and pN atoms survive. During the next interval of time, each 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, a 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.

6.6 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 $e^{-\lambda t_{1/2}} = 1/2$ or

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

This equation relates λ and $t_{1/2}$. The value of $t_{1/2}$ is a standard technical characteristic of radioactive atoms, it is given in reference books. Given $t_{1/2}$, one can find $\lambda = \ln 2/t_{1/2}$.

Note that $P(X > t_{1/2}) = 1/2$, $P(X > 2t_{1/2}) = 1/4$, $P(X > 3t_{1/2}) = 1/8$, etc., as we saw in 6.5.

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

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

6.8 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 we have the equation $F_X(m) = 1/2$. Note that $P(X \leq m) = P(X > m) = 1/2$. In this sense, m exactly divides the probability distribution of X in half.

Note: The half-life $t_{1/2}$ is the median of any exponential random variable.

6.9 Example. New York Times has reported in 1999 that the median of the prices of houses in the South of the United States is \$135,000. What does this mean? Half of the houses are sold below \$135,000 and half of the houses are sold for more than \$135,000.

6.10* Percentiles. 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

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

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).

6.11 Failure rate. Back to exponential random variable X . The parameter λ is often referred to as *failure rate* (or decay rate, or death rate, depending on the situation). To see why, consider the number of atoms disintegrating per a small interval of time Δt . The probability of disintegration is

$$P(X < \Delta t) = 1 - e^{-\lambda \Delta t} \approx \lambda \Delta t$$

So, approximately $\lambda \Delta t N$ atoms disintegrate during a small interval Δt . This explains why λ is interpreted as the decay rate.

6.12 Remark. It is not a coincidence that we denote by the same symbol λ the parameters of Poisson and exponential random variables. We will see later, in Section 17, a process where both random variables are involved and their parameters coincide.

7 Functions of Random Variables

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

7.1 $Y=g(X)$. 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.

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

Solution. We have

$$F_Y(y) = P(Y \leq y) = P(12X - 6 \leq y) = P(X \leq (y+6)/12) = F_X((y+6)/12)$$

Since $F_X(x) = x$, see 5.20, we have $F_Y(y) = (y + 6)/12$. By differentiating, one gets $f_Y(y) = 1/12$. Of course, one needs to specify where these formulas for F_Y and f_Y are valid. One simply needs to find the values the variable Y takes. Since $0 < X < 1$, we have $0 < 12X < 12$ and $-6 < Y < 6$. Finally, one gets $F_Y(y) = (y + 6)/12$ and $f_Y(y) = 1/12$ for $-6 < y < 6$. It is advisable to find the range (all possible values) of the random variable Y first, this may simplify calculations.

7.3 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 :

- Find the range (interval of possible values) for the variable Y ;
- Start with $F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$, then solve the inequality $g(X) \leq y$ for X (this is the most tricky and confusing part!);
- Express the resulting probability in terms of the distribution function F_X and use the given (known) function F_X to obtain F_Y ;
- Differentiate F_Y to get f_Y .

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

Solution. Since $-1 < X < 1$, then $0 < X + 1 < 2$ and so $1/2 < 1/(X + 1) < \infty$. Now compute $F_Y(y)$ for $1/2 < y < \infty$:

$$F_Y(y) = P(Y \leq y) = P(1/(X + 1) \leq y) = P(X + 1 \geq 1/y)$$

Note: the inequality $1/(X + 1) \leq y$ is transformed into $X + 1 \geq 1/y$ because $X + 1 > 0$ and $y > 0$ (otherwise the inequality might have been reversed –

dividing or multiplying both sides by a negative number reverses the inequality). Now recall that $F(x) = (x+1)/2$ by 5.19 and complete the calculation:

$$F_Y(y) = P(X \geq y^{-1} - 1) = 1 - F(y^{-1} - 1) = 1 - (y^{-1} - 1 + 1)/2 = 1 - (2y)^{-1}$$

for $1/2 < y < \infty$. By differentiating, $f_Y(y) = 1/(2y^2)$. If one is not sure about the answer (the formulas look too weird), just verify that in the given interval the function F_Y increases from 0 to 1. In our case the interval is $1/2 < y < \infty$, so we verify that $F_Y(1/2) = 1 - 1^{-1} = 0$, indeed, and $F_Y(\infty) = 1 - 0 = 1$, all right.

7.5 Example. Let $X = U(0, 1)$ and $Y = -\lambda^{-1} \ln(1 - X)$ for some constant $\lambda > 0$.

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

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(-\lambda^{-1} \ln(1 - X) \leq y) = P(\ln(1 - X) \geq -\lambda y) \\ &= P(1 - X \geq e^{-\lambda y}) = P(X \leq 1 - e^{-\lambda y}) = F_X(1 - e^{-\lambda y}) = 1 - e^{-\lambda y} \end{aligned}$$

Amazing, this is the distribution function from 6.2. So, Y is an exponential random variable. Now, by differentiating, $f_Y(y) = \lambda e^{-\lambda y}$.

7.6 Generating exponential random variables. The last example shows how to generate an exponential random variable by a computer. Simply call an RNG (5.21) that returns a value X in $(0, 1)$, then compute $Y = -\ln(1 - X)/\lambda$. Generally speaking, one can generate any random variable by using the RNG described in 5.21, see also below.

7.7 Example. Let $X = \text{exponential}(\lambda)$ and $Y = \sqrt{X}$.

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

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

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

7.8 Example. Let $X = U(-1, 1)$ and $Y = X^2$.

Solution. There is a catch here! First, $-1 < X < 1$, then $0 < X^2 < 1$, hence $0 < Y < 1$. Now, for all $0 < y < 1$ we have

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y)$$

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

$$F_Y(y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) = (1 + \sqrt{y})/2 - (1 - \sqrt{y})/2 = \sqrt{y}$$

and $f_Y(y) = F'_Y(y) = 1/(2\sqrt{y})$.

7.9 Linear transformation. Let X be an arbitrary random variable with distribution function F_X and density f_X . Consider $Y = a + bX$, where a and $b > 0$ are constants.

Solution. We have

$$F_Y(y) = P(Y \leq y) = P(a + bX \leq y) = P\left(X \leq \frac{y - a}{b}\right) = F_X\left(\frac{y - a}{b}\right)$$

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

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

7.10 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.

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

Solution. First, $0 \leq F_X(X) \leq 1$, by the first property in 5.6. Now we have for $0 \leq y \leq 1$

$$F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y)$$

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

$$F_Y(y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y$$

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

7.12* Generating continuous random variables by computer. Let X be an arbitrary continuous random variable. According to 7.11, the variable $Y = F_X(X)$ is $U(0, 1)$. In other words, if $Y = U(0, 1)$, 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 (5.21) 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 of F_X is simple, an exact solution can be found, often by hands. If F_X is complicated, one can find an approximate solution by using a special computer program.

8 Normal Random Variables

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

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

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

Note that $f_Z(x)$ is defined and 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.

The graph of $f_Z(x)$ is a bell-shaped curve, symmetric about the y-axis. This curve is called a gaussian curve. Its maximum is at $x = 0$, then it decreases on both sides of its top point. Actually, it decreases very fast. One can easily check that $f_Z(0) \approx 0.399$, $f_Z(1) \approx 0.242$, $f_Z(2) \approx 0.054$, $f_Z(3) \approx 0.0044$, $f_Z(4) \approx 0.00013$, $f_Z(5) \approx 10^{-6}$, etc. For larger x , the function $e^{-x^2/2}$ approaches zero dramatically rapidly, so for all practical purposes $f_Z(x) = 0$ for something like $|x| > 5$.

8.2 The Φ function. The distribution function of Z is denoted by $\Phi(x)$. According to 5.11,

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

Unfortunately, there is no formula for $\Phi(x)$ in terms of elementary functions. The above expression is the best one can write for $\Phi(x)$.

Since there is no convenient formula for $\Phi(x)$, the values of this function cannot be easily computed. We will use Table on page 266 of the textbook to find the values of the function $\Phi(x)$.

8.3 Table of $\Phi(x)$. Now we learn how to use Table on p. 266. For $0 \leq x \leq 3.09$ this is obvious. For negative x , specifically for $-3.09 \leq x \leq 0$, we can use the symmetry rule:

$$\Phi(-x) = 1 - \Phi(x) \quad x > 0$$

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

Finally, for all $x > 3.09$ 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 quite an accurate assumption.

8.4 Examples. Compute $\Phi(1)$, $\Phi(2.36)$, $\Phi(-1.25)$, $\Phi(3.7)$.

Answers: $\Phi(1) = 0.8413$, $\Phi(2.36) = 0.9909$, $\Phi(-1.25) = 1 - \Phi(1.25) = 1 - 0.8943 = 0.1057$, $\Phi(3.7) = 1$.

8.5 Examples. Compute $P(Z < 2.87)$, $P(Z > 0.76)$, $P(Z < -0.76)$, $P(Z > -2)$, $P(-0.6 < Z < 1.3)$, $P(|Z| < 2)$, $P(|Z| < 3)$, $P(|Z| > 4)$.

Solution. We have

$$P(Z < 2.87) = \Phi(2.87) = 0.9979$$

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

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

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

$$P(-0.6 < Z < 1.3) = \Phi(1.3) - \Phi(-0.6) = \Phi(1.3) - 1 + \Phi(0.6) = 0.9032 - 1 + 0.7257 = 0.6289$$

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

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

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

8.6 Normal random variables. Now we are ready to introduce a (general) normal random variable. It has two parameters: μ and $\sigma > 0$. It is 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. The normal random variable Y is denoted by $N(\mu, \sigma^2)$, i.e. we write $Y = N(\mu, \sigma^2)$. Note also that we have $Z = N(0, 1)$ is this notation.

8.7 Density and distribution function of $N(\mu, \sigma^2)$. By 7.9, the density of $Y = N(\mu, \sigma^2)$ is

$$f_Y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$$

and the distribution function is

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

8.8 Remarks. Note that the density function $f_Y(x)$ is positive for all x , has a peak at $x = \mu$ and goes down on both sides of the peak. It is a bell-shaped curve, just like $f_Z(x)$ in 8.1, but shifted (centered) to the point μ . The other parameter, σ affects the shape of the curve: for smaller σ , the peak is high but narrow, for larger σ the peak is low but wide. Remember that, in any case, the total area under the graph of $f_Y(x)$ is the same, it equals one by the normalization rule (5.2).

8.9 Examples. Let $Y = N(5, 4)$. Compute $P(Y < 7)$ and $P(3 < Y < 6)$.

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

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

and

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

8.10 Manipulations with a normal random variable. Let $Y = N(\mu, \sigma^2)$ 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

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

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

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

The moral is: when you rescale a normal random variable, you get another normal random variable.

Note: if $Y = N(\mu, \sigma^2)$, then the variable $W = -Y$ is normal and $W = N(-\mu, \sigma^2)$ (this is a case of the above rule with $a = 0$ and $b = -1$).

8.11 Rule of three sigmas. We have seen in 8.5 that $P(|Z| < 3) = 99.72\%$. Now, for any normal random variable $Y = N(\mu, \sigma^2)$ we have

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

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 ”.

8.12 Standard normal variable squared. Find the distribution and density function of $W = Z^2$.

Solution. This is similar to 7.8. Obviously, $W > 0$. Now we have, for all $x > 0$,

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

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.

8.13 Remark. The importance of normal random variables will be demonstrated later, in Section 15. Right now we can just say that many random variables in practical applications are approximately normal. Consider the height (or weight) of students in a big class. Most of them will have close to

the average height (weight) with a few exceptions that are well over or well under the average. Plotting the density function will give something close to a bell-shaped curve. Another example: errors in many experimental data have approximately normal distribution, again we will see that Section 15.

8.14* Error function. 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 $\Phi(x)$ in 8.2 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.

8.15* Approximations to $\Phi(x)$. In some very precise calculations, one needs the values of $\Phi(x)$ for $x > 3$. The following formula gives a very good approximation:

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

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

8.16* Remark. A law in physics shows that the normal distribution is unique. A gas (the air in the room, for example) consists of billions and billions of molecules (something like 10^{25} or 10^{30} molecules in the room). They all move chaotically at various speed and in various directions. Pick one molecule at random, then its velocity vector $\mathbf{v} = (v_x, v_y, v_z)$ is practically a random vector. The components v_x, v_y, v_z are random variables. They have the same distribution, since there is apparently no difference between the x- y- and z-direction in the molecular chaos. Moreover, if one rotates the coordinate frame, then the new components v_x, v_y, v_z , even though measured differently, will have the same distribution. Such distributions are said to be *spherically symmetric*, they are independent of the direction of the x, y, z axes. Another look at this model suggests that v_x, v_y, v_z are independent –

this seems intuitively quite reasonable. Let us adopt these requirements: a spherical symmetry and the independence of v_x, v_y, v_z . It turns out, quite surprisingly, that the only distribution that satisfy these two requirements is normal! This is called the *Maxwell law* in physics. The following formula describes it:

$$f(x, y, z) = \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\frac{x^2+y^2+z^2}{2\sigma^2}}$$

This formula will make more sense in 9.13.

8.17 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 $N(0, 1)$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	$\Phi(x)$	$-\infty < x < \infty$
normal $N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$	$-\infty < x < \infty$

9 Joint Distributions

Here we consider experiments that involve two (or more) random variables at a time.

9.1 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 $P(X = Y)$. Find $P(X < Y)$.

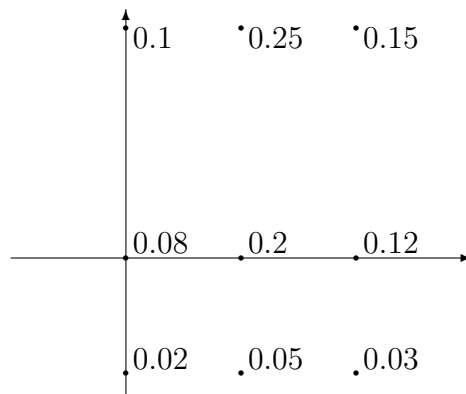
Solution. Possible pairs of values are $(0, -1), (0, 0), (0, 2), (1, -1), \dots, (2, 2)$. The probabilities are computed by the multiplication rule

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

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 corresponding points on the xy plane and write the probabilities next to the points. 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 $y = x$, in this case the points $(0, 0)$ and $(2, 2)$. Hence, $P(X = Y) = 0.08 + 0.15 = 0.23$.

The event $\{X < Y\}$ contains all the points above the diagonal $y = x$, i.e. the points $(0, 2)$ and $(1, 2)$. Hence, $P(X < Y) = 0.25 + 0.15 = 0.4$.

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 $1/36$.

9.2 Discrete pairs 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.

9.3 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

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

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

Example 9.1 continued. Here are some values of the joint distribution function in Example 9.1: $F(1.1, 0.8) = 0.35$ (the quadrant to the left and below the point $(1.1, 0.8)$ covers four pairs of (X, Y) , with the total probability 0.35), also $F(5, -0.4) = 0.1$, $F(4, 7) = 1$, $F(-2, 8) = 0$, etc. We will not attempt to describe this function completely, it is not of much use in this

example.

9.4 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:

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

We know that $F_X = x$ for $0 < x < 1$ and $F_Y(y) = y$ for $0 < y < 1$, see 5.20. 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 X, Y .

9.5 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 IV, should not worry much. We will use elements of multivariate calculus (partial derivative and double integrals) only of the simplest forms, in 9.7 we give precise instructions.

Example 9.4 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).

9.6 Computation of probabilities. Generalizing the rules 5.11 (that the probability is the integral of the density function) we now have

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

This is a double integral of the function f over the region R .

9.7 Rule for constant density functions. Let the joint density function be constant: $f_{X,Y}(x, y) = c$ over the region R (as in Example 9.4, where $f(x, y) = 1$ over the unit square). Then the double integral in 9.6 simply equals c times the area of R . Hence,

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

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.

Example 9.4 continued.

(a) Find the probability $P(X + Y < 1)$.

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

(b) Find $P(X^2 + Y^2 < 1)$.

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, $P(X^2 + Y^2 < 1) = \pi/4$.

(c) Find $P(|X - Y| < 0.1)$.

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 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 $P(|X - Y| < 0.1) = 0.19$.

9.4 b



9.4 c



9.9



9.8 Remark. It is interesting that the problem (b) above suggests a method of determining the number π to any precision (at least theoretically). Simply generate pairs of uniform random numbers by an RNG (see 5.21), 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 Section 15 we will learn how many pairs of random numbers one needs to generate to obtain k correct digits of the number $\pi/4$.

9.9 Example. Let X, Y have the joint density function $f(x, y) = 2$ for $0 < y < x < 1$. Find $P(X - Y > 0.4)$.

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, $P(X - Y > 0.4) = 2 \times 0.18 = 0.36$.

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

9.10 A weird example. 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 individual point (x, y) we have $P(X = x, Y = y) = 0$. On the other hand, it does not have a joint density function. We will not study such “weird” examples in detail.

9.11 Multiplication rules for independent r.v.. Let X and Y be two independent random variables. Then we have simple multiplication rules:

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

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 Example 9.4, the second follows by differentiation.

9.12 Three and more random variables. Let X_1, \dots, X_n be n random variables. One can also call (X_1, \dots, 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, \dots, X_n . The multiplication rules of 9.11 work for any number of independent random variables.

9.13* Maxwell's distribution. Let X, Y, Z be three independent normal random variables $N(0, \sigma^2)$. Then, by the multiplication rule and the formula of 8.7, their joint density function is

$$f_{X,Y,Z}(x, y, z) = \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\frac{x^2+y^2+z^2}{2\sigma^2}}$$

This is exactly Maxwell's law given in 8.16!

9.14 Min/max of two random variables. Let X and Y be two independent random variables, and F_X, 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,

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

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

$$F_W(x) = 1 - P(W > x) = 1 - P(X > x)P(Y > x) = 1 - (1 - F_X(x))(1 - F_Y(x))$$

In particular, if X and Y have the same distribution function F , then

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

9.15 Independent identically distributed (i.i.d.) random variables.

Let X_1, \dots, 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 frequent in probability theory.

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

Solution. Very much like in 9.14, 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, \dots, 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 using the chain rule:

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

and similarly

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

9.17 Example. Let X_1, \dots, X_n be i.i.d. random variables uniform on $(0, 1)$. Find the distribution and density functions of $V = \max\{X_1, \dots, X_n\}$ and $W = \min\{X_1, \dots, X_n\}$.

Solution. The common distribution function of X_1, \dots, 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 $0 < x < 1$. Graph the density functions $f_V(x)$ and $f_W(x)$ and you 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, \dots, X_n , most likely takes values close to 1. On the contrary, W , the minimum of X_1, \dots, X_n , most likely takes values close to 0.

9.18 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 , $1 \leq i \leq n$, the lifetime (time to failure) of the i th component. Then X_1, \dots, 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 $T_{\text{fr}} = \min\{X_1, \dots, X_n\}$. The other uses connection in parallel. That type of system is “robust”, it works if at least one component is functioning, so that $T_{\text{rb}} = \max\{X_1, \dots, X_n\}$.

According to 9.17, we have

$$F_{T_{\text{fr}}}(x) = 1 - (1 - F(x))^n \quad \text{and} \quad F_{T_{\text{rb}}}(x) = F^n(x)$$

9.19 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.

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

$$F_{T_{\text{fr}}}(x) = 1 - e^{-n\lambda x} \quad \text{and} \quad F_{T_{\text{rb}}}(x) = (1 - e^{-\lambda x})^n$$

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

9.20* More on multicomponent systems. 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 $F_T(x) = P(T < x)$. 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 $p = P(X_i > x) = 1 - F(x)$, and the probability of failure (of 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 $Y \leq k - 1$, hence

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

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

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

This formula is practical if k is small, $k < n/2$. If $k > n/2$, one can better use the “complement” formula

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

9.21* Example. Let a system consist of 6 components whose lifetime is uniform on $(0, 20)$. 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

$$f_T(x) = F'_T(x) = 3(1 - x/20)^2(x/20)^3$$

9.22* Remark. 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 cancellation of all the terms but the last one and gives

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

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

10 Mean Value

10.1 Example: a roulette. A roulette wheel has 18 black spots, 18 red spots and 2 green spots. You can bet \$1 on black or on red and win \$1 if that color comes up or lose \$1 if not (a green spot is always a casino win, and so you 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 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. Does this make sense?

10.2 Example: a 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 $(1 + \dots + 6)/6 = 3.5$. In 100 rolls, you then expect to accumulate $100 \times 3.5 = 350$ total. All right? Note that now you expect to get, approximately, 3.5 per a die roll.

10.3 Concept of the 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 Example 10.1, 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.

Another way to look at the mean value is this: if you obtain n experimental values x_1, \dots, x_n of a random variable X , then their average $(x_1 + \dots + x_n)/n$ should be approximately the mean value of X . While this rule does not necessarily apply to small n , it is quite rigid and stable for large n 's (of order of thousands and millions).

10.4 Rule for the mean values of discrete random variables. If X is a discrete random variable that takes values x_1, x_2, \dots with the corresponding

probabilities p_1, p_2, \dots , then the *mean value* (or *expectation*) of X is

$$E(X) = x_1 p_1 + x_2 p_2 + \dots$$

Note: in Example 10.2 we had $(1 + \dots + 6)/6 = 1 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6}$, so our calculations were consistent with the above rule.

10.5 Mean value of a discrete uniform r.v. Let X take values $1, \dots, n$, each with the same probability $1/n$ (see 4.11). Then

$$E(X) = (1 + \dots + n)/n = \frac{n(n+1)}{2} \cdot \frac{1}{n} = \frac{n+1}{2}$$

10.6 Mean value of a geometric r.v. A geometric random variable X takes values $n = 1, 2, \dots$ with probabilities $P(X = n) = pq^{n-1}$, see 4.7. Its mean value is

$$E(X) = 1 \cdot p + 2 \cdot pq + 3 \cdot pq^2 + 4 \cdot pq^3 + \dots$$

We use the following trick to compute it:

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

In each row, we have a geometric series whose sum can be computed by the rule

$$1 + q + q^2 + q^3 + \dots = 1/(1 - q) = 1/p$$

already used in 4.7. Hence, we obtain

$$E(X) = p/p + pq/p + pq^2/p + pq^3/p + \dots = 1/p$$

10.7* Remark. There is an alternative way to compute $E(X)$ based on the formula

$$E(X) = \sum_{k=1}^{\infty} P(X \geq k)$$

which is valid for any discrete random variable whose values are nonnegative integers $0, 1, 2, \dots$. For the geometric random variable X we can use Problem 4.8 and obtain

$$E(X) = \sum_{k=1}^{\infty} q^{k-1} = 1/(1-q) = 1/p$$

10.8 Mean value of a binomial r.v.. A binomial random variable $X = b(n, p)$ takes values $k = 0, 1, \dots, n$ with probabilities $P(X = k) = C_{n,k} p^k q^{n-k}$, see 4.6. Then

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

where the last sum simply equals one because it adds the probabilities of the random variable $X = b(n-1, p)$, see 4.6. Hence, $E(X) = np$.

10.9 Mean value of a Poisson r.v. A Poisson random variable X takes values $k = 0, 1, 2, \dots$ with probabilities $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, see 4.15. Then

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

The last sum simply equals one because it adds the probabilities of the random variable X (see also our calculation in 4.15). Hence, $E(X) = \lambda$.

10.10 Remark. In the end of 4.14 and 4.16, we already remarked that $\lambda = np$ had an intuitively clear meaning of the average number of successes. Now we see that np , is, indeed, the average (mean) value of $X = b(n, p)$ and λ is, indeed, the average (mean) value of $X = \text{poisson}(\lambda)$. Now we can say that our approximation of a binomial random variable by a Poisson random variable in 4.14 is based on “matching” of their mean values: $\lambda = np$. A very simple rule!

10.11 Rule for the mean values of continuous random variables. If X is a continuous random variable with density function $f(x)$, then

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx$$

If there is a minimum and/or a maximum value of X , then $-\infty$ can be replaced by the minimum and ∞ by the maximum, thus simplifying the integration. (This is similar to the min/max rules in 5.13.)

10.12 Remark. Note the similarity between 10.4 and 10.11: in both cases we multiply the actual values of the random variable, denoted by x , by the probability density, and then the products are added up (the integration is simply a calculus equivalent of summation!).

10.13 Mean value of a uniform r.v.. If $X = U(a, b)$, then $f(x) = 1/(b-a)$ for $a < x < b$, see 5.19. Then

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

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

10.14 Mean value of an exponential r.v.. Let $X = \text{exponential}(\lambda)$. Then integration by parts gives

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

The first term is zero, since $e^{-\infty} = 0$ and then also $x = 0$. The second integral is $\lambda^{-1} \int_0^{\infty} \lambda e^{-\lambda x} dx$, where now the integral equals one (it is the integral of a density function, so by the normalization rule (5.2) it must be equal to one). Hence, $E(X) = 1/\lambda$.

Example 9.19 continued. We have seen in 9.19 that the lifetime of a fragile system is $\text{exponential}(n\lambda)$, if its n components have an $\text{exponential}(\lambda)$ lifetime. Hence, the mean lifetime of the system is $1/(\lambda n)$ while the mean

lifetime of each component is $1/\lambda$. Thus, the system lives, on the average, n times shorter than each component! No wonder we call it fragile.

10.15 Mean value of a standard normal r.v.. Let $Z = N(0, 1)$ be a standard normal random variable. Then

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

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, $E(Z) = 0$.

10.16 Mean value of a Cauchy r.v.. A 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

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

It was involved in the homework problem 3.3.8. What is so special about it? Its density $f_X(x)$ is also even (as the one in 10.15), but its mean value

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

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

What does it mean in practice? Recall: according to our concept of the mean value in 10.3, $E(X)$ is approximately the average of experimental values x_1, \dots, x_n of the random variable X . Now, if you have a sequence of normal $N(0, 1)$ random values, z_1, \dots, z_n , generated by computer or obtained experimentally, then their average $(z_1 + \dots + z_n)/n$ will, indeed, nicely converge to zero (to $E(Z)$) as $n \rightarrow \infty$. On the contrary, if you have a sequence of Cauchy random values, x_1, \dots, x_n , generated by computer or measured experimentally, then their average $(x_1 + \dots + x_n)/n$ will not converge to anything, it will oscillate wildly 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 wonderful process!... So, this is a practical difference between the normal random variable Z in 10.15 and the Cauchy random variable in 10.16.

10.17 Rules for mean values. Since the mean value $E(X)$ is given by an integral, it has properties similar to those of integrals.

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

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

Rule 3. If $Y = X_1 + X_2$, then $E(Y) = E(X_1) + E(X_2)$. Also, if $Y = X_1 - X_2$, then $E(Y) = E(X_1) - E(X_2)$. (This property is called additivity.)

Rule 4. If X is a constant, i.e. takes one value, c , with probability one, then $E(X) = c$.

An example of how these rules can be used: $E(2X - 4Y + 7) = 2 E(X) - 4 E(Y) + 7$.

10.18 Mean value of a normal r.v. Recall that an arbitrary normal random variable $Y = N(\mu, \sigma^2)$ satisfies $Y = \mu + \sigma Z$, cf. 8.6. Then by Rules 1-4 we have $E(Y) = \mu + \sigma E(Z) = \mu$.

10.19 Bernoulli random variable. Recall (4.3) that a Bernoulli trial is a simple experiment with two possible outcomes: a success (labelled S) and a failure (labelled F). Success occurs with probability p and failure with probability $q = 1 - p$. Let us mark successes by 1 and failures by 0. Then we get a random variable X that takes two values: 1 (with probability p) and 0 (with probability q). This is called a Bernoulli random variable. Its mean value is, obviously, $E(X) = 1 \cdot p + 0 \cdot q = p$.

10.20 Mean value of a binomial r.v., alternatively. Recall (4.4, 4.6) that a binomial random variable is the number of successes in n independent Bernoulli trials. According to 10.19, with n independent Bernoulli trials we associate n independent Bernoulli random variables X_1, \dots, 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 $X_1 + \dots + X_n$ gives exactly the number of

successes in n Bernoulli trials! Therefore,

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

where $X = b(n, p)$ is the binomial random variable and X_i are independent Bernoulli random variables. Now, using the additive rule 3 in 10.17 yields $E(X) = E(X_1) + \cdots + E(X_n) = np$, since $E(X_i) = p$ by 10.19.

10.21 Mean value of a function of r.v. Let X be a random variable, and $y = g(x)$ a function. Then $Y = g(X)$ is another random variable, as in 7.1. Here we provide rules to compute the mean value $E(Y)$.

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

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

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

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

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

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

Solution. Since X only takes values 0 and 1, then $X^k = X$, so $E(X^k) = E(X) = p$.

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

Solution. Recall (5.20) that $f(x) = 1$ for $0 < x < 1$. Then

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

10.25 Special rule for independent r.v. If X and Y are independent random variables, then $E(XY) = E(X) \cdot E(Y)$.

This rule works for independent random variables and usually fails for dependent ones. For example, let X be Bernoulli random variable and $Y = X$. Then $E(XY) = E(X^2) = p$, rather than $E(X) \cdot E(Y) = [E(X)]^2 = p^2$.

10.26* Remark. Just like in Remark 10.7, there is also an alternative way to compute the mean value of a continuous nonnegative random variable $X \geq 0$:

$$E(X) = \int_0^{\infty} P(X > x) dx$$

This can simplify, for example, the integration in 10.14: there $P(X > x) = 1 - F(x) = e^{-\lambda x}$, hence

$$E(X) = \int_0^{\infty} e^{-\lambda x} dx = 1/\lambda$$

11 Variance

11.1 Motivation. We have seen that the mean value $E(X)$ of a random variable X gives the most fair expectation of X . Is this enough to predict X in practice?

Suppose you are going to visit Montana next March and wonder what the temperature there might be. The climatological data from a reference book say that the average daily temperature in Montana in March is 42° F. This is exactly like knowing the mean value of a random variable. Is this enough for you? You quickly realize that the actual daily temperature might fluctuate around 42° . If typical fluctuations are small, then 42° can be a pretty accurate estimate. 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 42° 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 42° with the range of typical fluctuations (variations). For example, 42 ± 3 would say that the weather is stable and the temperature between 39 and 45 degrees can be reasonably expected. Or, on the contrary, 42 ± 23 would tell you that the weather is very unstable and you should expect anything from 19° F to 65° F. The conclusion is that the range of typical fluctuations around the mean value is practically as important as the mean value itself.

11.2 Measuring variations. The difference between the actual value of X and its mean value is $X - E(X)$. Should we just find the average of this difference? Let us try this: $E[X - E(X)]$ =(by the rules of 10.17)= $E(X) - E(X) = 0$. It gives us nothing! The reason is clear: positive fluctuations ($X - E(X) > 0$) and negative fluctuations ($X - E(X) < 0$) cancel out in the end. Let us square $X - E(X)$ to ensure that all fluctuations are taken as positive numbers.

11.3 Variance. The variance of a random variable X is defined to be

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

Alternatively, one can use

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

To see that these two formulas are equivalent, let us expand the square in the first one and use the rules of 10.17:

$$E[X - E(X)]^2 = E(X^2 - 2X \cdot E(X) - [E(X)]^2) = E(X^2) - 2[E(X)]^2 + [E(X)]^2$$

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

11.4 Examples.

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

(b) Let $X = U(0, 1)$. Then, recall 10.24, $\text{Var}(X) = 1/3 - 1/4 = 1/12$.

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

Note: In these examples one can easily find all possible deviations of X from its mean value $E(X)$, 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 $\text{Var}(X)$ found above? The main reason is that $\text{Var}(X)$ measures *squared* deviations, rather than deviations. So, we need to take the square root of $\text{Var}(X)$, to estimate the deviations of X .

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

$$\sigma_X = \sqrt{\text{Var}(X)}$$

Examples 11.4 continued. The standard deviations in those examples are: in (a) $\sigma_X = 1/2$, in (b) $\sigma_X \approx 0.29$, in (c) $\sigma_X \approx 1.71$.

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 in 11.5 rather than with precisely computed average deviation.

In any case, it is traditional in probability theory to work with standard deviations.

11.6 Rules for variance and standard deviation.

Rule 1. If X is a constant, i.e. takes one value, c , with probability one, then $\text{Var}(X) = 0$ and $\sigma_X = 0$.

Rule 2. If $Y = aX$, where a is a constant, then $\text{Var}(Y) = a^2 \text{Var}(X)$. (That is, a constant must be squared before it can be factored out. For example, $\text{Var}(2X) = 4 \text{Var}(X)$, $\text{Var}(-5X) = 25 \text{Var}(X)$, $\text{Var}(-X) = \text{Var}(X)$, etc.). In this case, also, $\sigma_Y = |a|\sigma_X$.

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

Rule 4. If $Y = X_1 + X_2$, and X_1 and X_2 are independent, then $\text{Var}(Y) = \text{Var}(X_1) + \text{Var}(X_2)$. 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 $\text{Var}(X) = 0$. In Rule 3, adding a constant means shifting all values of X by a fixed amount. In this case the mean value $E(X)$ is shifted by the same amount, so all the deviations of X from its mean value will not change, this is why $\text{Var}(X)$ and σ_X remain unchanged. In Rule 4, does it remind you the Pythagorean theorem? There is, indeed, a deep connection, but we will not explore it.

Rule 4 can be obtained by the following calculation:

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

Now, by 10.25, we have $E(X_1X_2) = E(X_1) \cdot E(X_2)$, so the terms $2E(X_1X_2)$ and $2E(X_1) \cdot E(X_2)$ cancel out. The remaining terms can be easily grouped to make $\text{Var}(X_1) + \text{Var}(X_2)$.

11.7 A tricky question. Let X and Y be independent. Is it true that $\text{Var}(X - Y) = \text{Var}(X) - \text{Var}(Y)$?

Answer. No. This is an incorrect “application” of Rule 4. The correct application is

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

where the last step was done due to Rule 2.

We note also that $\text{Var}(X) \geq 0$ and $\sigma_X \geq 0$. Moreover, $\text{Var}(X) = 0$ and $\sigma_X = 0$ only if X is constant.

11.8 Problem. Can we have a random variable with $E(X) = 4$ and $E(X^2) = 13$?

Solution. Such a random variable would have $\text{Var}(X) = 13 - 4^2 = -3$, which is impossible since $\text{Var}(X) \geq 0$.

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

Solution. Using the rules of 10.16 gives $E(Z) = 3 \cdot E(X) - 2 \cdot E(Y) - 2 = 15 + 6 - 2 = 19$. Using the rules of 11.6 gives

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

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

11.10 Variance of a Bernoulli r.v. Recall (10.19) that a Bernoulli random variable X takes values 1 and 0 with probabilities p and q , respectively. Also note that $X^2 = X$. Then $E(X^2) = E(X) = p$, so $\text{Var}(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1 - p) = pq$.

11.11 Variance of a binomial r.v. Recall (10.20) that a binomial random variable $X = b(n, p)$ is the sum $X = X_1 + \cdots + X_n$ of n independent Bernoulli random variables. Hence, by Rule 4 of 11.6

$$\text{Var}(X) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = npq$$

11.12 Meaning of σ_X . The standard deviation σ_X is a typical (average) deviation of the random variable X from its mean value $E(X)$. Hence, one can expect that X typically takes values $E(X) \pm \sigma_X$, in a way described in 11.1. We will express this by a “formula”

$$X \approx E(X) \pm \sigma_X$$

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.

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

11.13 Binomial r.v. for large n . Let $X = b(n, p)$ be a binomial random variable. We know already that $E(X) = np$ and $\sigma_X = \sqrt{npq}$, see 11.11. Hence, by the pattern of 11.12, 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 $p = q = 1/2$ (like in tossing of a coin, where X is 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 much more slowly than n and $E(X) = n/2$ do. So, relative to $E(X)$, the deviations become less and less visible.

11.14 Relative frequency. The contrast between $E(X)$ and σ_X in the above example 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, $E(\bar{X}_n) = E(X)/n = p$ and $\text{Var}(\bar{X}_n) = \text{Var}(X)/n^2 = pq/n$. Hence,

$$\bar{X}_n = p \pm \sqrt{pq}/\sqrt{n}$$

For an illustration, again let $p = q = 1/2$. Then

$$\bar{X}_n = \frac{1}{2} \pm \frac{1}{2\sqrt{n}}$$

We see that as n increases, the typical deviations decrease to zero. Hence, \bar{X}_n concentrates more and more tightly near its mean value $1/2$. So, as

$n \rightarrow \infty$, 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 Section 14.

12 Moment Generating Function

12.1 Moment generating function. Let X be a random variable, and $g(x) = e^{tx}$ a function of x (here t is an additional parameter). Then $Y = g(X) = e^{tX}$ is another random variable. The rule 10.21 tells us how to compute $E(Y)$. Note that the value of $E(Y) = E(e^{tX})$ 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 :

$$M_X(t) = E(e^{tX})$$

The parameter t becomes the argument of this function.

12.2 Calculating a moment generating function. According to 10.21, if X is a discrete random variable and takes values x_1, x_2, \dots with probabilities p_1, p_2, \dots , then

$$M_X(t) = E(e^{tX}) = e^{tx_1}p_1 + e^{tx_2}p_2 + \dots$$

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

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

12.3 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 12.2, with $x_1 = -1$, $x_2 = 5$ and $p_1 = 1/3$, $p_2 = 2/3$. (b) No, because the sum of coefficients $1/2 + 1/3$ is not equal to one.

12.4 Generating moments. What is the use of the m.g.f. $M_X(t)$? Let us differentiate it and substitute $t = 0$. It is easier done with discrete random variables:

$$M'_X(t) = e^{tx_1}x_1p_1 + e^{tx_2}x_2p_2 + \dots$$

and the substitution $t = 0$ eliminates all the exponential factors, leaving only $x_1p_1 + x_2p_2 + \dots$, which is $E(X)$. So we arrive at $M'_X(0) = E(X)$. Differentiating once more we get

$$M''_X(t) = e^{tx_1}x_1^2p_1 + e^{tx_2}x_2^2p_2 + \dots$$

and the substitution $t = 0$ gives $x_1^2 p_1 + x_2^2 p_2 + \dots$, which is $E(X^2)$, so $M_X''(0) = E(X^2)$, which is the second moment of X . In the same way we get

$$M^{(k)}(0) = E(X^k)$$

Hence, to compute the k th moment of X we can differentiate the function $M_X(t)$ k times and then substitute $t = 0$. This is its main use – 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.

12.5 M.g.f. for an exponential r.v.. If X is an exponential r.v., then $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, and

$$M_X(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \int_0^\infty \lambda e^{-(\lambda-t)x} dx = \frac{\lambda}{\lambda-t}$$

Taking derivatives and substituting $t = 0$ gives

$$E(X) = M_X'(0) = \frac{\lambda}{(\lambda-t)^2} \Big|_{t=0} = \frac{1}{\lambda}$$

(which we know already, recall 10.14), and

$$E(X^2) = M_X''(0) = \frac{2\lambda}{(\lambda-t)^3} \Big|_{t=0} = \frac{2}{\lambda^2}$$

Therefore, $\text{Var}(X) = 2/\lambda^2 - (1/\lambda)^2 = 1/\lambda^2$ and $\sigma_X = 1/\lambda$.

12.6 Remark: unpredictability of an exponential r.v.. In the spirit of 11.11, we can represent an exponential random variable as $X = E(X) \pm \sigma_X = \lambda^{-1} \pm \lambda^{-1}$. Stop! Does it make sense that the typical deviations are about the same as the mean value? This means that the exponential random variable is completely unpredictable! For example, suppose on a long highway, state trooper patrol cars are deployed randomly, approximately one every 20 miles. If you drive on the highway, then the distance to the next patrol car can be regarded as an exponential random variable, with the mean value of 20 miles (which it is, as we will see later in Section 17). But then the typical fluctuations about the mean value are also about 20 miles! So, the expected

distance to the next patrol car is 20 ± 20 miles. If you (accidentally) go over the legal speed limit and a trooper happens to be nearby and pulls you over, no need to complain about bad luck: zero distance to the trooper is within the expected range 20 ± 20 .

12.7 M.g.f. for a standard normal r.v.. If Z is a standard normal random variable, then

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} e^{t^2/2} dx$$

The factor $e^{t^2/2}$ 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 $N(t, 1)$ (recall 8.7). Hence, the remaining integral will be equal to one by the normalization rule (5.2). We thus get

$$M_Z(t) = e^{t^2/2}$$

12.8 A rule for m.g.f. Let X be a random variable with moment generating function $M_X(t)$, and $Y = a + bX$, where a and $b > 0$ are constants. Then

$$M_Y(t) = e^{at} \cdot M_X(bt)$$

Indeed, by 12.1,

$$M_Y(t) = E(e^{tY}) = E(e^{at+btX}) = E(e^{at} e^{btX})$$

Now, e^{at} does not contain X , so it can be factored out by Rule 1 in 10.17. Also, $E(e^{btX}) = M_X(bt)$ by the formula in 12.1, where t is replaced by bt .

12.9 M.g.f. for a normal r.v.. If X is a normal random variable $N(\mu, \sigma^2)$, then $X = \mu + \sigma Z$ by 8.6. Now, by the rule of 12.8 and the formula of 12.6 we get

$$M_X(t) = e^{\mu t} e^{(\sigma t)^2/2} = e^{\mu t + \sigma^2 t^2/2}$$

12.10 Variance of a normal r.v.. Differentiating $M_X(t)$ in 12.9 gives

$$M'_X(t) = (\mu + \sigma^2 t) e^{\mu t + \sigma^2 t^2/2}$$

and so $E(X) = M'_X(0) = \mu$, which we know already (10.18).

With a little extra work, we can differentiate once again:

$$M''_X(t) = (\sigma^2 + \mu^2 + 2\mu\sigma^2t + \sigma^4t^2)e^{\mu t + \sigma^2t^2/2}$$

and so $E(X^2) = M''_X(0) = \sigma^2 + \mu^2$. Therefore,

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \sigma^2$$

and then $\sigma_X = \sqrt{\sigma^2} = \sigma$.

12.11 Notational remark. Now we see that the second parameter of a normal random variable $X = N(\mu, \sigma^2)$ is its standard deviation $\sigma = \sigma_X$. It is now clear why it is denoted by σ , it so conveniently matches the more general notation σ_X . Moreover, in many textbooks even the mean value of random variables X is denoted by μ_X , again to match the first parameter of the normal random variable.

12.12 M.g.f. for a Poisson r.v.. A Poisson random variable X takes values $k = 0, 1, 2, \dots$ with probabilities $P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$, see 4.15. Then

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

The factor $e^{\lambda e^t - \lambda}$ 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,

$$M_X(t) = e^{\lambda e^t - \lambda} = e^{\lambda(e^t - 1)}$$

a “double exponential” function.

12.13 Variance of a Poisson r.v.. Differentiating $M_X(t)$ in 12.12 gives

$$M'_X(t) = \lambda e^t e^{\lambda(e^t - 1)}$$

and so $E(X) = M'_X(0) = \lambda$, which we know already.

With a little extra work, we can differentiate once again:

$$M_X''(t) = (\lambda^2 + \lambda) e^t e^{\lambda(e^t-1)}$$

and so $E(X^2) = M_X''(0) = \lambda^2 + \lambda$. Therefore

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \lambda$$

and then $\sigma_X = \sqrt{\lambda}$.

12.14 M.g.f. for a geometric r.v.. A geometric random variable X takes values $n = 1, 2, \dots$ with probabilities $P(X = n) = pq^{n-1}$, see 4.7. Then

$$M_X(t) = \sum_{n=1}^{\infty} pq^{n-1} e^{tn} = pe^t \sum_{n=1}^{\infty} (qe^t)^{n-1}$$

This is a geometric series, so

$$M_X(t) = \frac{pe^t}{1 - qe^t}$$

By differentiating this function (we omit tedious details) one gets $E(X) = 1/p$ and $\text{Var}(X) = q/p^2$.

12.15 Summary. The above results are summarized in a chart below:

	$E(X)$	$\text{Var}(X)$	σ_X	$M_X(t)$
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 $N(0, 1)$	0	1	1	$e^{-t^2/2}$
normal $N(\mu, \sigma^2)$	μ	σ^2	σ	$e^{\mu t + \sigma^2 t^2/2}$

Remark: in the chart 12.15, the variance of $U(a, b)$ can be easily found by direct integration: if $X = U(a, b)$, then

$$\begin{aligned}\text{Var}(X) &= E(X^2) - [E(X)]^2 = \int_a^b \frac{x^2}{b-a} dx - \frac{(a+b)^2}{4} \\ &= \frac{b^3 - a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{b^2 + a^2 - 2ab}{12}\end{aligned}$$

12.16 A special rule for m.g.f.. If X and Y are independent random variables, then

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

Indeed, we have

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) = (\text{by 10.25}) = E(e^{tX}) \cdot E(e^{tY}) = M_X(t)M_Y(t)$$

12.17 Stable distributions. The special rule 12.16 can be used to find the distribution of $X + Y$, if X and Y are given independent random variables. For example, if $X = N(\mu_1, \sigma_1^2)$ and $Y = N(\mu_2, \sigma_2^2)$ are two independent **normal** random variables, then $X + Y$ has m.g.f.

$$M_{X+Y}(t) = e^{\mu_1 t + \sigma_1^2 t^2 / 2} \cdot e^{\mu_2 t + \sigma_2^2 t^2 / 2} = e^{(\mu_1 + \mu_2)t + (\sigma_1^2 + \sigma_2^2)t^2 / 2}$$

By 12.15, we got the m.g.f. of the normal random variable $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Hence, the sum of two independent normal random variables is also normal, and we get a rule:

$$N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Similarly, if $X = \text{poisson}(\lambda_1)$ and $Y = \text{poisson}(\lambda_2)$, then

$$M_{X+Y}(t) = e^{\lambda_1(e^t - 1)} e^{\lambda_2(e^t - 1)} = e^{(\lambda_1 + \lambda_2)(e^t - 1)}$$

hence $X + Y = \text{poisson}(\lambda_1 + \lambda_2)$. Lastly, if $X = b(n_1, p)$ and $Y = b(n_2, p)$, then

$$M_{X+Y}(t) = (pe^t + q)^{n_1} (pe^t + q)^{n_2} = (pe^t + q)^{n_1 + n_2}$$

hence $X + Y = b(n_1 + n_2, p)$. The last conclusion is not surprising, though, since $X + Y$ is the total number of successes in two series of trials, of lengths n_1 and n_2 , respectively. The trials are independent, so we have $n_1 + n_2$

similar trials (with probability of success p in each). Now it is clear that $X + Y = b(n_1 + n_2, p)$.

If the type of distribution is preserved when two independent random variables with distributions of that type are added, we call such distributions stable. Hence, binomial, Poisson, and normal distributions are stable. The others (uniform, geometric, exponential) are not.

12.18 Manipulations with normal random variables. We know from Section 8 that when you rescale a normal random variable, you get another normal random variable. Now we know that if you add two independent normal random variables, you get another normal random variable. Combining these two rules, we can state:

12.19 Rule for normals. Let $X = N(\mu_1, \sigma_1^2)$ and $Y = N(\mu_2, \sigma_2^2)$ be independent normal random variables, and a, b two constants. Then $W = aX + bY$ is a normal random variable

$$W = N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

12.20 Example. Let $X = N(-1, 3)$ and $Y = N(2, 1)$ be independent random variables. Compute the probabilities of the following events: (a) $P(X + 2Y > 2)$, (b) $P(X > Y)$, (c) $P(2X < Y - 2)$.

Solution. We have $W = X + 2Y = N(-1 + 2 \cdot 2, 3 + 2^2 \cdot 1) = N(3, 7)$, hence

$$P(X + 2Y > 2) = P(W > 2) = 1 - \Phi\left(\frac{2 - 3}{\sqrt{7}}\right) = 0.6480$$

Similarly, in the case (b) we have $V = X - Y = N(-1 - 2, 3 + 1) = N(-3, 4)$, hence

$$P(X > Y) = P(X - Y > 0) = 1 - \Phi\left(\frac{0 - (-3)}{\sqrt{4}}\right) = 0.0668$$

In the case (c), $R = 2X - Y = N(2(-1) - 2, 2^2 \cdot 3 + 1) = N(-4, 13)$, hence

$$P(2X < Y - 2) = P(2X - Y < -2) = \Phi\left(\frac{(-2) - (-4)}{\sqrt{13}}\right) = 0.7088$$

13 Covariance and Correlation

This is a measure of dependence between random variables.

13.1 Covariance. In 11.6, we derived a rule for the variance $\text{Var}(X_1 + X_2)$ and obtained

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2E(X_1X_2) - 2E(X_1) \cdot E(X_2)$$

By the special rule 10.25 for independent random variables $E(X_1X_2) = E(X_1) \cdot E(X_2)$, and then the last two terms cancel, thus we had the Rule 4 in 11.6. Here we are going to consider dependent random variables. Then the quantity

$$\text{Cov}(X_1, X_2) = E(X_1X_2) - E(X_1) \cdot E(X_2) \quad (13.1)$$

may not be zero. We call it the *covariance* between X_1 and X_2 . Alternatively, one can put

$$\text{Cov}(X_1, X_2) = E[(X_1 - E(X_1))(X_2 - E(X_2))] \quad (13.2)$$

The equivalence of (13.1) and (13.2) can be easily verified just like in 11.3.

13.2 A rule for variance. Now, the rule for the variance can be stated as

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$$

This rule is valid for any random variables X_1 and X_2 , independent or not. Note that the above equation resembles a simple algebraic formula $(a + b)^2 = a^2 + b^2 + 2ab$.

13.3 Covariance and dependence. For independent random variables X_1 and X_2 we have $\text{Cov}(X_1, X_2) = 0$. 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 may have zero covariance (see 13.7 below).

13.4 Rules for covariance.

(a) For any random variable X

$$\text{Cov}(X, X) = \text{Var}(X)$$

(b) The covariance is symmetric:

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

(c) The covariance is linear in both arguments:

$$\text{Cov}(a_1X_1 + a_2X_2, Y) = a_1 \text{Cov}(X_1, Y) + a_2 \text{Cov}(X_2, Y)$$

$$\text{Cov}(X, b_1Y_1 + b_2Y_2) = b_1 \text{Cov}(X, Y_1) + b_2 \text{Cov}(X, Y_2)$$

(d) If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

(e) If X is a constant, i.e. takes one value with probability one, then $\text{Cov}(X, Y) = 0$ for any r.v. Y .

One can verify these rules easily based on the definition of covariance in 13.1.

13.5 Example. Let $X = U(0, 1)$. Find $\text{Cov}(X, X^2)$.

Solution. Here we use (13.1) and 10.24:

$$\text{Cov}(X, X^2) = E(X^3) - E(X) \cdot E(X^2) = \frac{1}{4} - \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{12}$$

13.6 Example. Let X and Y be independent uniform $U(0, 1)$ random variables. Find $\text{Cov}(X + 2Y, X^2 - Y)$.

Solution. Here we use the rules of 13.4:

$$\begin{aligned} \text{Cov}(X + 2Y, X^2 - Y) &= \text{Cov}(X, X^2) + 2\text{Cov}(Y, X^2) - \text{Cov}(X, Y) - 2\text{Cov}(Y, Y) \\ &= \frac{1}{12} + 0 + 0 - 2 \times \frac{1}{12} = -\frac{1}{12} \end{aligned}$$

13.7 Example. A random variable X takes three values $-2, 0$ and 2 , with probability $1/3$ each. Let $Y = X^2$. Compute $\text{Cov}(X, Y)$.

Solution. It is easy to see that $E(XY) = E(X^3) = 0$, and also $E(X) = 0$ and $E(Y) = 8/3$ (the value of $E(Y)$ is not important, though). Then $\text{Cov}(X, Y) = 0$.

Note: in this example X and Y are obviously dependent (having X one can compute Y precisely). But, for some strange reason, X and Y have no covariance. This “mystery” is cleared in 13.12 below.

13.8 Correlation. If $\text{Cov}(X, Y) \neq 0$, it tells us that X and Y are dependent. But how strong is the dependence? The following modification of covariance estimates the strength of dependence quite accurately:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

This is called the *correlation*, or *correlation coefficient*, between X and Y .

13.9 Rules for correlation.

The correlation only takes values in the interval $[-1, 1]$, that is

$$-1 \leq \rho(X, Y) \leq 1$$

(b) If X and Y are proportional, i.e. $Y = aX + b$ with some constants a and b , then $\rho(X, Y) = \pm 1$. Precisely, $\rho(X, Y) = 1$ if $a > 0$ and $\rho(X, Y) = -1$ if $a < 0$.

Remark: It is interesting that the converse of 13.9(b) is true: if $\rho(X, Y) = \pm 1$, then X and Y are proportional, i.e. $Y = aX + b$ with probability one.

13.10 Correlation as measure of dependence. The above rules justify the following common interpretation of the correlation coefficient:

- (a) If $\rho(X, Y) = 0$, then X and Y are *practically* independent (one often says that X and Y are *uncorrelated*).
- (b) 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).
- (c) 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 13.5 continued. Compute $\rho(X, X^2)$.

Solution. Recall that $\text{Var}(X) = 1/12$. By the rules of 10.24, $\text{Var}(X^2) = E(X^4) - [E(X^2)]^2 = 1/5 - (1/3)^2 = 4/45$. Hence,

$$\rho(X, X^2) = \frac{\text{Cov}(X, X^2)}{\sigma_X \cdot \sigma_{X^2}} = \frac{1/12}{\sqrt{1/12} \cdot \sqrt{4/45}} \approx 0.968$$

A very strong dependence! Indeed, X and X^2 are obviously dependent: knowing X one can compute $X^2 = X \cdot X$, and knowing X^2 one can compute

$$X = \sqrt{X^2}.$$

13.11 The sign of correlation. If $\rho(X, Y) > 0$, then X and Y are positively correlated. This means that if X happens to be above its mean value (fluctuates upward), then Y is also very likely to be above its mean value. If X is below its mean value, then the same probably happens to Y .

On the contrary, if $\rho(X, Y) < 0$, then X and Y are negatively correlated, i.e. fluctuate in opposite directions: when X goes up Y goes down, and vice versa.

13.12 Examples. Recall that in Example 13.5 we have $\rho(X, X^2) = 0.968$, a very high positive value. Indeed, X and X^2 are positively correlated: the higher X , the higher X^2 .

In Example 13.7, we have $\rho(X, Y) = 0$. 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 from its mean value sends Y in one direction – upward. This is a good enough reason why X and Y are uncorrelated!

We now turn to two more, quite special, numerical characteristics of random variables.

13.13 Skewness. The skewness of a random variable X is

$$\beta_1 = \left(\frac{E[X - E(X)]^3}{\sigma_X^3} \right)^2$$

It characterizes the degree of *asymmetry* of the density of X about the mean value $E(X)$. For example, $\beta_1 = 0$ for any normal random variable $N(\mu, \sigma^2)$, because its density is perfectly symmetric about its mean value μ . The same is true for any uniform random variable $U(a, b)$.

13.14 Kurtosis. The kurtosis of a random variable X is

$$\beta_2 = \frac{E[X - E(X)]^4}{\sigma_X^4}$$

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 $E(X)$. More precisely, if

$f(x)$ has heavy (thick) tails far from the mean value $E(X)$, then the kurtosis is high. 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 $E(X)$. For normal random variables we have $\beta_2 = 3$.

14 Law of Large Numbers

14.1 Motivation: relative frequency. We resume our discussion in 11.14. Recall that we had $X = b(n, p)$, and then $\bar{X}_n = X/n$ was the relative frequency of successes in a series of n Bernoulli trials. Its mean value was $E(\bar{X}_n) = p$ and its variance $\text{Var}(\bar{X}_n) = pq/n$. We concluded in 11.14 that \bar{X}_n concentrated more and more tightly near its mean value p , as n grows. Here we investigate this issue further. It is one of the central issues in probability theory.

We need to estimate precisely by how much a random variable can deviate from its mean value.

14.2 Markov inequality. Let $X \geq 0$ be a random variable, and $t > 0$ a real number. Then

$$P(X \geq t) \leq \frac{E(X)}{t}$$

This inequality estimates the probability that X takes large values.

14.3 Example. Let X be nonnegative and $E(X) = \mu$. What can we say about the probability $P(X \geq 10\mu)$?

Solution: By Markov inequality,

$$P(X \geq 10\mu) \leq \frac{\mu}{10\mu} = \frac{1}{10}$$

Hence, there is not much chance (at most 10%) that a random variable exceeds ten times its mean value.

14.4 Chebyshev's inequality. Let X be a random variable, and $y > 0$ a real number. Then

$$P(|X - E(X)| \geq y) \leq \frac{\text{Var}(X)}{y^2}$$

This inequality estimates the probability that X deviates by more than y from its mean value $E(X)$.

14.5 Example. Let $X = N(\mu, \sigma^2)$. We can estimate the probability $P(|X - \mu| \geq 3\sigma)$ by Chebyshev's inequality: $P(|X - \mu| \geq 3\sigma) \leq \sigma^2/(3\sigma)^2 =$

1/9. We also know that $P(|X - \mu| \geq 3\sigma) = 2(1 - \Phi(3)) = 0.0028$, from the table for the function Φ . Since $0.0028 < 1/9$, the estimate is correct.

14.6 Remark. In the above example, the estimate $1/9 = 0.1111$ is quite inaccurate, 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 (per month). Such a statement is correct but practically useless, if not ridiculous. Back to Chebyshev's inequality, indeed, for most random variables it is an overkill - it overestimates $P(|X - E(X)| \geq y)$, it gives a bound well above the actual value. On the other hand, Chebyshev's inequality is universal, it applies to *any* random variable. It cannot be improved without sacrificing universality, because for every $y > 0$ there is an "ugly" random variable X for which $P(|X - E(X)| \geq y) = \text{Var}(X)/y^2$, i.e. Chebyshev's inequality becomes identity. We will see that below.

14.7 The proof of Markov inequality is quite simple and instructive. Given X , define a new random variable X_1 by

$$X_1 = \begin{cases} t & \text{if } X \geq t \\ 0 & \text{otherwise} \end{cases}$$

Note that, obviously, $X_1 \leq X$, hence $E(X_1) \leq E(X)$. On the other hand, X_1 takes only two values (0 and t), so its mean value can be easily computed:

$$E(X_1) = t \cdot P(X_1 = t) = t \cdot P(X \geq t)$$

Combining these two facts gives Markov inequality.

14.8 The proof of Chebyshev's inequality is also simple and instructive. Given X , define a new random variable X_1 by

$$X_1 = \begin{cases} y^2 & \text{if } |X - E(X)| \geq y \\ 0 & \text{otherwise} \end{cases}$$

Note that, obviously, $X_1 \leq |X - E(X)|^2$, hence $E(X_1) \leq E(|X - E(X)|^2) = \text{Var}(X)$. On the other hand, $E(X_1) = y^2 \cdot P(X_1 = y^2) = y^2 \cdot P(|X - E(X)| \geq y)$. Combining these two facts gives Chebyshev's inequality.

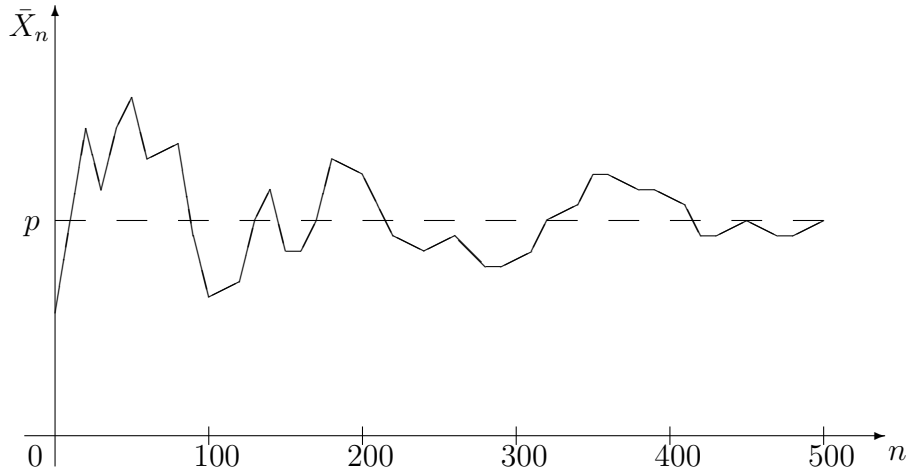
The proof also suggests an idea how to construct a random variable X for which $P(|X - E(X)| \geq y) = \text{Var}(X)/y^2$. Such a random variable must

satisfy the condition $X_1 = |X - E(X)|^2$, where X_1 is defined above. In other words, X can only take two values: $E(X) + y$ and $E(X) - y$. The rest of the construction is simple and left to the students.

14.9 Example. Suppose that a random variable X has mean value $E(X) = 10$ and variance $\text{Var}(X) = 9$. By using Chebyshev's inequality, estimate the probability $P(X \geq 25)$.

Solution. The event $\{X \geq 25\}$ can be represented as $\{X \geq 10 + 15\}$ or $\{X - 10 \geq 15\}$. It is a subset of the event $\{|X - 10| \geq 15\}$. Then by Chebyshev's inequality

$$P(X \geq 25) \leq P(|X - 10| \geq 15) \leq \frac{9}{15^2} = \frac{1}{25}$$



The convergence of \bar{X}_n to $p = 1/2$

14.10 Back to relative frequency. As in 11.14, 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 $E(\bar{X}_n) = p$ and $\text{Var}(\bar{X}_n) = pq/n$. Let $y > 0$ be any number. By Chebyshev's inequality

$$P(|\bar{X}_n - p| \geq y) \leq \frac{pq}{y^2 n}$$

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

$n \rightarrow \infty$. This is true for all $y > 0$. Hence, all deviations of \bar{X}_n from p vanish as $n \rightarrow \infty$. The random variable \bar{X}_n indeed converges to p as $n \rightarrow \infty$, as we guessed in 11.14.

We observed in 10.20 that a binomial random variable $X = b(n, p)$ 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 $\bar{X}_n = (X_1 + \cdots + X_n)/n$, i.e. \bar{X}_n is the experimental average of the Bernoulli random variables X_1, \dots, X_n . The above fact then says that the experimental average of X_1, \dots, X_n converges to the (theoretical) mean value $E(X_i) = p$ as $n \rightarrow \infty$. In this form, the fact can be extended to more general random variables:

14.11 Law of Large Numbers. Let X_1, X_2, \dots be independent identically distributed (i.i.d) random variables. Let $\mu = E(X_i)$ be the common mean value of all X_i 's, and $\sigma^2 = \text{Var}(X_i)$ the common variance of all X_i 's. For each n let

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \quad (14.1)$$

be the (experimental) average of X_i 's up to n . Then the random variable \bar{X}_n converges to μ as $n \rightarrow \infty$. Precisely, for every real number $y > 0$ we have

$$P(|\bar{X}_n - \mu| \geq y) \leq \frac{\sigma^2}{y^2 n} \rightarrow 0$$

as $n \rightarrow \infty$.

Indeed, the above inequality is simply Chebyshev's inequality, because $\text{Var}(\bar{X}_n) = [\text{Var}(X_1) + \cdots + \text{Var}(X_n)]/n^2 = \sigma^2/n$.

14.12 Time average. Recall that one gets a sequence of i.i.d. random variables every time a random experiment repeated under the same conditions. Then the obtained experimental data X_1, \dots, X_n are i.i.d. random variables. Here n plays the role of time, and the experimental average (14.1) can be called the *time average*. Then the law of large numbers says that the time average approaches, as time goes on, the mean value $\mu = E(X_i)$. That is, \bar{X}_n is getting closer and closer to μ and sooner or later becomes practically indistinguishable from μ .

14.13* Monte-Carlo Integration. The law of large numbers can be used

to integrate functions. For example, consider a double integral

$$\iint_R f(x, y) dx dy \quad (14.2)$$

over a region R (for simplicity, let R be a part of the unit square $0 \leq x, y \leq 1$). Such integrals are often hard to compute, because either the function f or the region R , or both, may be 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 $(x_1, y_1), (x_2, y_2), \dots$. Each pair (x_i, y_i) is generated by calling an RNG, cf. 5.21, 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

$$\frac{f_1 + \dots + f_n}{n} \quad (14.3)$$

approximates the integral (14.2). The larger n the better approximation. Easy, isn't it?

Since this algorithm is based on random numbers (i.e., resembles playing roulette), it is called *Monte-Carlo integration*.

Note that our Remark 9.8 describing a way to compute the number π with the help of an RNG was exactly a case of the 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 (14.2) actually coincided with the area of R , see again 9.7).

A drawback of the Monte-Carlo integration is a slow convergence of the average (14.3) to the integral (14.2). One really needs to compute a lot of f_i 's in (14.3) to obtain an accurate value of (14.2). We will see that in the next section.

15 Central Limit Theorem

15.1 Binomial r.v. for large n . Let $X = b(n, p)$ be a binomial random variable with large n . Its density function is $P(X = k) = C_{n,k}p^k(1-p)^{n-k}$ for $0 \leq k \leq n$, cf. (4.1). As we pointed out in 4.14, this formula for $P(X = k)$ is practically useless for large n . We found in 4.14 a simple approximation formula (Poisson law) that works for large n and small p . But what if p is not small?

Remarkably, then the binomial random variable $X = b(n, p)$ can be well approximated by a normal random variable, $Y = N(\mu, \sigma^2)$. The calculations that lead to this approximation are quite complicated, so we skip them. Instead, let us consider the proper choice of μ and σ^2 , the parameters of Y . Recall that $\mu = E(Y)$ and $\sigma^2 = \text{Var}(Y)$. Now if we want the normal random variable $Y = N(\mu, \sigma^2)$ to match the binomial random variable X , then we expect that their mean values are equal, $E(X) = E(Y)$, and their variances are equal, $\text{Var}(X) = \text{Var}(Y)$. This gives the choice

$$\mu = E(Y) = E(X) = np \quad \text{and} \quad \sigma^2 = \text{Var}(Y) = \text{Var}(X) = npq$$

15.2 De Moivre-Laplace Theorem. The binomial random variable $X = b(n, p)$ is approximated by a normal $Y = N(\mu, \sigma^2)$ with $\mu = np$ and $\sigma^2 = npq$.

15.3 Example. Toss a coin 100 times and let X be the number of Heads. Find the probability $P(X \leq 50)$.

Solution. Since $X = b(100, 0.5)$, then by 15.2 we have $X \approx Y = N(50, 25)$. Now we proceed as

$$P(X \leq 50) \approx P(Y \leq 50) = \Phi\left(\frac{50 - 50}{\sqrt{25}}\right) = \Phi(0) = 0.5$$

Wait a minute. Is this right? Not quite. Indeed, if this was right, then $P(X \geq 51) = 1 - P(X \leq 50) = 0.5$. On the other hand, the above same method applied directly yields

$$P(X \geq 51) \approx P(Y \geq 51) = 1 - \Phi\left(\frac{51 - 50}{\sqrt{25}}\right) = 1 - \Phi(0.2) = 0.4207$$

Of course, $0.5 \neq 0.4207$, 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, it takes only integer values. So, the events $X \leq 50$ and $X \geq 51$ are complementary to each other. The open interval in between, $50 < X < 51$, is irrelevant, its probability is zero. On the contrary, $Y = N(\mu, \sigma^2)$ is continuous, and the probability $P(50 < Y < 51)$ is positive. It is exactly this probability that we overlooked.

Note that the interval $(50, 51)$ is a “border” interval, a gap between the event $X \leq 50$ and its complement $X \geq 51$. To take a proper care of this border interval, we divide it in half and include one half into the event $X \leq 50$ and the other half into the complement $X \geq 51$. In other words, the event and its complement “split up” the border interval.

Now we correct our solution. The proper range for Y is $Y \leq 50.5$. Hence

$$P(X \leq 50) \approx P(Y \leq 50.5) = \Phi\left(\frac{50.5 - 50}{\sqrt{25}}\right) = \Phi(0.1) = 0.5398$$

15.4 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 .

15.5 Example. Toss a coin 100 times and let X be the number of Heads. Find the probability $P(40 \leq X \leq 55)$.

Solution. As in 15.3, $X = b(100, 0.5)$ and $X \approx Y = N(50, 25)$. Now we apply correction for continuity. The event in question is $40 \leq X \leq 55$. The complement is $X \leq 39$ and $X \geq 56$. There are two border intervals: $(39, 40)$ and $(55, 56)$. Then the proper range for Y is $39.5 \leq Y \leq 55.5$, and

$$\begin{aligned} P(40 \leq X \leq 55) &\approx P(39.5 \leq Y \leq 55.5) \\ &= \Phi\left(\frac{55.5 - 50}{\sqrt{25}}\right) - \Phi\left(\frac{39.5 - 50}{\sqrt{25}}\right) \\ &= \Phi(1.1) - \Phi(-2.1) = 0.8643 - 0.0179 = 0.7464 \end{aligned}$$

15.6 Example. Toss a coin 100 times and let X be the number of Heads. Find the probability $P(X = 50)$.

Solution. As before, $X = b(100, 0.5)$ and $X \approx Y = N(50, 25)$. Now we apply correction for continuity. The event in question is $X = 50$ or $50 \leq X \leq$

50. The complement is $X \leq 49$ and $X \geq 51$. There are two border intervals: (49, 50) and (50, 51). Then the proper range for Y is $49.5 \leq Y \leq 50.5$, and

$$\begin{aligned} P(50 \leq X \leq 50) &\approx P(49.5 \leq Y \leq 50.5) \\ &= \Phi\left(\frac{50.5 - 50}{\sqrt{25}}\right) - \Phi\left(\frac{49.5 - 50}{\sqrt{25}}\right) \\ &= \Phi(0.1) - \Phi(-0.1) = 0.5398 - 0.4602 = 0.0796 \end{aligned}$$

This finally answers the question 1.13!

15.7 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 $X = b(12, 0.75)$. By 15.2, $X \approx Y = N(9, 9/4)$. 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

$$P(X \geq 10) \approx P(Y \geq 9.5) = 1 - \Phi\left(\frac{9.5 - 9}{\sqrt{9/4}}\right) = 1 - \Phi(0.33) = 1 - 0.6293 = 0.3707$$

15.8 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 better. We begin with a logical conclusion: 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 the event looks familiar. 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, $X = b(24, 0.75)$. By 15.2, $X \approx Y = N(18, 9/2)$. 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

$$P(X \leq 19) \approx P(Y \leq 19.5) = \Phi\left(\frac{19.5 - 18}{\sqrt{9/2}}\right) = \Phi(0.71) = 0.7611$$

15.9 Remark. De Moivre-Laplace theorem requires n be large. In practice, the theorem works well whenever $n \geq 25$. Even for smaller values of n the theorem is often applied with good results, as we did in Examples 15.7 and 15.8. When n is very small, like $n = 5$ or $n = 10$, it may be actually easier to compute the binomial probabilities directly by (4.1), and get an exact answer.

15.10 Generalizing 15.2. We observed in 10.20 that a binomial random variable $X = b(n, p)$ is the sum of n independent Bernoulli random variables: $X = X_1 + \cdots + X_n$. Therefore, $E(X) = n E(X_1)$ and $\text{Var}(X) = n \text{Var}(X_1)$. Now De Moivre-Laplace theorem can be restated as the fact that $X = X_1 + \cdots + X_n$ is approximately a normal $Y = N(\mu, \sigma^2)$ with $\mu = n E(X_1)$ and $\sigma^2 = n \text{Var}(X_1)$. In this form the theorem can be extended to more general random variables:

15.11 Central limit theorem for S_n . Let X_1, X_2, \dots, X_n be independent identically distributed (i.i.d) random variables. Let $\mu_X = E(X_i)$ be the common mean value of all X_i 's, and $\sigma_X^2 = \text{Var}(X_i)$ the common variance of all X_i 's. For each n let $S_n = X_1 + \cdots + X_n$. Then for large n the variable S_n is approximately normal $Y = N(\mu, \sigma^2)$ with $\mu = E(S_n) = n\mu_X$ and $\sigma^2 = \text{Var}(S_n) = n\sigma_X^2$:

$$S_n \approx N(\mu, \sigma^2) \quad \text{with } \mu = E(S_n) = n\mu_X, \quad \sigma^2 = \text{Var}(S_n) = n\sigma_X^2 \quad (15.1)$$

15.12 Central limit theorem for \bar{X}_n . In the context of 15.11, let $\bar{X}_n = S_n/n = (X_1 + \cdots + X_n)/n$ be the experimental average of X_i 's. Then for large n the variable \bar{X}_n is approximately normal $Y = N(\mu, \sigma^2)$ with $\mu = E(\bar{X}_n) = \mu_X$ and $\sigma^2 = \text{Var}(\bar{X}_n) = \sigma_X^2/n$:

$$\bar{X}_n \approx N(\mu, \sigma^2) \quad \text{with } \mu = E(\bar{X}_n) = \mu_X, \quad \sigma^2 = \text{Var}(\bar{X}_n) = \sigma_X^2/n \quad (15.2)$$

15.13 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 $E(X_i) = 3.5$ by 10.4 and $\text{Var}(X_i) = 2.92$ by 11.4(c). By the central limit theorem 15.11 we have $S_{50} \approx Y = N(175, 146)$. 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

$$\begin{aligned} P(150 \leq X \leq 190) &\approx P(149.5 \leq Y \leq 190.5) \\ &= \Phi\left(\frac{190.5 - 175}{\sqrt{146}}\right) - \Phi\left(\frac{149.5 - 175}{\sqrt{146}}\right) \\ &= \Phi(1.28) - \Phi(-2.11) = 0.8997 - 0.0174 = 0.8823 \end{aligned}$$

15.14 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 the central limit theorem the total weight W is approximately normal $Y = N(100, 100 \times 0.13^2) = N(100, 1.69)$. Since W is (obviously) continuous, we do not apply correction for continuity. Hence

$$P(W > 102) \approx P(Y > 102) = 1 - \Phi\left(\frac{102 - 100}{\sqrt{1.69}}\right) = 1 - \Phi(1.54) = 1 - 0.9382 = 0.0618$$

Note that we do not know the distribution of the weights of individual bolts, that is irrelevant! All we need to know is the mean weight and the standard deviation.

15.15 Remark. The above example illustrates the universality of the central limit theorem. Basically, any random variable that is the sum of many independent small components is approximately normal. Or, put it differently, any experimental value that is the result of a combination of many small factors is approximately normal. This explains why it is customary to assume that experimental data have normal distribution: like the height of an adult person, the lifetime of a person, the weight of a fish caught at random in a pond, the air temperature in weather forecasts, etc. It is not too much of an exaggeration to say: *everything random in nature and sciences has normal distribution, unless some specific constrains apply.*

15.16 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 successes. Estimate $P(|\bar{X} - 0.8| < 0.01)$.

Solution. The total number of successes in 1000 throws is $X = b(1000, 0.8)$, and $\bar{X} = X/1000$. Hence, $E(\bar{X}) = 0.8$ and $\text{Var}(\bar{X}) = 0.16/1000 = 0.00016$. By 15.12, $X \approx Y = N(0.8, 0.00016)$. Then

$$\begin{aligned} P(|X - 0.8| < 0.01) &\approx P(|Y - 0.8| < 0.01) \\ &= \Phi\left(\frac{0.81 - 0.8}{\sqrt{0.00016}}\right) - \Phi\left(\frac{0.79 - 0.8}{\sqrt{0.00016}}\right) \\ &= \Phi(0.79) - \Phi(-0.79) \approx 0.7852 - 0.2148 = 0.5704 \end{aligned}$$

15.17 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 $X_i = \text{exponential}(1/5)$. (Here $\lambda = 1/E(X_i) = 1/5$ by 10.14.) Now we have $\text{Var}(X_i) = 1/\lambda^2 = 25$ according to 12.5. By 15.11 we have $S_{22} \approx Y = N(110, 550)$. Hence,

$$P(S_{22} \leq 100) \approx P(Y \leq 100) = \Phi\left(\frac{100 - 110}{\sqrt{550}}\right) = \Phi(-0.43) = 0.3336$$

Note: it is not so small, 33%. How come? 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, this may be not enough with probability 33%... The reason why this strange result takes place is the unpredictability of exponential random variables we have seen in 12.6. Indeed, some (or several) of those light bulbs can burn down very quickly, and this is not at all unusual, as 12.6 explains. Devices with exponential lifetime are very risky!

15.18 Normal approximation to Poisson. Let $X = \text{poisson}(\lambda)$ with a large λ . Now we want to approximate X by a normal. According to 12.17, X is the sum of many independent Poisson random variables (with smaller parameters). More precisely, $X = X_1 + \cdots + X_n$ where $X_i = \text{poisson}(\lambda/n)$. Hence, by 15.11 we have $X \approx N(\mu, \sigma^2)$ with $\mu = E(X) = \lambda$ and $\sigma^2 = E(X) = \lambda$. So we can write

$$\text{poisson}(\lambda) \approx N(\lambda, \lambda)$$

for large λ . When you use this rule, do not forget about correction for continuity – Poisson random variables are discrete!

15.19 Example. Let S_n be the sum of n independent uniform $U(0,1)$ random variables. Approximate S_n by a normal.

Solution. If $X = U(0,1)$, then $E(X) = 1/2$ and $\text{Var}(X) = 1/12$ (by 11.14b). Hence, $E(S_n) = n/2$ and $\text{Var}(S_n) = n/12$. We thus get

$$S_n \approx N(n/2, n/12)$$

When you use this rule, do not apply correction for continuity (uniform random variables are continuous).

15.20 Generating a normal r.v. by computer. The above example suggests a simple, reliable, and hence widely used in practice method of generating a standard normal random variable $Z = N(0,1)$ by computer. It is convenient to pick $n = 12$, then $S_{12} \approx N(6,1)$, so that $S_{12} - 6 \approx N(0,1)$. 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 $Y = N(\mu, \sigma^2)$, 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.

<pre>FORTRAN: Z=-6.0 DO 1 I=1,12 1 Z=Z+RAND Y=MU+SIGMA*Z</pre>	<pre>C: Z=-6.0 for (i=0;i<12;i++) Z=Z+RAND Y=mu+sigma*Z</pre>	
---	---	--

15.21 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 games. Approximate S_n by a normal.

Solution. First, $S_n = X_1 + \cdots + X_n$, where $X_i = \pm 1$ are gains/losses in individual games. Let X_i be a random variable that takes two values: $+1$ with probability p and -1 with probability $q = 1 - p$. We have

$$E(X_i) = 1 \times p + (-1) \times q = p - q = 2p - 1$$

and

$$\text{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = 1 - (2p - 1)^2 = 4p - 4p^2 = 4pq$$

Hence, $E(S_n) = (p - q)n$ and $\text{Var}(S_n) = 4pqn$. We thus get

$$S_n \approx N((p - q)n, 4pqn)$$

16 Random Walks (Gambler's Ruin)

16.1 Back to Example 15.21. Suppose a gambler plays a game repeatedly, each time either winning \$1 with probability p or losing \$1 with probability $q = 1 - p$. He/she starts with x dollars, and after n games possesses S_n dollars. Approximate S_n by a normal.

Solution. The only difference here from Example 15.21 is the given initial capital of x dollars. Hence, after n games the gambler has

$$S_n = x + X_1 + \cdots + X_n \quad (16.1)$$

dollars. Here $X_i = \pm 1$ are his/her gains/losses in individual games for $i = 1, \dots, n$. As we found in Example 15.21, $E(X_i) = p - q$ and $\text{Var}(X_i) = 4pq$. Hence,

$$E(S_n) = x + (p - q)n \quad \text{and} \quad \text{Var}(S_n) = 4pqn$$

We thus get

$$S_n \approx N(x + (p - q)n, 4pqn) \quad (16.2)$$

16.2 Asymptotics in the symmetric case $p = q$. It is interesting to see how S_n evolves in the distant future, i.e. asymptotically as $n \rightarrow \infty$. We first consider the “fair game” situation, when $p = q = 1/2$. It is a symmetric case, where S_n is equally likely to go either way, up or down. The expression (16.2) takes form

$$S_n \approx N(x, n) \quad (16.3)$$

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, on the average),
- (b) the typical values of S_n are $x \pm \sqrt{n}$, i.e. typical total gains or losses grow as \sqrt{n} .

The typical values of S_n are getting farther and farther away from x (in both directions) as n grows. More precisely, 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 getting farther and longer in time. This is a process called “diffusion” in some physical models. The evolution of S_n resembles the diffusion of a test molecule in a gas.

16.3 Asymptotics in the asymmetric case $p \neq q$. We now consider the “unfair game” situation, when $p \neq q$. The evolution of S_n is described by (16.2). Of course, the cases $p > q$ and $p < q$ are symmetric to each other. Let us 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 customers, the casino is more likely to win than lose.)

Put $\mu = p - q$ (mean gain per game) and $\sigma = \sqrt{4pq}$. Then (16.2) reads

$$S_n \approx N(x + \mu n, \sigma^2 n) \quad (16.4)$$

Now S_n is approximately normal with a growing mean value ($= x + \mu n$) and a growing standard deviation $\sigma\sqrt{n}$. That is, $S_n \approx x + \mu n \pm \sigma\sqrt{n}$. This means that S_n grows on the average, and its typical deviations from the average value grow, 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, right?

From the point of view of the casino owners, these are two competing processes: one (growth on the average) is good, it makes their business stronger. The other (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: the steady growth on the average or random fluctuations?

The 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 . This will be illustrated in 16.10 and 16.13.

Employing once again the physical model of a test molecule in a gas, we have now a diffusion combined with a “drift” in a given (positive) direction with a constant speed $\mu > 0$. Such a diffusion occurs when the test molecule is placed in a moving gas (under wind): it drifts with the gas and at the same time wanders about randomly among other molecules.

16.4 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 out 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 consider a random walk on a 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 space: a drunk person (no, let's say just a moving point) moves forward or backward, left or right, up or down, with the same probability $1/6$. This model describes, quite accurately, the motion of a test molecule in a real 3D gas.

16.4 Restricted random walks. Back to 16.1-16.3. In practice, random walks are often restricted. A casino owner cannot afford to have his capital S_n drop below zero - if this happens he will have to close down his casino. The same is true for a gambler - if he/she loses the entire initial capital ($S_n = 0$), then it is all over. Most gamblers (at least the wiser ones) set an upper limit, too, - they decide in advance to stop playing 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 levels: 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 (or absorbing). If just one bound is set (lower or upper), then we have a one-sided restricted random walk.

16.5 Parameters of restricted random walks. For a restricted random walk, the normal approximation (16.2) does not fully apply. In fact, if the random walk has two restrictions, $a \leq S_n \leq b$, 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 $T \geq 1$ and its value S_T is final, it will never change again. Such a walk can be then characterized by three parameters: the probabilities $P_a = P(S_T = a)$ and $P_b = P(S_T = b)$ of stopping at a and b , respectively, and the mean lifetime $E(T)$. 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, $S_n \geq a$, then it may either stop at a at a random time T with probability P_a , or evolve indefinitely (without going down to a). Again, the mean lifetime $E(T)$ is a relevant parameter.

16.6 Wald's identity. According to (16.1), at the stopping time $n = T$ we have $S_T = x + X_1 + \dots + X_T$. Recall also that $E(X) = p - q = \mu$. The

following equation then looks natural (but we omit the argument):

$$E(S_T) = x + \mu E(T) \quad (16.5)$$

It is known as Wald's identity. In the case of two restrictions, S_T only takes values a and b , hence $E(S_T) = aP_a + bP_b$, so Wald's identity reads

$$aP_a + bP_b = x + \mu E(T) \quad (16.6)$$

This is a very helpful relation.

16.7 Two sided restrictions: symmetric case. In the case $p = q = 1/2$ we have $\mu = 0$, so the Wald's identity (16.6) 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

$$P_a = \frac{b-x}{b-a} \quad \text{and} \quad P_b = \frac{x-a}{b-a}$$

The mean lifetime $E(T)$ can be found from another Wald's identity, $\text{Var}(S_T) = \sigma^2 E(T)$. From this, it is easy to find

$$E(T) = (x-a)(b-x)$$

16.8 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 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 to win 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 $E(T) = (10-0)(50-10) = 400$. 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 $P_{\text{win}} = 0.2$.

16.9 Two sided restrictions: asymmetric case. In the case $p \neq q$ we have $\mu \neq 0$, so the Wald's identity (16.6) gives

$$E(T) = \frac{aP_a + bP_b - x}{p - q} \quad (16.7)$$

It remains to find P_a and P_b , for which special formulas exist:

$$P_a = \frac{(q/p)^{x-a} - (q/p)^{b-a}}{1 - (q/p)^{b-a}} \quad \text{and} \quad P_b = \frac{1 - (q/p)^{x-a}}{1 - (q/p)^{b-a}}$$

16.10 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, wait a minute. As in 16.8, we have $x = 10$, $a = 0$ and $b = 50$. Then the probability to win is

$$P_{50} = \frac{1 - (10/9)^{10}}{1 - (10/9)^{50}} = 0.0097$$

and the mean number of games is

$$E(T) = \frac{0 \cdot (1 - 0.0097) + 50 \cdot 0.0097 - 10}{-1/19} = 180.8$$

Now compare these results to those in Example 16.8. The chances to win dropped from 20% to below 1%! The game only seemed 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 16.8, 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 Example 16.8. The moral of this story is this: if you have to risk in an unfavorable situation, risk “big”. The longer you play (trying to achieve your goal in small steps), the more chances against you accumulate, and you will almost certainly lose.

16.11 One sided restrictions: symmetric case. We consider a one-sided restriction $S_n \geq a$ (a lower bound is set, no upper bound). The other case, when only the 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 $E(T)$ in 16.7 and take the limit as $b \rightarrow \infty$. We obtain

$$P_a = 1 \quad \text{and} \quad E(T) = \infty$$

This means that in a fair game with one restriction, the random walk hits the bound and stops, sooner or later. This is consistent, by the way, with the diffusive character of the random walk observed in 16.3: 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.

A surprise comes with the formula $E(T) = \infty$. This means that in practice, it takes arbitrary long (better to 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.

16.12 One sided restrictions: asymmetric case. Consider again a one-sided restriction $S_n \geq a$ (a lower bound is set, no upper bound). In the case $p \neq q$ (an asymmetric walk) we use the formulas for P_a and $E(T)$ in 16.9 and take the limit as $b \rightarrow \infty$. There are two distinct cases here.

Assuming $p < q$ we obtain

$$P_a = 1 \quad \text{and} \quad E(T) = \frac{x - a}{q - p}$$

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

$$P_a = (q/p)^{x-a} \quad \text{and} \quad E(T) = \infty$$

Now the random walk does NOT have to hit the lower bound at all! 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 in 16.3: 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 drive up and never come close to a again.

16.13 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 than 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.

16.14 Hit probabilities in a symmetric random walk. Consider a symmetric random walk starting out at x with 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 in 16.7 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 , and 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 in 16.7 we have

$$P(x, y) = \frac{b - x}{b - y}$$

Example 16.8 continued. 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

$$P(10, 40) = \frac{10 - 0}{40 - 0} = 0.25$$

16.15 Return probabilities in a symmetric random walk. Consider a symmetric random walk starting out at x with two-sided restrictions $a \leq S_n \leq b$. 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 16.8 back again. 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 have exactly \$10 again?

Solution. We have

$$P(10, 10) = 1 - \frac{50 - 0}{2(50 - 10)(10 - 0)} = \frac{15}{16}$$

16.16 Number of returns in a symmetric random walk. Consider a symmetric random walk starting out at x with two-sided restrictions $a \leq S_n \leq b$. 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 $P(x, x)$ found in 16.15. 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 16.8 once again. 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 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% (Example 16.8). 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...

16.17 Returns in a symmetric random walk without restrictions.

Consider a symmetric random walk starting out 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 $a \rightarrow -\infty$ and $b \rightarrow \infty$ in the expressions obtained in 16.15 and 16.16. We get

$$P(x, x) = 1 \quad \text{and} \quad G(x, x) = \infty$$

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 in 16.3: the deviations from x must always go both ways, left and right. In order to go from left to right or vice versa the walk has to cross the point x .

16.18 Returns in a symmetric random walk without restrictions in 2D and 3D.

In 16.4 we mentioned 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 linear case in 16.17, the walk does not have to cross x in order to change the direction of deviation: it can come back close to x , go around x , and then evolve in another direction, for example. So, it is not quite clear whether the symmetric random walk in 2D or 3D is recurrent. You can make your best guess.

The answer is that the 2D walk (on the plane) is still recurrent, the walk comes back with probability one infinitely many times. But the 3D walk (in the space) is not recurrent anymore. In 3D, the probability to come back to the original point x is less than one (about 60%) and the average number of returns is finite (less than 3).

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 nonrecurrence of 3D random walks. Maybe this has something to do with the physics of gases and fluids...

17 Poisson Process

17.1 Examples. In 4.17 we discussed a telephone operator that received about 5 calls per hour, on the average. The incoming calls arrive at random times, and the intervals between successive calls are random, too. In 12.6 we discussed state trooper patrol cars located on a long stretch of a highway approximately one per 20 miles. This means that patrol cars can be located anywhere, randomly, and “one car per 20 miles” just refers to an average interval between patrol cars. The actual intervals vary, they are random.

These two examples have one feature in common: some events (points) appear randomly on a line (the line can be a time axis). The location of points is random, the intervals between points are random, too, and even the number of points on any given segment of the line is random. All we know is just the average rate at which those points appear, per unit length.

17.2 More examples. Many other practical situations fit that model well. Customers arriving at a supermarket or a car repair shop make a random sequence of arrival times on the time axis. Failures in a long cable line can appear randomly at any point, and the number of failures is random, too. Accidents on a long highway is another example.

17.3 Poisson process. Any random sequence (collection) of points on a line with the above properties is called a Poisson process. We emphasize the main features: points appear anywhere randomly and independently of each other, all we know is the rate, or the average density of points per unit length, we call it λ . The random points are often called the *events* of the Poisson process.

17.4 Number of points on a given segment. Let (a, b) be a given interval (segment) on the line, and denote by $N_{(a,b)}$ the number of random points (events) in this segment. We have seen in Chapter 4 (see 4.17, for example), that $N_{(a,b)}$ is a Poisson random variable with parameter $\lambda(b - a)$. Remember this:

$$N_{(a,b)} = \text{poisson}(\lambda(b - a))$$

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.

17.5 Waiting times (inter-arrival times). Intervals between successive points in a Poisson process are called waiting times or inter-arrival times. If $0 < P_1 < P_2 < \dots$ are the successive points (events), then $W_1 = P_1$, $W_2 = P_2 - P_1, \dots$ 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”.

Let us find the distribution function of a waiting time $W_k = P_k - P_{k-1}$. We have $F_{W_k}(x) = P(W_k \leq x) = 1 - P(W_k > x)$. Now the event $P_k > x$ precisely means that there are no points of the Poisson process on the interval $I = (P_{k-1}, P_{k-1} + x)$. The length of this interval is x , and the number of points in it is a Poisson random variable N_I , according to 17.4. Therefore,

$$P(N_I = 0) = P(\text{poisson}(\lambda x) = 0) = \frac{(\lambda x)^0}{0!} e^{-\lambda x} = e^{-\lambda x}$$

We obtain

$$F_W(x) = 1 - P(W_k > x) = 1 - P(N_I = 0) = 1 - e^{-\lambda x}$$

Thus, 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, $W_k = \text{exponential}(\lambda)$. Also, the waiting times W_1, W_2, \dots are independent.

17.6 Remark. Note: the parameter λ of the Poisson process (the average density of events) serves as the parameter for both the Poisson random variables involved in 17.4 and the exponential random variables involved in 17.5. This explains why we have denoted the parameters of these two types of random variables by one symbol: λ (this fact was noted in 6.12).

17.7 Remark. Let $(a, a + T)$ be a long interval in the Poisson process. The number of events on this interval is a Poisson random variable with parameter λT . Its mean value is λT . So, we expect, on the average, λT events 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.

17.8 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. By 17.5, the waiting times are independent of each other, so it does not matter how many accidents you passed. The probability that the interval to the next accident (“the waiting time”) is longer than 40 miles is

$$P(W > 40) = 1 - F_W(40) = 1 - (1 - e^{-\lambda \cdot 40}) = e^{-2}$$

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 17.8 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,

$$P(N > 1) = 1 - P(N = 0) - P(N = 1) = 1 - e^{-0.5} - 0.5e^{-0.5} = 0.09$$

17.9 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. How come?

17.10 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.

17.11 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 (events) 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 (events) 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.

17.12 Number of points in a 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 (events) of the process in this region. Just as in 17.4, now 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.

17.13 Problem. 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 $\text{Cov}(N_{R_1}, N_{R_2})$.

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,

$$\begin{aligned}\text{Cov}(N_{R_1}, N_{R_2}) &= \text{Cov}(N_{D_0} + N_{D_1}, N_{D_0} + N_{D_2}) \\ &= \text{Cov}(N_{D_0}, N_{D_0}) + 0 + 0 + 0 = \text{Var}(N_{D_0}) = \lambda|D_0|\end{aligned}$$

In the last step we remembered that the variance of a Poisson random variable with parameter λ was equal to λ .

17.14 Problem. 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 event in the process. Find the distribution function of the random variable X .

Solution. We have

$$F_X(x) = P(X \leq x) = 1 - P(X > x)$$

The condition $X > x$ means that the nearest point (even) 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,

$$F_X(x) = 1 - P(N_{B_x} = 0) = 1 - e^{-\lambda|B_x|} = 1 - e^{-\frac{4}{3}\lambda\pi x^3}$$

In the last step we used the fact from elementary geometry that the volume of a ball of radius x was $\frac{4}{3}\pi x^3$.