Chapter 1

Combinatorial Analysis

1.1 Introduction

Many problems in probability theory can be solved simply by counting the number of different ways that a certain event can occur. Effective methods for counting would then be useful in our study of probability. The mathematical theory of counting is formally known as *combinatorial analysis*.

1.2 Basic Principle of Counting

The basic principle of counting is a simple but useful result for all our work.

(**The Basic Principle of Counting**)
Suppose that two experiments are to be performed. If

- experiment 1 can result in any one of m possible outcomes; and
- experiment 2 can result in any one of n possible outcomes;

then together there are mn possible outcomes of the two experiments.

Proof. Make a list:

$$\begin{array}{cccc} (1,1) & (1,2) & \dots & (1,\, n) \\ (2,1) & (2,2) & \dots & (2,\, n) \\ \vdots & \vdots & \vdots & \vdots \\ (m,1) & (m,2) & \dots & (m,\, n) \end{array}$$

□

Example 1.1. A small community consists of 10 women, each of whom has 3 children. If one woman and one of her children are to be chosen as mother and child of the year, how many different choices are possible?

Solution:
By regarding the choice of the woman as the outcome of the first experiment and the subsequent choice of one of her children as the outcome of the second experiment, we see from the basic principle that there are $10 \times 3 = 30$ possible choices.

Example 1.2. In a class of 40 students, we choose a president and a vice president. There are

$$40 \times 39 = 1560$$

possible choices.

For more than two experiments, we have

(**The Generalized Basic Principle of Counting**)
Suppose that r experiments are to be performed. If

- experiment 1 results in n_1 possible outcomes;
- experiment 2 results in n_2 possible outcomes;
- $\cdots$
- experiment r results in n_r possible outcomes;

then together there are $n_1 n_2 \cdots n_r$ possible outcomes of the r experiments.

Example 1.3. How many different 7-place license plates are possible if the first 3 places for letters and the final 4 places by numbers?

Solution:
First 3 places each has 26 ways, and final 4 places each has 10 ways. Therefore, the total possible number of ways is

$$26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 175,760,000.$$

Example 1.4. How many functions defined on a domain with n elements are possible if each functional value is either 0 or 1?

Solution:
Let's label the elements of the domain: $1, 2, \ldots, n$. Since $f(i)$ is either 0 or 1 for each $i = 1, 2, \ldots, n$, it follows that there are

$$2 \times 2 \times \cdots \times 2 = 2^n$$

possible functions.

Example 1.5. In Example 1.3, how many license plates would be possible if repetition among letters or numbers were prohibited?

Solution:
In this case, there would be $26 \times 25 \times 24 \times 10 \times 9 \times 8 \times 7 = 78,624,000$ possible license plates.

1.3 Permutations

How many different arrangements of the letters a, b and c are possible?

Solution:
Direct enumeration: 6 possible arrangements. They are

$$abc, acb, bac, bca, cab, \text{ and } cba.$$

Why 6?

$$6 = 3 \times 2 \times 1.$$

(**General Principle**)
Suppose there are n (distinct) objects, then the total number of different arrangements is

$$n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1 = n!$$

with the convention that

$$0! = 1.$$

Example 1.6. Seating arrangement in a row: 9 people sitting in a row. There are $9! = 362,880$ ways.

Example 1.7. 6 men and 4 women in a class. Students ranked according to test result (assume no students obtained the same score).

(a) How many different rankings are possible?

Solution:

$$10! = 3,628,800.$$

(b) Ranking men among themselves, ranking women among themselves.

Solution:
There will be

$$6! \times 4! = 720 \cdot 24 = 17280$$

different rankings.

We shall now determine the number of permutations of a set of n objects when some of them are indistinguishable from one another. Consider the following example.

Example 1.8. How many different letter arrangements can be formed using the letters M I S S?

Solution:
Trick: Make the S's distinct, call them S_1 and S_2.

$$\begin{array}{cc} MIS_1S_2 & MIS_2S_1 \\ MS_1IS_2 & MS_2IS_1 \\ MS_1S_2I & MS_2S_1I \\ \cdots & \cdots \end{array}$$

If they were all different, there would be 4! ways.
We double counted by 2! ways, so the actual number of different arrangements is

$$\frac{4!}{2!} = 12.$$

(**General Principle**)
For n objects of which n_1 are alike, n_2 are alike, ..., n_r are alike, there are

$$\frac{n!}{n_1!n_2!\cdots n_r!}$$

different permutations of the n objects.

Example 1.9. How many ways to rearrange Mississippi?

Solution:
11 letters of which 1 M, 4 I's, 4 S's and 2 P's.
There are

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

Example 1.10 (Seating in circle). 10 people sitting around a round dining table. It is the relative positions that really matters – who is on your left, on your right. No. of seating arrangements is

$$\frac{10!}{10} = 9!.$$

Generally, for n people sitting in a circle, there are

$$\frac{n!}{n} = (n-1)!$$

possible arrangements.

Example 1.11 (Making necklaces). n different pearls string in a necklace. Number of ways of stringing the pearls is

$$\frac{(n-1)!}{2}.$$

1.4 Combinations

In how many ways can we choose 3 items from 5 items: A, B, C, D and E?

Solution:
5 ways to choose first item,
4 ways to choose second item, and
3 ways to choose third item
So the number of ways (in this order) is $5 \cdot 4 \cdot 3$.
However,

$$ABC, ACB, BAC, BCA, CAB \text{ and } CBA$$

will be considered as the same group.
So the number of different groups (order not important) is

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1}.$$

Generally, if there are n distinct objects, of which we choose a group of r items,

$$
\begin{aligned}
&\text{Number of possible groups} \\
&= \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} \\
&= \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} \\
&\qquad \times \frac{(n-r)(n-r-1)\cdots 3\cdot 2\cdot 1}{(n-r)(n-r-1)\cdots 3\cdot 2\cdot 1} \\
&= \frac{n!}{r!(n-r)!}.
\end{aligned}
$$

Remark 1. (i) **Notation:**
Number of ways of choosing r items from n items: ${}_nC_r$ or $\binom{n}{r}$.

(ii) For $r = 0, 1, \ldots, n$,

$$\binom{n}{r} = \binom{n}{n-r}.$$

(iii)

$$\binom{n}{0} = \binom{n}{n} = 1.$$

(iv) **Convention:**
When n is a nonnegative integer, and $r < 0$ or $r > n$, take

$$\binom{n}{r} = 0.$$

Example 1.12. A committee of 3 is to be formed from a group of 20 people.

1. How many possible committees can be formed?
2. Suppose further that, two guys: Peter and Paul refuse to serve in the same committee. How many possible committees can be formed with the restriction that these two guys don't serve together?

Solution:

1. No. of different committees that can be formed = $\binom{20}{3} = 1140$.

2. Two possible cases:

   Case 1. Both of them are not in the committee.
   Ways to do that $= \binom{18}{3} = 816$.

   Case 2. One of them in.
   Ways to form $= \binom{2}{1}\binom{18}{2} = 306$.

   Total $= 816 + 306 = 1122$.

   Alternative solution: (sketch)

$$\binom{20}{3} - \binom{2}{2}\binom{18}{1} = 1140 - 18 = 1122.$$

Example 1.13. Consider a set of n antennas of which m are defective and $n-m$ are functional and assume that all of the defectives and all of the functionals are considered indistinguishable. How many linear orderings are there in which no two defectives are consecutive?

Solution:
Line up the $n-m$ functional antennas among themselves. In order that no two defectives are consecutive, we insert each defective antenna into one of the $n-m+1$ spaces between the functional antennas. Hence, there are $\binom{n-m+1}{m}$ possible orderings in which there is at least one functional antenna between any two defective ones.

Some Useful Combinatorial Identities

For $1 \leq r \leq n$,

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

Proof.
Algebraic:

$$\begin{aligned} \text{RHS} &= \frac{(n-1)!}{(r-1)!(n-r)!} + \frac{(n-1)!}{r!(n-r-1)!} \\ &= \frac{(n-1)!}{r!(n-r)!}[r+(n-r)] = \frac{n!}{r!(n-r)!}. \end{aligned}$$

Combinatorial:
Consider the cases where the first object (i) is chosen, (ii) not chosen:

$$\binom{1}{1}\cdot\binom{n-1}{r-1}+\binom{1}{0}\cdot\binom{n-1}{r}.$$

□

(**The Binomial Theorem**)
Let n be a nonnegative integer, then

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

In view of the binomial theorem, $\binom{n}{k}$ is often referred to as the binomial coefficient.

Example 1.14. How many subsets are there of a set consisting of n elements?

Solution:
Since there are $\binom{n}{k}$ subsets of size k, the desired answer is

$$\sum_{k=0}^{n}\binom{n}{k}=(1+1)^n=2^n.$$

This result could also have been obtained by considering whether each element in the set is being chosen or not (to be part of a subset). As there are 2^n possible assignments, the result follows.

Example 1.15.

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

Solution:
Let $x=-1$, $y=1$ in the binomial theorem.

Example 1.16.

$$\binom{n}{0}+\binom{n}{2}+\binom{n}{4}+\cdots=\binom{n}{1}+\binom{n}{3}+\binom{n}{5}+\cdots.$$

Solution:
Start with the previous example. Move the negative terms to the right hand side of the equation.

1.5 Multinomial Coefficients

A set of n distinct items is to be divided into r distinct groups of respective sizes $n_1, n_2, \ldots, n_r$, where $\sum_{i=1}^{r} n_i = n$. How many different divisions are possible?

Solution:
Note that there are $\binom{n}{n_1}$ possible choices for the first group; for each choice of the first group, there are $\binom{n-n_1}{n_2}$ possible choices for the second group; for each choice of the first two groups, there are $\binom{n-n_1-n_2}{n_3}$ possible choices for the third group; and so on. It then follows from the generalized version of the basic counting principle that there are

$$\begin{aligned}
&\binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \cdots \times \binom{n-n_1-n_2-\cdots-n_{r-1}}{n_r} \\
&= \frac{n!}{(n-n_1)!n_1!} \times \frac{(n-n_1)!}{(n-n_1-n_2)!n_2!} \times \cdots \times \frac{(n-n_1-n_2-\cdots-n_{r-1})!}{0!n_r!} \\
&= \frac{n!}{n_1!n_2!\cdots n_r!}
\end{aligned}$$

possible divisions.

(**Notation**)
If $n_1 + n_2 + \cdots + n_r = n$, we define $\binom{n}{n_1, n_2, \ldots, n_r}$ by

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!n_2!\cdots n_r!}.$$

Thus $\binom{n}{n_1, n_2, \ldots, n_r}$ represents the number of possible divisions of n distinct objects into r distinct groups of respective sizes $n_1, n_2, \ldots, n_r$.

Example 1.17. A police department in a small city consists of 10 officers. If the department policy is to have 5 of the officers patrolling the streets, 2 of the officers working full time at the station, and 3 of the officers on reserve at the station, how many different divisions of the 10 officers into the 3 groups are possible?

Solution:
There are $\frac{10!}{5!2!3!} = 2520$ possible divisions.

Example 1.18. Ten children are to be divided into an A team and a B team of 5 each. The A team will play in one league and the B team in another. How many different divisions are possible?

Solution:
There are $\frac{10!}{5!5!} = 252$ possible divisions.

Example 1.19. In order to play a game of basketball, 10 children at a playground divide themselves into two teams of 5 each. How many different divisions are possible?

Solution:
Note that this example is different from the previous one because now the order of the two teams is irrelevant. That is, there is no A and B team, but just a division consisting of 2 groups of 5 each. Hence, the desired answer is

$$\frac{10!/(5!5!)}{2!} = 126.$$

(**The Multinomial Theorem**)

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{(n_1,\ldots,n_r):n_1+\cdots+n_r=n} \binom{n}{n_1, n_2, \ldots, n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}$$

Note that the sum is over all nonnegative integer-valued vectors $(n_1, n_2, \ldots, n_r)$ such that $n_1 + n_2 + \cdots + n_r = n$.

Example 1.20.

$$\begin{aligned}(x_1 + x_2 + x_3)^2 &= \binom{2}{2,0,0} x_1^2 x_2^0 x_3^0 + \binom{2}{0,2,0} x_1^0 x_2^2 x_3^0 + \binom{2}{0,0,2} x_1^0 x_2^0 x_3^2 \\ &\quad + \binom{2}{1,1,0} x_1^1 x_2^1 x_3^0 + \binom{2}{1,0,1} x_1^1 x_2^0 x_3^1 + \binom{2}{0,1,1} x_1^0 x_2^1 x_3^1 \\ &= x_1^2 + x_2^2 + x_3^2 + 2x_1\varkappa_2 + 2x_1x_3 + 2x_2x_3.\end{aligned}$$

Remark 2. The preceding result shows that there is a one-to-one correspondence between the set of the possible tournament results and the set of permutations of $1, \ldots, n$. For more details on how such a correspondence can be constructed, refer to [Ross].

1.6 The Number Of Integer Solutions Of Equations

Proposition 1.21. *There are* $\binom{n-1}{r-1}$ *distinct positive integer-valued vectors* $(x_1, x_2, \ldots, x_r)$ *that satisfies the equation*

$$x_1 + x_2 + \cdots + x_r = n,$$

where $x_i > 0$ *for* $i = 1, \ldots, r$.

Proof. Consider n distinct objects. We want to divide them into r nonempty groups. To do so, we can select $r-1$ of the $n-1$ spaces between the n objects as our dividing points.

$$\mathrm{X} \wedge \mathrm{X} \wedge \mathrm{X} \wedge \cdots \wedge \mathrm{X} \wedge \mathrm{X} \wedge \mathrm{X}$$

In the diagram there are $n-1$ spaces in between the n objects represented by X. We choose $r-1$ of them to divide the n objects into r nonempty groups.

So there are $\binom{n-1}{r-1}$ distinct positive integer-valued vectors $(x_1, x_2, \ldots, x_r)$ that satisfies the equation. □

Proposition 1.22. *There are* $\binom{n+r-1}{r-1}$ *distinct non-negative integer-valued vectors* $(x_1, x_2, \ldots, x_r)$ *that satisfies the equation*

$$x_1 + x_2 + \cdots + x_r = n,$$

where $x_i \geq 0$ *for* $i = 1, \ldots, r$.

Proof. Let $y_i = x_i + 1$, then $y_i > 0$ and the number of non-negative solutions of

$$x_1 + x_2 + \cdots + x_r = n$$

is the same as the number of positive solutions of

$$(y_1 - 1) + (y_2 - 1) + \cdots + (y_r - 1) = n$$

i.e.,

$$y_1 + y_2 + \cdots + y_r = n + r,$$

which is $\binom{n+r-1}{r-1}$. □

Example 1.23. How many distinct nonnegative integer-valued solutions of $x_1 + x_2 = 3$ are possible?

Solution:
There are $\binom{3+2-1}{2-1} = 4$ such solutions: $(0,3),(1,2),(2,1),(3,0)$.

Example 1.24. An investor has 20 thousand dollars to invest among 4 possible investments. Each investment must be in units of a thousand dollars. If the total 20 thousand is to be invested, how many different investment strategies are possible? What if not all the money need be invested?

Solution:
If we let $x_i, i = 1,2,3,4$, denote the number of thousands invested in investment i, then, when all is to be invested, x_1, x_2, x_3, x_4 are integers satisfying equation

$$x_1 + x_2 + x_3 + x_4 = 20, \quad x_i \geq 0$$

Hence, by Proposition 1.22, there are $\binom{23}{3} = 1771$ possible investment strategies. If not all of the money need be invested, then if we let x_5 denote the amount kept in reserve, a strategy is a nonnegative integer-valued vector x_1, x_2, x_3, x_4, x_5 satisfying the equation

$$x_1 + x_2 + x_3 + x_4 + x_5 = 20$$

Hence, by Proposition 1.22, there are now $\binom{24}{4} = 10,626$ possible strategies.

Example 1.25. How many terms are there in the multinomial expansion of $(x_1 + x_2 + \cdots + x_r)^n$?

Solution:

$$(x_1 + x_2 + \cdots + x_r)^n = \sum \binom{n}{n_1, n_2, \ldots, n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}$$

where the sum is over all nonnegative integer-valued $(n_1, \ldots, n_r)$ such that $n_1 + \cdots + n_r = n$. Hence, by Proposition 1.22, there are $\binom{n+r-1}{r-1}$ such terms.

Example 1.26. Let us consider again Example 1.13. Imagine that the defective items are lined up among themselves and the functional ones are now to be put in position. Let us denote x_1 as the number of functional items to the left of the first defective, x_2 as the number of functional items between the first two defectives, and so on. That is, schematically, we have

$$x_1 \quad 0 \quad x_2 \quad 0 \quad \cdots \quad 0 \quad x_m \quad 0 \quad x_{m+1}$$

Now, there will be at least one functional item between any pair of defectives as long as $x_i > 0$, $i = 2, \ldots, m$. Hence, the number of outcomes satisfying the condition is the number of vectors $x_1, \ldots, x_{m+1}$ that satisfy the equation

$$x_1 + \cdots + x_{m+1} = n - m, \quad x_1 \geq 0, \quad x_{m+1} \geq 0, \quad x_i > 0, \quad i = 2, \ldots, m.$$

Let $y_1 = x_1 + 1$, $y_i = x_i$ for $i = 2, \ldots, m$ and $y_{m+1} = x_{m+1} + 1$, we see that this number is equal to the number of positive vectors $(y_1, \ldots, y_{m+1})$ that satisfy the equation

$$y_1 + y_2 + \cdots + y_{m+1} = n - m + 2.$$

By Proposition 1.21, there are $\binom{n-m+1}{m}$ such outcomes.

Suppose now that we are interested in the number of outcomes in which each pair of defective items is separated by at least 2 functional items. By the same reasoning as that applied previously, this would equal the number of vectors satisfying the equation

$$x_1 + \cdots + x_{m+1} = n - m, \quad x_1 \geq 0, \quad x_{m+1} \geq 0, \quad x_i \geq 2, \quad i = 2, \ldots, m.$$

Let $y_1 = x_1 + 1$, $y_i = x_i - 1$ for $i = 2, \ldots, m$ and $y_{m+1} = x_{m+1} + 1$, we see that this is the same as the number of positive solutions of the equation

$$y_1 + y_2 + \cdots + y_{m+1} = n - 2m + 3.$$

By Proposition 1.21, there are $\binom{n-2m+2}{m}$ such outcomes.

Chapter 2

Axioms of Probability

2.1 Introduction

In this chapter we introduce the basic terminology of probability theory: experiment, outcomes, sample space, events. Next we define what the probability of an event is and proceed to show how it is computed in a variety of examples.

2.2 Sample Space and Events

The basic object of probability is an **experiment**: an activity or procedure that produces distinct, well-defined possibilities called **outcomes**. The **sample space** is the set of all possible outcomes of an experiment, usually denoted by S.

Example 2.1. The sample space of tossing a coin:

$$S = \{ \text{ head, tail } \}.$$

Example 2.2. The sample space of positions of a 7-horse race:

$$S = \{\text{all permutations of } \{1,2,3,4,5,6,7\}\}.$$

Example 2.3. Tossing two dice:

$$\begin{aligned} S &= \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\} \\ &= \{(i,j) : 1 \leq i,j \leq 6\}. \end{aligned}$$

Example 2.4. The lifetime of a transistor:

$$S = [0,\infty).$$

Any subset E of the sample space is an **event**.

Example 2.5.

(a) For Example 2.1, $E = \{$ head $\}$ is an possible event.

(b) For Example 2.2, $E = \{$ horse 2 comes out first$\}$ is an possible event. Here E has 6! elements.

(c) For Example 2.3,

$$\begin{aligned} E &= \{\text{ sum of 2 dice is } 7\} \\ &= \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \end{aligned}$$

is an possible event.

(d) For Example 2.4, $E = \{x : 0 \leq x \leq 5\}$ is an possible event.

2.3 Axioms of Probability

Abstractly, a **probability** is a function that assigns numbers to events, and satisfies the following

(Axioms of Probability:)
Probability, denoted by P, is a function on the collection of events satisfying

(i) For any event A,

$$0 \leq P(A) \leq 1.$$

(ii) Let S be the sample space, then

$$P(S) = 1.$$

(iii) For any sequence of mutually exclusive events $A_1, A_2, \ldots$ (that is, $A_i A_j = \emptyset$ when $i \neq j$),

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

These are also known as the **Kolmogorov axioms**, named after Andrey Kolmogorov.

Remark 3. We call $P(A)$ the **probability** of the event A.
For (iii), RHS should naturally give the probability of the event $\bigcup_{i=1}^{\infty} A_i$. LHS is the assigned value of $\bigcup_{i=1}^{\infty} A_i$.

2.4 Properties of Probability

Using only the axioms, we now establish a few very useful propositions.

Proposition 2.6. $P(\emptyset) = 0$.

Proof. Take $A_1 = S$, $A_2 = A_3 = \cdots = \emptyset$, then $A_1, A_2, \ldots$ are mutually exclusive(why?). Axiom (iii) says that

$$P(S) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(S) + \sum_{i=2}^{\infty} P(\emptyset),$$

which implies that $P(\emptyset) = 0$. □

Proposition 2.7. *For any finite sequence of mutually exclusive events $A_1, A_2, \ldots, A_n$,*

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

Proof. Let $A_{n+1} = A_{n+2} = \cdots = \emptyset$, then the sequence $A_1, A_2 \ldots$ is still mutually exclusive, and hence

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= P\left(\bigcup_{i=1}^{\infty} A_i\right) \\ &= \sum_{i=1}^{\infty} P(A_i) \\ &= \sum_{i=1}^{n} P(A_i) \quad \text{since } P(A_i) = 0, i \geq n+1. \end{aligned}$$

□

Proposition 2.8. *Let A be an event, then*

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

Proof. Since $S = A \cup A^c$, it follows that

$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c).$$

Rearranging the terms, we proved the result. □

Proposition 2.9. *If $A \subset B$, then*

$$P(A) + P(BA^c) = P(B).$$

Hence, $P(A) \leq P(B)$. (Bigger event, bigger probability.)

Proof. Write $B = A \cup BA^c$, then

$$P(B) = P(A) + P(BA^c).$$

□

Example 2.10. Linda is 31 years old, single, outspoken, and very bright. She majored in Philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
Which of the following alternatives is more probable? And why?

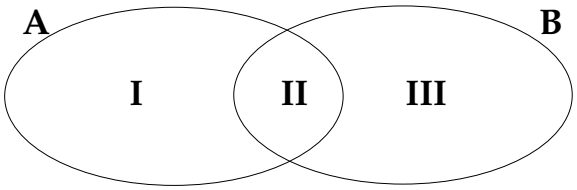
(a) Linda is a bank teller.

(b) Linda is a bank teller and active in feminist movement.

Proposition 2.11. *Let A and B be any two events, then*

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

Proof. Decompose $A \cup B$ into 3 mutually exclusive regions as shown:



then

$$\begin{aligned} P(A) &= P(I) + P(II) \\ P(B) &= P(II) + P(III) \\ P(AB) &= P(II) \\ P(A \cup B) &= P(I) + P(II) + P(III). \end{aligned}$$

Eliminating $P(I), P(II)$ and $P(III)$ from above, we get

$$P(A \cup B) = P(A) + P(B) - P(II) = P(A) + P(B) - P(AB).$$

□

Example 2.12. J is taking two books along on her holiday vacation. With probability 0.5, she will like the first book; with probability 0.4, she will like the second book; and with probability 0.3, she will like both books. What is the probability that she likes neither book?

Solution:
Let B_i denote the event that J likes book i, where $i = 1,2$. Then the probability that she likes at least one of the books is

$$P(B_1 \cup B_2) = P(B_1) + P(B_2) - P(B_1B_2) = 0.5 + 0.4 - 0.3 = 0.6.$$

The event that she likes neither book is represented by $B_1^C B_2^C$, and its probability is given as

$$P\left(B_1^C B_2^C\right) = P\left((B_1 \cup B_2)^C\right) = 1 - P(B_1 \cup B_2) = 0.4.$$

Proposition 2.13 (Inclusion-Exclusion Principle)**.** *Let $A_1, A_2, \ldots, A_n$ be any events, then*

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) &= \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i_1 < i_2 \le n} P(A_{i_1}A_{i_2}) + \cdots \\ &\quad + (-1)^{r+1} \sum_{1 \le i_1 < \cdots < i_r \le n} P(A_{i_1} \cdots A_{i_r}) \\ &\quad + \cdots + (-1)^{n+1} P(A_1 \cdots A_n). \end{aligned}$$

Example 2.14. Explanation of the summations. Suppose $n = 4$,

$\sum_{1 \le i_1 < i_2 \le 4} P(A_{i_1}A_{i_2})$ means

$$\begin{aligned} &P(A_1A_2) + P(A_1A_3) + P(A_1A_4) + P(A_2A_3) \\ &\qquad + P(A_2A_4) + P(A_3A_4). \end{aligned}$$

$\sum_{1 \le i_1 < i_2 < i_3 \le 4} P(A_{i_1}A_{i_2}A_{i_3})$ means

$$P(A_1A_2A_3) + P(A_1A_2A_4) + P(A_1A_3A_4) + P(A_2A_3A_4).$$

$\sum_{1 \le i_1 < i_2 < i_3 < i_4 \le 4} P(A_{i_1}A_{i_2}A_{i_3}A_{i_4})$ means

$$P(A_1A_2A_3A_4).$$

2.5 Sample Spaces Having Equally Likely Outcomes

For many experiments, it is natural to assume that all outcomes in the sample space (of finite number of elements) are equally likely to occur.
For example,

- tossing a fair coin: $P(\{\text{ head }\}) = P(\{\text{ tail }\})$
- tossing a pair of fair dice: $P(\{(1,1)\}) = P(\{(1,2)\}) = \cdots$.

Write $S = \{s_1, s_2, \ldots s_N\}$ where N denotes the number of outcomes of S. (We will use $|S|$ to denote the number of outcomes of S.) Since outcomes are assumed to be equally likely to occur, write $P(\{s_i\}) = c$, for $i = 1, 2, \ldots, N$. As

$$\begin{aligned} 1 &= P(S) \\ &= P(\cup_{i=1}^{N} \{s_i\}) \\ &= \sum_{i=1}^{N} P(\{s_i\}) \\ &= Nc, \end{aligned}$$

we get $c = 1/N$. In other words,

$$P(\{s_i\}) = \frac{1}{|S|}.$$

Similarly, if event A has $|A|$ outcomes, then $P(A) = |A|c$, that is,

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S}$$

when the outcomes of the sample space are equally likely.
This is why we need all those counting methods.

Example 2.15. A pair of fair dice is tossed. What is the probability of getting a sum of 7?

Solution:
As the dice are fair, we assume all outcomes are equally likely. So

$$\begin{aligned} A &= \{\text{ sum is } 7\} \\ &= \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\} \end{aligned}$$

therefore,

$$P(A) = \frac{|A|}{|S|} = \frac{6}{36} = \frac{1}{6}.$$

Example 2.16. If 3 balls are "randomly drawn" from an urn containing 6 white and 5 black balls, what is the probability that one of the drawn balls is white and the other two black?

Solution:
(I) Regard the order in which the balls are selected as being relevant.
There are a total of $11 \cdot 10 \cdot 9 = 990$ ways to draw 3 balls, this is $|S|$.
To get one white two black, there are 3 cases:

first ball is white. Number of ways $= 6 \cdot 5 \cdot 4 = 120$.

second ball is white. Number of ways $= 5 \cdot 6 \cdot 4 = 120$.

third ball is white. Number of ways $= 5 \cdot 4 \cdot 6 = 120$.

Therefore, $|A| = 120 + 120 + 120 = 360$. It follows that

$$P(A) = \frac{360}{990} = \frac{4}{11}.$$

(II) Regard the order in which the balls are selected as being irrelevant.
There are a total of $\binom{11}{3} = 165$ ways to draw 3 balls, this is $|S|$.
To get one white two black, there are $\binom{5}{2}\binom{6}{1} = 60$ ways.
Hence

$$P(A) = \frac{60}{165} = \frac{4}{11}.$$

Example 2.17. A poker hand consists of 5 cards. If the cards have distinct consecutive values and are not all of the same suit, we say that the hand is a straight. What is the probability that one is dealt a straight?

Solution:
We assume that all $\binom{52}{5}$ possible poker hands are equally likely. Let us first determine the number of possible outcomes for which the poker hand consists of an ace, two, three, four, and five (the suits being irrelevant).

The ace can be any 1 of the 4 possible aces, and similarly for the two, three, four, and five. It follows that there are 4^5 outcomes leading to exactly one ace, two, three, four, and five. In 4 of these outcomes all the cards will be of the same suit, it follows that there are $4^5 - 4$ hands that make up a straight of the form ace, two, three, four, and five.

Similarly, there are $4^5 - 4$ hands that make up a straight of the form two, three, four, five and six. So there are $10(4^5 - 4)$ hands that are straights.
Thus the desired probability is

$$\frac{10(4^5 - 4)}{\binom{52}{5}} \approx 0.0039.$$

Example 2.18. There are 40 students in a canoe class, of which there are 20 boys and 20 girls . These 40 students are paired into groups of 2 for determining partners in the canoe practice.
If the pairing is done at random,

(a) what is the probability that there is no boy-girl pairs?

(b) what is the probability that there are exactly 2 boy-girl pairs?

(c) what is the probability that there are exactly $2i$ boy-girl pairs?

Solution:

(a) There are

$$\binom{40}{2}\binom{38}{2}\cdots\binom{4}{2}\binom{2}{2} = \frac{40!}{2^{20}}$$

ways of dividing the 40 students into 20 ordered pairs of two each.

Hence there are

$$\frac{40!}{2^{20}20!}$$

ways of dividing the students into unordered pairs of 2. That is,

$$|S| = \frac{40!}{2^{20}20!}.$$

Let A be the event that there is no boy-girl pairs. For A to occur, we could have

(i) Boys pair with boys:
Number of ways is

$$\binom{20}{2}\binom{18}{2}\cdots\binom{2}{2}/10! = \frac{20!}{2^{10}10!}.$$

(ii) Girls pair with girls:
Number of ways is

$$\frac{20!}{2^{10}10!}.$$

Therefore, the probability of having no boy-girl pairs is

$$\frac{\left(\frac{20!}{2^{10}10!}\right)^2}{\frac{40!}{2^{20}20!}} = \frac{(20!)^3}{(10!)^2 40!},$$

which works out to be 1.34×10^{-6}.

(b) There are

$\binom{20}{2}$ ways of choosing 2 boys;

$\binom{20}{2}$ ways of choosing 2 girls;

and there are 2 ways of forming the two boy-girl pairs.

For the remaining 18 boys, they will paired among themselves. Number of ways of doing so is

$$\binom{18}{2}\binom{16}{2}\cdots\binom{2}{2}/9! = \frac{18!}{2^9 9!}$$

Similarly, for the remaining 18 girls, they will pair among themselves. Number of ways is

$$\frac{18!}{2^9 9!}.$$

Summing up, the probability of having exactly 2 boy-girl pairs is

$$\frac{\binom{20}{2}^2 \cdot 2\left(\frac{18!}{2^9 9!}\right)^2}{\frac{40!}{2^{20}20!}}.$$

(c) Refer to Ross p.39.

Example 2.19 (Birthday problem). There are n people in a room, what is the probability that there are at least two people share the same birthday? (We assume each day is equally likely to be a birthday of everyone, and there is no leap year.)

Solution:
A person can have his birthday on any of the 365 days.
There is a total of $(365)^n$ of outcomes, ie $|S|$.
Let A denote the event that there are at least two people among the n people sharing the same birthday.

We will work out A^c first. To count A^c, note that

$$|A^c| = 365 \cdot 364 \cdots [365 - (n-1)].$$

Therefore,

$$P(A^c) = \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n}.$$

Or another way to write it:

$$P(A^c) = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right).$$

Therefore,

$$P(A) = 1 - \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right).$$

Let $q_n = P(A^c)$ when there are n people, and $p_n = P(A) = 1 - q_n$. The values of p_n and q_n for selected values of n are tabulated below:

n	q_n	p_n
1	1	0
2	0.99726	0.00274
3	0.99180	0.00820
10	0.88305	0.11695
15	0.74710	0.25290
20	0.58856	0.41144
21	0.55631	0.44369
22	0.52430	0.47570
23	0.49270	0.50730
24	0.46166	0.53834
25	0.43130	0.56870
30	0.29368	0.70632
35	0.18562	0.81438
40	0.10877	0.89123
45	0.059024	0.940976
50	0.029626	0.979374
100	3.0725×10^{-7}	1
253	6.9854×10^{-53}	1

Take note that the probability of having two people sharing the same birthday exceeds $1/2$ once you have 23 people.

Example 2.20 (Birthday problem II). How large must the group be so that there is a probability of greater than 0.5 that someone will have the same birthday as you do?

Solution:
Assuming 365 equally likely birthdays. In a group of n people the probability that all will fail to have your birthday is $(364/365)^n$. Setting this equal to 0.5 and solving

$$n = \frac{\log(0.5)}{\log(364/365)} = 252.7.$$

So we need 253 people.

Remark 4. In our first encounter with the birthday problem it was surprising that the size needed to have 2 people with the same birthday was so small. This time the surprise goes in the other direction. A naive guess is that 183 people will be enough.

The "problem" is that many people in the group will have the same birthday, so the number of different birthdays is smaller than the size of the group.

Example 2.21. A deck of 52 playing cards is shuffled and the cards turned up one at a time until the first ace appears. Is the next card – that is, the card following the first ace – more likely to be the ace of spades or the two of clubs?

Solution:
We first need to calculate how many of the 52! possible ordering of the cards have the ace of spades immediately following the first ace. To begin, note that each ordering of the 52 cards can be obtained by first ordering the 51 cards different from the ace of spades and then inserting the ace of spades into that ordering.

Furthermore, for each of the 51! orderings of the other cards, there is only one place where the ace of spades can be placed so that it follows the first ace. For instance, if the ordering of the other 51 cards is

$$4c,\ 6h,\ Jd,\ 5s,\ Ac,\ 7d,\ \ldots,\ Kh.$$

then the only insertion of the ace of spaces into this ordering that results it following the first ace is

$$4c,\ 6h,\ Jd,\ 5s,\ Ac,\ As,\ 7d,\ \ldots,\ Kh.$$

Therefore, we see that there are 51! orderings that result in the ace of spades following the first ace, so

$$P(\{\text{the ace of spades follows the first ace}\}) = \frac{51!}{52!} = \frac{1}{52}.$$

In fact, by exactly the same argument, it follows that the probability that the two of clubs (or any other specified card) follows the first ace is also $\frac{1}{52}$. In other words, each of the 52 cards of the deck is equally likely to be the one that follows the first ace.

Example 2.22 (Matching problem). Suppose each of N men at a party throws his hat into the center of the room. The hats are first mixed up, and then each man randomly selects a hat. What is the probability that none of them selects his own hat?

Solution:
Let A_k denote the event that the kth man gets back his hat. Then, (a) is indeed asking for $P(A_1^c A_2^c \cdots A_N^c)$.
By De Morgan's laws,

$$(A_1 \cup A_2 \cup \cdots \cup A_N)^c = A_1^c A_2^c \cdots A_N^c.$$

We are then interested to compute

$$\begin{aligned} & P(A_1 \cup A_2 \cup \cdots \cup A_N) \\ = & \sum_{i=1}^{N} P(A_i) - \sum_{1 \le i_1 < i_2 \le N} P(A_{i_1} A_{i_2}) + \cdots \\ & + (-1)^{r+1} \sum_{1 \le i_1 < \cdots < i_r \le N} P(A_{i_1} \cdots A_{i_r}) \\ & + \cdots + (-1)^{N+1} P(A_1 \cdots A_N). \end{aligned}$$

To go through the action slowly, we consider

(i) $\sum_{i=1}^{N} P(A_i)$:
Notice that, by relabelling the N men if necessary, we see that for any $1 \le i \le N$,

$$P(A_i) = P(A_1) = \frac{1 \cdot (N-1)(N-2) \cdots 1}{N!} = \frac{1}{N}.$$

Therefore,

$$\sum_{i=1}^{N} P(A_i) = N \frac{1}{N} = 1.$$

(ii) $\sum_{1\le i_1<i_2\le N} P(A_{i_1}A_{i_2})$:
For any $1 \le i_1 < i_2 \le N$,

$$\begin{aligned} P(A_{i_1}A_{i_2}) = P(A_1A_2) &= \frac{1^2 \cdot (N-2)(N-3)\cdots 1}{N!} \\ &= \frac{(N-2)!}{N!}. \end{aligned}$$

Therefore,

$$\sum_{1\le i_1<i_2\le N} P(A_{i_1}A_{i_2}) = \frac{(N-2)!}{N!} \times \binom{N}{2} = \frac{1}{2!}.$$

(iii) Calculating the general term, $\sum_{1\le i_1<\cdots<i_r\le N} P(A_{i_1}\cdots A_{i_r})$:
First note that for any $1 \le i_1 < i_2 < \cdots < i_r \le N$,

$$\begin{aligned} P(A_{i_1}\cdots A_{i_r}) &= P(A_1\cdots A_r) \\ &= \frac{1^r \cdot (N-r)(N-r-1)\cdots 1}{N!} \\ &= \frac{(N-r)!}{N!} \end{aligned}$$

hence

$$\sum_{1\le i_1<\cdots<i_r\le N} P(A_{i_1}\cdots A_{i_r}) = \frac{(N-r)!}{N!}\binom{N}{r} = \frac{1}{r!}.$$

Therefore,

$$P(A_1 \cup A_2 \cup \cdots \cup A_N) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + (-1)^{N+1}\frac{1}{N!}.$$

So the probability that none of the men selects his own hat is

$$\begin{aligned} P(A_1^c\cdots A_N^c) &= P((A_1 \cup A_2 \cup \cdots \cup A_N)^c) \\ &= 1 - P(A_1 \cup A_2 \cup \cdots \cup A_N) \\ &= 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots + (-1)^N\frac{1}{N!} \\ &\approx e^{-1} = 0.3679 \quad \text{if } N \text{ is large.} \end{aligned}$$

2.6 Probability as a Continuous Set Function

Definition 2.23. *A sequence of events* $\{E_n\}$, $n \ge 1$ *is said to be an* increasing *sequence if*

$$E_1 \subset E_2 \subset \cdots \subset E_n \subset E_{n+1} \subset \cdots$$

whereas it is said to be a decreasing *sequence if*

$$E_1 \supset E_2 \supset \cdots \supset E_n \supset E_{n+1} \supset \cdots$$

Definition 2.24. *If* $\{E_n\}$, $n \geq 1$ *is an increasing sequence of events, then we define a new event, denoted by* $\lim_{n\to\infty} E_n$ *as*

$$\lim_{n\to\infty} E_n = \cup_{i=1}^{\infty} E_i.$$

Similarly, if $\{E_n\}$, $n \geq 1$ *is a decreasing sequence of events, then we define a new event, denoted by* $\lim_{n\to\infty} E_n$ *as*

$$\lim_{n\to\infty} E_n = \cap_{i=1}^{\infty} E_i.$$

Proposition 2.25. *If* $\{E_n\}$, $n \geq 1$ *is either an increasing or a decreasing sequence of events, then*

$$\lim_{n\to\infty} P(E_n) = P\left(\lim_{n\to\infty} E_n\right).$$

Proof. For a proof, refer to Ross p. 44. □

Example 2.26. Consider an infinite sequence of tosses of a fair coin. Compute the probability that there is at least one head and one tail in this infinite sequence of tosses.

Solution:
Let A_n denote the event that there are no heads in the first n tosses. Then $P(A_n) = 1/2^n$.
Now,

$$A_1 \supset A_2 \supset \cdots \supset A_n \supset A_{n+1} \supset \cdots,$$

therefore

$$P(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} \frac{1}{2^n} = 0.$$

However, $\cap_{n=1}^{\infty} A_n = \{\text{No heads in the infinite sequence of tosses}\}$, so

$$P(\text{At least one head in the infinite sequence of tosses}) = 1.$$

Using the same argument,

$$P(\text{At least one tail in the infinite sequence of tosses}) = 1.$$

It can be shown that if $P(A) = 1 = P(B)$, then $P(AB) = 1$. Thus

$$P(\text{At least one head and one tail in the infinite sequence of tosses}) = 1.$$

In fact, this probability is the same for any biased coin.

Chapter 3

Conditional Probability and Independence

3.1 Introduction

In many problems, we are interested in event B. However, we have some partial information, namely, that an event A has occurred. How to make use of this information?
In calculating $P(B)$, there are occasions that we can consider it under different cases. How do we compute the probability of B under these cases? How do we combine them to give $P(B)$?

3.2 Conditional Probabilities

Definition 3.1. *Let A and B be two events. Suppose that $P(A) > 0$, the conditional probability of B given A is defined as*

$$\frac{P(AB)}{P(A)}$$

and is denoted by $P(B|A)$.

Remark 5. $P(B|A)$ can also be read as: the conditional probability that B occurs given that A has occurred.

Since we know that A has occurred, think of A as our new, or **reduced sample space**. The probability that the event AB occurs will equal the probability of AB relative to the probability of A.

Example 3.2. A fair coin is flipped twice. All outcomes in

$$S = \{(H,H),(H,T),(T,H),(T,T)\}$$

are equally likely. What is the probability that both flips result in heads given that the first flip does?

Solution:
Let $B = \{(H,H)\}$ and $A = \{(H,H),(H,T)\}$.
Now $P(A) = |A|/|S| = 2/4 = 1/2$, and $P(AB) = P(\{(H,H)\}) = 1/4$. So

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{1/4}{1/2} = 1/2.$$

Alternatively:
Because the first flip lands on heads, we are left with the 2 equally likely outcomes $(H,H),(H,T)$, only one of which results in both flips landing on heads. So the required probability is $1/2$.

Example 3.3. In the card game bridge, the 52 cards are dealt out equally to 4 players – called East, West, North, and South. If North and South have a total of 8 spades among them, what is the probability that East has 3 of the remaining 5 spades?

Solution:
We work with the reduced sample space.

Since North-South have a total of 8 spades among their 26 cards, there are a total of 26 cards, 5 of them being spades, to be distributed among the East-West hands.

So the conditional probability that East will have exactly 3 spades is

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

Example 3.4. A bin contains 25 light bulbs, of which 5 are good and function at least 30 days, 10 are partially defective and will fail in the second day of use, while the rest are totally defective and won't light up at all. Given that a randomly chosen bulb initially lights up, what is the probability that it will still be working after one week?

Solution:
Let G be event that the randomly chosen bulb is in good condition, T the event that the randomly chosen bulb is totally defective.

Given that the chosen bulb lights up initially, that is, given that T^c occurs. So the conditional probability required is

$$P(G|T^c) = \frac{P(GT^c)}{P(T^c)} = \frac{5/25}{15/25} = 1/3.$$

Exercise: Solve this using the reduced sample space approach.

(Multiplication Rule)
Suppose that $P(A) > 0$, then

$$P(AB) = P(A)P(B|A).$$

Example 3.5. Suppose that an urn contains 8 red balls and 4 white balls. We draw 2 balls, one at a time, from the urn without replacement. If we assume that at each draw each ball in the urn is equally likely to be chosen, what is the probability that both balls drawn are red?

Now suppose that the balls have different weights, with each red ball having weight r and each white ball having weight w. Suppose that the probability that a given ball in the urn is the next one selected is its weight divided by the sum of the weights of all balls currently in the urn. Now what is the probability that both balls are red?

Solution:
Let R_1 and R_2 denote respectively the events that the first and second ball drawn are red. We want to compute $P(R_1R_2)$.
As $P(R_1) = 8/12$ and $P(R_2|R_1) = 7/11$,

$$P(R_1R_2) = P(R_1)P(R_2|R_1) = 8/12 \cdot 7/11 = 14/33.$$

Alternatively

$$P(R_1R_2) = \frac{\binom{8}{2}}{\binom{12}{2}} = \frac{8 \cdot 7}{12 \cdot 11} = 14/33.$$

For the second part, number the red balls, and let $B_i, i = 1, \ldots, 8$ be the event that the first ball drawn is red ball number i. Then

$$P(R_1) = P\left(\bigcup_{i=1}^{8} B_i\right) = \sum_{i=1}^{8} P(B_j) = 8 \times \frac{r}{8r+4w}.$$

When the first ball is red, the urn then contains 7 red and 4 white balls. Thus

$$P(R_2|R_1) = \frac{7r}{7r+4w}.$$

The probability that both balls are red is then

$$P(R_1R_2) = \frac{8r}{8r+4w} \times \frac{7r}{7r+4w}.$$

(General Multiplication Rule)
Let $A_1, A_2, \ldots, A_n$ be n events, then

$$\begin{aligned} & P(A_1A_2\cdots A_n) \\ = \; & P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1}). \end{aligned}$$

Proof. (Sketch)

$$RHS \;=\; P(A_1)\frac{P(A_1A_2)}{P(A_1)}\frac{P(A_1A_2A_3)}{P(A_1A_2)}\cdots\frac{P(A_1\cdots A_n)}{P(A_1\cdots A_{n-1})}$$

□

Example 3.6. In the matching problem stated in Example 2.22 of Chapter 2 (page 25), it was shown that P_N, the probability that there are no matches when N people randomly select from among their own N hats, is given by

$$P_N = \sum_{i=0}^{N} (-1)^i/i!$$

What is the probability that exactly k of the N people have matches?

Solution:
Consider a particular set of k people. We will determine the probability that these k individuals have matches and no one else does.

Let E denote the event that everyone in this set has a match, and G the event that none of the other $N-k$ people have a match. We have

$$P(EG) = P(E)P(G|E).$$

Now, let $F_i, i = 1, \ldots, k$, be the event that the ith member of the set has a match. Then

$$\begin{aligned} P(E) &= P(F_1 F_2 \cdots F_k) \\ &= P(F_1)P(F_2|F_1)P(F_3|F_1F_2)\cdots P(F_k|F_1\cdots F_{k-1}) \\ &= \frac{1}{N} \times \frac{1}{N-1} \times \frac{1}{N-2} \times \cdots \times \frac{1}{N-k+1} \\ &= \frac{(N-k)!}{N!}. \end{aligned}$$

Given that everyone in the set of k has a match, the probability that none of the other $N-k$ people has a match is equal to the probability of no matches in a problem having $N-k$ people choosing among their own $N-k$ hats. Therefore,

$$P(G|E) = P_{N-k} = \sum_{i=0}^{N-k} (-1)^i / i!$$

showing that the probability that a specified set of k people have matches and no one else does is

$$P(EG) = \frac{(N-k)!}{N!} P_{N-k}.$$

Since there are $\binom{N}{k}$ sets of k individuals, the desired probability is

$$P(\text{exactly } k \text{ matches}) = P_{N-k}/k!$$

which is approximately $e^{-1}/k!$ when N is large.

Example 3.7. An ordinary deck of 52 playing cards is randomly divided into 4 **distinct** piles of 13 cards each. Compute the probability that each pile has exactly 1 ace.

Solution:
Define events E_i, $i = 1,2,3,4$ as follows:

$E_1 = \{\text{the ace of spades is in any one of the piles}\}$
$E_2 = \{\text{the ace of spades and the ace of hearts are in different piles}\}$
$E_3 = \{\text{the aces of spades, hearts and diamonds are all in different piles}\}$
$E_4 = \{\text{all 4 aces are in different piles}\}$

Note then that the required probability is given by

$$P(E_1E_2E_3E_4) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2)P(E_4|E_1E_2E_3).$$

It is obvious that $P(E_1) = 1$.

Also, $P(E_2|E_1) = \frac{39}{51}$ since the pile containing the ace of spades will receive 12 of the remaining 51 cards.

Similarly, $P(E_3|E_1E_2) = \frac{26}{50}$ since the piles containing the ace of spades and the ace of hearts will receive 24 of the remaining 50 cards.

Finally, $P(E_4|E_1E_2E_3) = \frac{13}{49}$ and so the required probability is

$$P(E_1E_2E_3E_4) = \frac{39 \cdot 26 \cdot 13}{51 \cdot 50 \cdot 49} \approx 0.105.$$

3.3 Bayes' Formulas

Let A and B be any two events, then

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

Proof.

$$\begin{aligned} P(B) &= P(B \cap (A \cup A^c)) \\ &= P(BA \cup BA^c) \\ &= P(BA) + P(BA^c) \\ &= P(B|A)P(A) + P(B|A^c)P(A^c), \end{aligned}$$

□

Remark 6. Weighted average of two cases: case 1 being A occurs; case 2 being A does not occur. The weights are placed according to how likely A, A^c occur.

Example 3.8. In answering a question on a multiple-choice test, a student either knows the answer or guesses the answer at random. Let p be the probability that the student knows the answer, and $1-p$ the probability that he doesn't. Suppose there are m alternatives in the question.

(a) What is the probability that he answered it correctly?

(b) What is the probability that the student knew the answer given that he answered it correctly?

Solution:
Let C denote the event that he answered it correctly, K that he knew the answer. So $P(K) = p$, and $P(K^c) = 1-p$. Also, $P(C|K) = 1$, and $P(C|K^c) = 1/m$.

(a)

$$\begin{aligned} P(C) &= P(C|K)P(K) + P(C|K^c)P(K^c) \\ &= 1 \cdot p + 1/m \cdot (1-p). \end{aligned}$$

(b) We are asked $P(K|C)$.

From $P(K|C) = P(KC)/P(C)$, we need to compute $P(CK)$.

However, $P(CK) = P(C|K)P(K) = p$, so

$$\begin{aligned} P(K|C) &= \frac{P(C|K)P(K)}{P(C)} \\ &= \frac{p}{p + (1-p)/m} \\ &= \frac{mp}{1 + (m-1)p}. \end{aligned}$$

Example 3.9. A laboratory test is 95% accurate in detecting a certain disease when it is, in fact, present. However, the test also yields a "false positive" result for 1% of the healthy persons tested. If 0.5% of the population actually has the disease, what is the probability that a person has the disease given that the test result is positive?

Solution:
Let D be the event that a person has the disease, and A the event that the test result is positive.
We are asked to find $P(D|A)$. Given $P(D) = 0.005$, $P(A|D) = 0.95$, and $P(A|D^c) = 0.01$. Now

$$\begin{aligned} P(D|A) &= \frac{P(AD)}{P(A)} \\ &= \frac{P(A|D)P(D)}{P(A|D)P(D) + P(A|D^c)P(D^c)} \\ &= \frac{(0.95)(0.005)}{(0.95)(0.005) + (0.01)(0.995)} \\ &= 0.323 \end{aligned}$$

This means that only 32.3% of those persons whose test results are positive actually has the disease!! Not 95%
However, note the increase from 0.5% (before the test) to 32.3% after the test indicates positive.
Try

(i) a more common disease (with everything the same): i.e. $P(D) = 10\%$;

(ii) a rarer disease 0.001% disease.

And compare these.

Definition 3.10. *We say that $A_1, A_2, \ldots, A_n$ **partitions** the sample space S if:*

(a) They are "mutually exclusive", meaning $A_i \cap A_j = \emptyset$, for all $i \neq j$.

(b) They are "collectively exhaustive", meaning $\cup_{i=1}^{n} A_i = S$.

Proposition 3.11 (Bayes' first formula)**.** *Suppose the events $A_1, A_2, \ldots, A_n$ partitions the sample space. Assume further that $P(A_i) > 0$ for $1 \leq i \leq n$. Let B be any event, then*

$$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n).$$

Proof. Observe that

$$B = B \cap S = B \cap (\cup_{i=1}^{n} A_i) = \cup_{i=1}^{n} BA_i$$

and that $BA_1, \ldots, BA_n$ are mutually exclusive.
Therefore,

$$\begin{aligned} P(B) &= P(\cup_{i=1}^{n} BA_i) \\ &= P(BA_1) + P(BA_2) + \cdots + P(BA_n) \\ &= P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n). \end{aligned}$$

□

Proposition 3.12 (Bayes' second formula)**.** *Suppose the events $A_1, A_2, \ldots, A_n$ partitions the sample space. Assume further that $P(A_i) > 0$ for $1 \leq i \leq n$. Let B be any event, then for any $1 \leq i \leq n$,*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)}.$$

Proof.

$$P(A_i|B) = \frac{P(A_iB)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)}$$

and use Bayes' first formula for the denominator. □

Remark 7. Bayes' first formula, also known as the **Law of Total Probability**, is useful in computing the probability of a composite event, that is, an event B which depends on a series of "causes" $A_1, A_2, \cdots$. The formula tells us how to compute $P(B)$ when the probabilities of the causes $A_1, A_2, \cdots$ and the conditional probabilities of the event B given the cause A_i are known.

Bayes' second formula has been interpreted as a formula for "inverse probabilities". If $A_1, A_2, \cdots$ is a series of causes and B is a possible effect, then $P(B|A_i)$ is the probability of B when it is known that A_i is the cause, whereas $P(A_i|B)$ is the probability that A_I is the cause when it is known that B is the effect.

Example 3.13. Thousands of little snails live in the Snail Garden between the University at Buffalo Natural Sciences Complex and Talbert Hall. A concrete stairway divides the garden into two sections.

When it rains, the snails bravely travel from one section of the garden to the other by crawling across the steps of the concrete stairway. It is a treacherous journey. Many of the snails are crushed by the feet of unthinking students and faculty walking up and down the stairway.

There are two types of snails, *smart* snails and *not-so-smart* snails. A *smart* snail crawls along the vertical portion of a step to avoid being crushed. A *not-so-smart* snail crawls on the horizontal part of a step.

Suppose that half of the snails are *smart* snails, and the other half are *not-so-smart* snails. The probability that a *smart* snail will be crushed is 0.10. The probability that a *not-so-smart* snail will be crushed is 0.40. The event that any one snail will be crushed is independent from all others.

(a) Find the probability that a single snail (chosen at random) will successfully cross the concrete steps uncrushed.

(b) Suppose that after the end of a rainstorm, you select, at random, a snail that has successfully crossed the concrete steps. What is the probability that it is a *smart* snail? [1]

Solution:
Define the events

$$S = \{\text{The snail is a } \textit{smart} \text{ snail}\}$$
$$S^C = \{\text{The snail is a } \textit{not-so-smart} \text{ snail}\}$$
$$A = \{\text{The snail is successful}\}$$

[1] No animals were harmed in the making of this example.

(a) Note that S, S^C is a partition of the sample space. Therefore, using the Law of Total Probability,

$$P(A) = P(A|S)P(S) + P(A|S^C)P(S^C) = (0.90)(0.50) + (0.60)(0.50) = 0.75.$$

(b) In this case, we want to find

$$\begin{aligned} P(S|A) &= \frac{P(A|S)P(S)}{P(A|S)P(S) + P(A|S^C)P(S^C)} \\ &= \frac{(0.90)(0.50)}{(0.90)(0.50) + (0.60)(0.50)} = 0.60. \end{aligned}$$

Example 3.14. A plane is missing, and it is presumed that it was likely to have gone down in any three of the possible regions. Let $1 - \beta_i$ denote the probability that the plane will be found in the ith region given that in fact the plane went down in ith region, $i = 1, 2, 3$. (The constants β_i are called overlook probabilities.) What is the conditional probability that the plane is in the 2nd region given that a search in region 1 is unsuccessful?

Solution:
Let R_1, R_2 and R_3 denote the event that the plane is in region 1, region 2 and region 3 respectively. Let F be the event that the search at region 1 is unsuccessful. We want $P(R_2|F)$.
We will need $P(R_1) = P(R_2) = P(R_3) = 1/3$, $P(F|R_1) = \beta_1$, $P(F|R_2) = 1$ and $P(F|R_3) = 1$.
Apply Bayes second formula:

$$\begin{aligned} & P(R_2|F) \\ &= \frac{P(F|R_2)P(R_2)}{P(F|R_1)P(R_1) + P(F|R_2)P(R_2) + P(F|R_3)P(R_3)} \\ &= \frac{1 \cdot 1/3}{\beta_1 \cdot 1/3 + 1 \cdot 1/3 + 1 \cdot 1/3} \\ &= \frac{1}{\beta_1 + 2}. \end{aligned}$$

Similarly, we can find (it is of interest too!) $P(R_1|F)$ and $P(R_3|F)$.

3.4 Independent Events

Definition 3.15. *Two events A and B are said to be* ***independent*** *if*

$$P(AB) = P(A)P(B).$$

They are said to be **dependent** if

$$P(AB) \neq P(A)P(B).$$

Remark 8. (1) Motivation: Suppose $P(B) > 0$. Then,

$$\begin{aligned} P(A|B) &= \frac{P(AB)}{P(B)} \\ &= \frac{P(A)P(B)}{P(B)} \\ &= P(A). \end{aligned}$$

That is, A is independent of B if knowledge that B has occurred does NOT change the probability that A occurs.

(2) The following phrases mean the same:

A and B are independent;

A is independent of B;

B is independent of A.

Example 3.16. A card is selected at random from a deck of 52 playing cards. If A is the event that the selected card is an ace, B is the event that it is a ♠, are events A and B independent?

Solution:
$A = \{A\spadesuit, A\heartsuit, A\diamondsuit, A\clubsuit\}$; and
$B = \{A\spadesuit, 2\spadesuit, 3\spadesuit, \ldots, J\spadesuit, Q\spadesuit, K\spadesuit\}$.
Therefore $AB = \{A\spadesuit\}$.
It follows that
$P(A) = \frac{4}{52} = \frac{1}{13}$;
$P(B) = \frac{13}{52} = \frac{1}{4}$; and
$P(AB) = \frac{1}{52}$.
And we see that $P(AB) = P(A)P(B)$. Therefore, A and B are independent.

Example 3.17. Suppose we toss 2 fair dice.

(a) Let A_6 denote the event that the sum of two dice is 6 and B denote the event that the first die equals 4.

Then,

$A_6 = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$ and

$B = \{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}$.

So, $A_6B = \{(4,2)\}$.

Hence, $P(A_6) = \frac{5}{36}$, $P(B) = \frac{6}{36} = \frac{1}{6}$ and $P(A_6B) = \frac{1}{36}$. It can be checked easily that

$$P(A_6B) \neq P(A_6)P(B).$$

That is, A_6 and B are **dependent**.

(b) Let A_7 denote the event that the sum of two dice is 7.

$P(A_7B) = \frac{1}{36}$, $P(A_7) = \frac{6}{36} = \frac{1}{6}$ and $P(B) = \frac{1}{6}$. Hence,

$$P(A_7B) = \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = P(A_7)P(B).$$

That is, A_7 and B are **independent**.

Proposition 3.18. *If A and B are independent, then so are*

(i) A and B^c;

(ii) A^c and B;

(iii) A^c and B^c.

Proof.

(i) Since

$$\begin{aligned} P(A) &= P(AB) + P(AB^c) \\ &= P(A)P(B) + P(AB^c) \quad \text{as } A, B \text{ are independent} \end{aligned}$$

hence

$$\begin{aligned} P(AB^c) &= P(A) - P(A)P(B) \\ &= P(A)[1 - P(B)] = P(A)P(B^c). \end{aligned}$$

(ii) Similarly for the parts (ii) and (iii).

□

Remark 9. If A is independent of B, and A is also independent of C, it is NOT necessarily true that A is independent of BC.

Example 3.19. Two fair dice are thrown. Let A be the event that the sum of the dice is 7; B the event that first die is 4; and C the event that the second die is 3.
$A = \{(1,6),(2,6),(3,4),(4,3),(5,2),(6,1)\}$.
$B = \{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}$.
$C = \{(1,3),(2,3),(3,3),(4,3),(5,3),(6,3)\}$.
$AB = \{(4,3)\}$
$AC = \{(4,3)\}$
$BC = \{(4,3)\}$ and
$ABC = \{(4,3)\}$
Then,

$$P(AB) = P(A)P(B)$$

and

$$P(AC) = P(A)P(C)$$

that is A and B are independent; also A and C are independent.
However,

$$P(ABC) = \frac{1}{36} \neq \frac{1}{216} = \frac{1}{6} \times \frac{1}{36} = P(A)P(BC).$$

Definition 3.20. *Three events A, B and C are said to be independent if the following 4 conditions hold:*

$$P(ABC) = P(A)P(B)P(C) \quad (1)$$
$$P(AB) = P(A)P(B) \quad (2)$$
$$P(AC) = P(A)P(C) \quad (3)$$
$$P(BC) = P(B)P(C) \quad (4)$$

Remark 10. Second condition implies A and B are independent;
Third condition implies A and C are independent; and
Fourth condition implies B and C are independent
That is, A, B and C are pairwise independent.

Example 3.21. It should be noted that if A, B and C are independent, then A is independent of any event formed from B and C.

(i) A is independent of $B \cup C$.

(ii) A is independent of $B \cap C$.

Solution:

(i) Consider

$$\begin{aligned}
P(A(B\cup C)) &= P(AB\cup AC)\\
&= P(AB)+P(AC)-P(AB\cap AC)\\
&= P(A)P(B)+P(A)P(C)-P(ABC) && \text{from (2), (3)}\\
&= P(A)P(B)+P(A)P(C)-P(A)P(B)P(C) && \text{from (1)}\\
&= P(A)[P(B)+P(C)-P(B)P(C)]\\
&= P(A)[P(B)+P(C)-P(BC)] && \text{from (4)}\\
&= P(A)P(B\cup C),
\end{aligned}$$

that is, A and $B\cup C$ are independent.

(ii) Follows from definition. (Exercise).

Definition 3.22. *Events $A_1, A_2, \ldots, A_n$ are said to be independent if, for every sub-collection of events $A_{i_1}, \ldots, A_{i_r}$, we have*

$$P(A_{i_1}\cdots A_{i_r}) = P(A_{i_1})\cdots P(A_{i_r}).$$

Explanation of the above –
For $n=4$, that is, 4 events A_1, A_2, A_3 and A_4 are independent if we can verify (i), (ii) and (iii) below:

(i) $r=4$:

$$P(A_1A_2A_3A_4) = P(A_1)P(A_2)P(A_3)P(A_4)$$

(ii) $r=3$: there are 4 conditions,

$$\begin{aligned}
P(A_1A_2A_3) &= P(A_1)P(A_2)P(A_3)\\
P(A_1A_2A_4) &= P(A_1)P(A_2)P(A_4)\\
P(A_1A_3A_4) &= P(A_1)P(A_3)P(A_4)\\
P(A_2A_3A_4) &= P(A_2)P(A_3)P(A_4)
\end{aligned}$$

(iii) $r=2$: there are 6 conditions,

$$\begin{aligned}
P(A_1A_2) &= P(A_1)P(A_2)\\
P(A_1A_3) &= P(A_1)P(A_3)\\
P(A_1A_4) &= P(A_1)P(A_4)\\
P(A_2A_3) &= P(A_2)P(A_3)\\
P(A_2A_4) &= P(A_2)P(A_4)\\
P(A_3A_4) &= P(A_3)P(A_4)
\end{aligned}$$

Example 3.23. We are given a loaded coin with probability of getting a head $= p$; with probability of getting a tail $= 1-p$.

(a) This loaded coin is tossed n times independently. What is the probability of getting

  (i) at least 1 head in these n tosses?

  (ii) exactly k heads in these n tosses?

(b) If this loaded coin is kept being tossed independently indefinitely, what is the probability that all trials result in heads?

Solution:

(a) Let A_i be the event that the ith toss results in a head.

  (i) Probability required is

$$\begin{aligned} & P(A_1 \cup \cdots \cup A_n) \\ = \; & 1 - P((A_1 \cup \cdots \cup A_n)^c) \\ = \; & 1 - P(A_1^c A_2^c \cdots A_n^c) \\ = \; & 1 - P(A_1^c)P(A_2^c)\cdots P(A_n^c) \quad \text{by independence} \\ = \; & 1 - (1-p)^n. \end{aligned}$$

  (ii) Probability of first k heads follows by $n-k$ tails is

$$\begin{aligned} & P(A_1 \cdots A_k A_{k+1}^c \cdots A_n^c) \\ = \; & P(A_1)\cdots P(A_k)P(A_{k+1}^c)\cdots P(A_n^c) \\ = \; & p^k(1-p)^{n-k}. \end{aligned}$$

  In order to have exactly k heads, there are

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

  ways of having k heads and $n-k$ tails. Hence the probability required is

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

(b) Denote the event that the first n tosses all result in heads by $E_n = \cap_{i=1}^{n} A_i$. Then $P(E_n) = p^n$. Note that $\{E_n\}$, $n \geq 1$ is a decreasing sequence of events.

Using Proposition 2.25 (page 27), the probability sought for is

$$\begin{aligned} P(\cap_{i=1}^{\infty} A_i) &= P\left(\lim_{n\to\infty} \cap_{i=1}^{n} A_i\right) \\ &= P\left(\lim_{n\to\infty} E_n\right) \\ &= \lim_{n\to\infty} P(E_n) \\ &= \lim_{n\to\infty} p^n \\ &= \begin{cases} 0, & \text{if } 0 \leq p < 1 \\ 1, & \text{if } p = 1 \end{cases}. \end{aligned}$$

Example 3.24. A system composed of n separate components is said to be a parallel system if it functions when at least one of the components functions. For such a system, if component i, independent of other components, functions with probability p_i, for $1 \leq i \leq n$. What is the probability that the system functions?

Solution:
Let A_i be the event that component i functions. Given that

$$P(A_i) = p_i, \quad \text{for } 1 \leq i \leq n.$$

Then

$$\begin{aligned} & P(\text{ system functions }) \\ = \; & P(A_1 \cup \cdots \cup A_n) \\ = \; & 1 - P([A_1 \cup \cdots \cup A_n]^c) \\ = \; & 1 - P(A_1^c A_2^c \cdots A_n^c) \\ = \; & 1 - P(A_1^c) P(A_2^c) \cdots P(A_n^c) \quad \text{by independence} \\ = \; & 1 - (1-p_1)(1-p_2)\cdots(1-p_n). \end{aligned}$$

Example 3.25. Independent trials consisting of rolling a pair of fair dice are performed. What is the probability that an outcome of 5 appears before an outcome of 7 when the outcome of a roll is the sum of the dice?

Solution:
Let E_n denote the event that no 5 or 7 appears on the first $n-1$ trials and a

5 appears on the nth trial, then the desired probability is

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

Now, since $P(5 \text{ on any trial}) = \frac{4}{36}$ and $P7 \text{ on any trial}) = \frac{6}{36}$, we obtain, by the independence of trials,

$$P(E_n) = \left(1 - \frac{10}{36}\right)^{n-1} \times \frac{4}{36}.$$

Thus,

$$\begin{aligned} P\left(\bigcup_{n=1}^{\infty} E_n\right) &= \frac{1}{9} \sum_{n=1}^{\infty} \left(\frac{13}{18}\right)^{n-1} \\ &= \frac{1}{9} \times \frac{1}{1 - \frac{13}{18}} \\ &= \frac{2}{5}. \end{aligned}$$

This result could also have been obtained by the use of conditional probabilities. If we let E be the event that a 5 occurs before a 7, then we can obtain the desired probability, $P(E)$, by conditioning on the outcome of the first trial, as follows: Let F be the event that the first trial results in a 5, let G be the event that it results in a 7, and let H be the event that the first trial results in neither a 5 nor a 7. Then, conditioning on which one of these events occurs gives

$$P(E) = P(E|F)P(F) + P(E|G)P(G) + P(E|H)P(H).$$

However,

$$P(E|F) = 1, \quad P(E|G) = 0 \quad \text{and} \quad P(E|H) = P(E).$$

The first two equalities are obvious. The third follows because if the first outcome results in neither a 5 nor a 7, then at that point the situation is exactly as it was when the problem first started – namely, the experimenter will continually roll a pair of fair dice until either a 5 or 7 appears.

Since $P(F) = \frac{4}{36}$, $P(G) = \frac{6}{36}$, and $P(H) = \frac{26}{36}$, it follows that

$$P(E) = \frac{1}{9} + \frac{13}{18} P(E)$$

which gives $P(E) = \frac{2}{5}$.

Note that the answer is quite intuitive. That is, because a 5 occurs on any roll with probability $\frac{4}{36}$ and a 7 with probability $\frac{6}{36}$, it seems intuitive that the odds that a 5 appears before a 7 should be 6 to 4 against. The probability should then be $\frac{4}{10}$.

Remark 11. The same argument shows that if E and F are mutually exclusive events of an experiment, then, when independent trials of the experiment are performed, the event E will occur before the event F with probability

$$\frac{P(E)}{P(E)+P(F)}.$$

de Méré - Pascal problem

A question posed in the mid-seventeenth century to Blaise Pascal by a French nobleman and inveterate gambler, the Chevalier de Méré's, marked the birth of probability theory. One of de Méré's favorite bets was that at least one six would appear during a total of four rolls of a die. From past experience, he knew that this gamble paid off more often than not. Then, for a change, he started betting that he would get a double-six on 24 rolls of two dice. However, he soon realized that his old approach to the game was more profitable. He asked his friend Pascal why. Pascal showed that the probability of getting at least one six in four rolls of a die is $1-(5/6)^4 \approx 0.5177$, which is slightly higher than the probability of at least one double-six in 24 throws of two dice, $1-(35/36)^24 \approx 0.4914$.

Example 3.26. Which is a better bet?
Gamble 1: Rolling at least a six in four throws of a single die.
Gamble 2: Rolling at least one double six in twenty four throws of a pair of fair dice.

Solution:
For Gamble 1:
Let $A_1, A_2, \ldots, A_4$ denote the event of getting a six in first, second, third and fourth throw respectively.
We want $P(A_1 \cup A_2 \cup A_3 \cup A_4)$.
First notice that $P(A_i) = \frac{1}{6}$, for $i = 1,2,3,4$.

So the probability of winning is

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup A_4) &= 1 - P(A_1^c A_2^c A_3^c A_4^c) \\ &= 1 - P(A_1^c)P(A_2^c)P(A_3^c)P(A_4^c) \\ &= 1 - (1 - P(A_1))^4 \\ &= 1 - \left(1 - \frac{1}{6}\right)^4 \\ &\approx 0.5177 \end{aligned}$$

For Gamble 2:
Let $B_1, B_2, \ldots, B_{24}$ denote the event of getting a double six in first, second, third, ... twenty-fourth throw respectively.
We want $P(B_1 \cup B_2 \cup \cdots \cup B_{24})$.
First notice that $P(B_i) = \frac{1}{36}$, for $1 \leq i \leq 24$.
So the probability of winning is

$$\begin{aligned} P(B_1 \cup B_2 \cup \cdots \cup B_{24}) &= 1 - P(B_1^c B_2^c \cdots B_{24}^c) \\ &= 1 - P(B_1^c)P(B_2^c) \cdots P(B_{24}^c) \\ &= 1 - (1 - P(B_1))^{24} \\ &= 1 - \left(1 - \frac{1}{36}\right)^{24} \\ &\approx 0.4914 \end{aligned}$$

So gamble 1 should be preferred. It has an advantage of 0.02635 over gamble 2.

This problem and others posed by de Méré's are thought to have been the original inspiration for a fruitful exchange of letters on probability between Pascal and Pierre de Fermat. Their combined work laid the foundations for probability theory as we know it today.

Gambler's Ruin Problem

Stating the problem

A is gambling against B, betting on the outcomes of successive "flips of a coin". On each flip, if the coin comes up heads, A wins and collects 1 unit from B, whereas if it comes up tails, A pays 1 unit to B.

They continue to do so until one of them runs out of money. (They are compulsive gamblers!!) It is assumed that the successive flips of the coin are independent and each flip results in a head with probability p. A starts with i units and B starts with $N-i$ units.
What is the probability that A ends up with all the money?

Solving the problem

We write $q = 1 - p$.

Step 1. Probabilistic derivation of the recursive relation, boundary values

For each $0 \leq k \leq N$, define

$$w_k = P(\text{ A starts with } k \text{ unit and ends up winning all}).$$

Note that we actually want to find w_i.

We employ the so called **first-step analysis:**
Conditioning on the result of the first flip, we have

$$w_k = pw_{k+1} + qw_{k-1}, \quad \text{for } 1 \leq k \leq N \qquad (*)$$

with $w_0 = 0$ and $w_N = 1$.

Why? Think in terms of a tree:

If the first flip is head, then A's fortune becomes $k+1$ (with probability p), then goes on to win all with probability w_{k+1}; – explaining the first summand.

Similarly, if the first flip is tail, then A's fortune becomes $k-1$ ...

Step 2. Solving the recursion relation

Recall that $q = 1 - p$, therefore $w_k = pw_k + qw_k$. After substituting this into the left hand side of (*), (*) becomes

$$p[w_{k+1} - w_k] = q[w_k - w_{k-1}]$$

or

$$w_{k+1} - w_k = r[w_k - w_{k-1}]$$

where $r = q/p$.

$$\begin{aligned}
w_2 - w_1 &= r[w_1 - w_0] = rw_1 \\
w_3 - w_2 &= r[w_2 - w_1] = r^2 w_1 \\
&\vdots \quad\quad \vdots \\
w_i - w_{i-1} &= r[w_{i-1} - w_{i-2}] = r^{i-1} w_1 \\
&\vdots \quad\quad \vdots \\
w_N - w_{N-1} &= r[w_{N-1} - w_{N-2}] = r^{N-1} w_1 .
\end{aligned}$$

Adding all of the above,

$$w_N - w_1 = [r + r^2 + \cdots + r^{N-1}]w_1$$

which is

$$\begin{aligned}
1 &= [1 + r + r^2 + \cdots + r^{N-1}]w_1 \\
&= \begin{cases} \frac{1-r^N}{1-r} w_1 & \text{if } r \neq 1 \\ N w_1 & \text{if } r = 1 \end{cases}
\end{aligned}$$

or

$$w_1 = \begin{cases} \frac{1-r}{1-r^N} & \text{if } r \neq 1 \\ \frac{1}{N} & \text{if } r = 1 \end{cases} .$$

Adding the first i equations give us

$$w_i - w_1 = [r + r^2 + \cdots + r^{i-1}]w_1$$

or

$$\begin{aligned}
w_i &= [1 + r + r^2 + \cdots + r^{i-1}]w_1 \\
&= \begin{cases} \frac{1-r^i}{1-r} w_1 & \text{if } r \neq 1 \\ i w_1 & \text{if } r = 1 \end{cases} \\
&= \begin{cases} \frac{1-r^i}{1-r^N} & \text{if } p \neq 1/2 \\ \frac{i}{N} & \text{if } p = 1/2 \end{cases}
\end{aligned}$$

since $r = 1$ is the same as $p = 1/2$.

Remark 12. 1. Let u_i denote the probability that B ends up with all the money when A starts with i and B starts with $N-i$. By symmetry, we get

$$u_i = \begin{cases} \frac{1-(p/q)^{N-i}}{1-(p/q)^N} & \text{if } q \neq 1/2 \\ \frac{N-i}{N} & \text{if } q = 1/2 \end{cases}.$$

2. Further, since $q = \frac{1}{2}$ is equivalent to $p = \frac{1}{2}$, we have, when $q \neq \frac{1}{2}$,

$$\begin{aligned} w_i + u_i &= \frac{1-(q/p)^i}{1-(q/p)^N} + \frac{1-(p/q)^{N-i}}{1-(p/q)^N} \\ &= \frac{p^N - p^N(q/p)^i}{p^N - q^N} + \frac{q^N - q^N(p/q)^{N-i}}{q^N - p^N} \\ &= \frac{p^N - p^{N-i}q^i - q^N + q^i p^{N-i}}{p^N - q^N} = 1. \end{aligned}$$

This result also holds when $p = q = \frac{1}{2}$, i.e.,

$$w_i + u_i = \frac{i}{N} + \frac{N-i}{N} = 1.$$

This equations states that with probability 1, either A or B will wind up with all of the money. In other words, the probability that the game continues indefinitely with A's fortune always being between 1 and $N-1$ is zero. We need to be aware that, a priori, there are three possible outcomes of this gambling game, not two: Either A wins, or B wins, or the game goes on forever with nobody winning. We have just shown that this last event has probability 0.

3. At the end of the game, either A wins all or A is ruined, therefore

$$\begin{aligned} & P(\text{ A is ruined, when he starts with } i\ ) \\ = \ & 1 - w_i \\ = \ & \begin{cases} \frac{r^i - r^N}{1-r^N} & \text{if } p \neq 1/2 \\ \frac{N-i}{N} & \text{if } p = 1/2 \end{cases}. \end{aligned}$$

4. Calculate for different possible cases:

   (a) $i = 10$, $N = 20$ for $p = 0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7$

   (b) $i = 10$, $N = 100$ for $p = 0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7$

   (c) $i = 10$, $N = 1000$ for $p = 0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7$;

- Table below shows probability of ruin of A with initial capital 10.
- Combined capital is N.

	N is		
p	20	100	1000
0.3	0.999791	1.000000	1.000000
0.4	0.982954	1.000000	1.000000
0.45	0.881499	1.000000	1.000000
0.5	0.500000	0.900000	0.990000
0.55	0.118501	0.134431	0.134431
0.6	0.017046	0.017342	0.017342
0.7	0.000209	0.000209	0.000209

5. Suppose A is playing against a very rich opponent B (N is very large).

   For $p = 1/2$, $\lim_{N\to\infty} w_i = 0$.

   And for $p < 1/2$, then $r > 1$ and hence

$$\lim_{N\to\infty} w_i = \lim_{N\to\infty} \frac{1-r^i}{1-r^N} = 0.$$

   And for $p > 1/2$, then $r < 1$ and hence

$$\lim_{N\to\infty} w_i = \lim_{N\to\infty} \frac{1-r^i}{1-r^N} = 1-r^i.$$

   Summarizing, we have

$$\lim_{N\to\infty} w_i = \begin{cases} 1-\left(\frac{q}{p}\right)^i & \text{for } p > 1/2 \\ 0 & \text{for } p \le 1/2 \end{cases}.$$

3.5 $P(\cdot|A)$ is a Probability

Proposition 3.27. *Let A be an event with $P(A) > 0$. Then the following three conditions hold.*

(i) For any event B, we have

$$0 \le P(B|A) \le 1.$$

(ii)

$$P(S|A) = 1.$$

(iii) Let $B_1, B_2, \ldots$ be a sequence of mutually exclusive events,then

$$P(\cup_{k=1}^{\infty} B_k|A) = \sum_{k=1}^{\infty} P(B_k|A).$$

Remark 13. That means, $P(\cdot|A)$ as a function of events satisfy the three axioms of probability. Hence, all of the propositions previously proved for probabilities apply to $P(\cdot|A)$2.

Proof.

(i) As $P(BA) \geq 0, P(A) > 0$, then it follows that

$$P(B|A) = \frac{P(BA)}{P(A)} \geq 0$$

proving the left inequality.

As $BA \subset A$, it follows by Proposition 2.9 in Chapter 2, that $P(BA) \leq P(A)$. So

$$P(B|A) = \frac{P(BA)}{P(A)} \leq 1,$$

proving the second inequality.

(ii) It is easy as

$$P(S|A) = \frac{P(SA)}{P(A)} = \frac{P(A)}{P(A)} = 1.$$

(iii) Let $B_1, B_2, \ldots$ be a sequence of mutually exclusive events

$$\begin{aligned} P(\cup_{k=1}^{\infty} B_k|A) &= \frac{P([\cup_{k=1}^{\infty} B_k] \cap A)}{P(A)} \\ &= \frac{P(\cup_{k=1}^{\infty} AB_k)}{P(A)} \\ &= \frac{\sum_{k=1}^{\infty} P(AB_k)}{P(A)} \quad \text{(why ??)} \\ &= \sum_{k=1}^{\infty} \frac{P(AB_k)}{P(A)} \\ &= \sum_{k=1}^{\infty} P(B_k|A). \end{aligned}$$

□

We return to the matching problem, Example 2.22 of Chapter 2 (page 25), and this time obtain a solution by using conditional probabilities.

Example 3.28. At a party, n men take off their hats. The hats are then mixed up, and each man randomly selects one. We say that a match occurs if a man selects his own hat. What is the probability of no matches?

Solution:
Let E denote the event that no matches occur in the case of n men and write $P_n = P(E)$. We start by conditioning on whether or not the first man selects his own hat – call these events M and M^c, respectively. Then

$$P_n = P(E) = P(E|M)P(M) + P(E|M^c)P(M^c).$$

Clearly $P(E|M) = 0$ so

$$P_n = P(E|M^c) \times \left(1 - \frac{1}{n}\right).$$

Here $P(E|M^c)$ can be interpreted as the probability of no matches when $n-1$ men select from a set of $n-1$ hats that does not contain the hat of one of these men.

This can happen in two mutually exclusive ways:

(a) there are no matches and the extra man does not select the extra hat (this being the hat of the man who chose first), or

(b) there are no matches and the extra man does select the extra hat.

The probability of the first of these events is just P_{n-1}, which is seen by regarding the extra hat as "belonging" to the extra man. Because the second event has probability $\frac{1}{n-1}P_{n-2}$, we have

$$P(E|M^c) = P_{n-1} + \frac{1}{n-1}P_{n-2}.$$

Thus

$$P_n = \frac{n-1}{n}P_{n-1} + \frac{1}{n}P_{n-2},$$

or equivalently,

$$P_n - P_{n-1} = -\frac{1}{n}(P_{n-1} - P_{n-2}).$$

We have

$$P_1 = 0, \quad P_2 = \frac{1}{2}$$

as the initial terms for the recursive equation, and so

$$\begin{aligned} P_3 - P_2 &= -\frac{(P_2 - P_1)}{3} = -\frac{1}{3!} \qquad \text{or} \quad P_3 = \frac{1}{2!} - \frac{1}{3!} \\ P_4 - P_3 &= -\frac{(P_3 - P_2)}{4} = \frac{1}{4!} \qquad \text{or} \quad P_4 = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} \end{aligned}$$

In general,

$$P_n = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots + \frac{(-1)^n}{n!}.$$

Not: The case of exactly k matches is done in Example 3.6.

Addendum

Define $Q(E) = P(E|F)$. Proposition 3.27 says that $Q(E)$ is a probability function defined on the events of S. All of the propositions previously proved for probabilities apply to $Q(E)$.

We have

$$Q(E_1 \cup E_2) = Q(E_1) + Q(E_2) - Q(E_1E_2)$$

or

$$P(E_1 \cup E_2|F) = P(E_1|F) + P(E_2|F) - P(E_1E_2|F).$$

This is the conditional probability version of the inclusion-exclusion principle (Proposition 2.11, page 17).

We can define the conditional probability $Q(E_1|E_2)$ by $Q(E_1|E_2) = Q(E_1E_2)/Q(E_2)$, then we have

$$Q(E_1) = Q(E_1|E_2)Q(E_2) + Q(E_1|E_2^c)Q(E_2^c),$$

analogous to Proposition 3.11 (Baye's First Formula).
Now

$$\begin{aligned} Q(E_1|E_2) &= \frac{Q(E_1E_2)}{Q(E_2)} \\ &= \frac{P(E_1E_2|F)}{P(E_2|F)} \\ &= \frac{\frac{P(E_1E_2F)}{P(F)}}{\frac{P(E_2F)}{P(F)}} = P(E_1|E_2F). \end{aligned}$$

So the above is equivalent to

$$P(E_1|F) = P(E_1|E_2F)P(E_2|F) + P(E_1|E_2^cF)P(E_2^c|F).$$

Chapter 4

Random Variables

4.1 Random Variables

In many situations when an experiment is performed, we are interested in some function of the outcome rather than the outcome itself. Here are some examples:

Example 4.1. When we roll a pair of dice, let's say we are not interested in the numbers that are obtained in each die but we are only interested in the sum of the numbers.

Example 4.2. There are 20 questions in a multiple choice paper. Each question has 5 alternatives. A student answers all 20 questions by randomly and independently choosing one alternative out of 5 in each question. We are interested in $X :=$ number of correct answers.

Definition 4.3. *A **random variable**, X, is a mapping from the sample space to real numbers.*

Example 4.4. Suppose we toss 3 fair coins. Let Y denote the number of heads appearing, then Y takes values 0, 1, 2, 3.
And

$$\begin{aligned} P(Y=0) &= P((T,T,T)) = \frac{1}{8} \\ P(Y=1) &= P((H,T,T),(T,H,T),(T,T,H)) = \frac{3}{8} \\ P(Y=2) &= P((H,H,T),(H,T,H),(T,H,H)) = \frac{3}{8} \\ P(Y=3) &= P((H,H,H)) = \frac{1}{8} \end{aligned}$$

Example 4.5. An urn contains 20 chips numbered from 1 to 20. Three chips are chosen at random from this urn. Let X be the largest number among the three chips drawn. Then X takes values from $3,4,\ldots,20$.
And, for $k = 3,4,\ldots,20$,

$$P(X=k) = \frac{\binom{k-1}{2}}{\binom{20}{3}}.$$

Suppose a game is that, you will win if the largest number obtained is at least 17. What is the probability of winning?

$$\begin{aligned} P(\text{win}) &= P(X=17)+P(X=18)+P(X=19)+P(X=20) \\ &= 0.15+0.134+0.119+0.105 = 0.508. \end{aligned}$$

Example 4.6. Suppose that there are N distinct types of coupons and that each time one obtains a coupon, it is, independently of previous selections, equally likely to be any one of the N types. Let T be the number of coupons that needs to be collected until one obtains a complete set of at least one of each type. Compute $P(T=n)$.

Solution:
Considering the probability $P(T>n)$. Once we know that, $P(T=n)$ will be given as

$$P(T=n) = P(T>n-1) - P(T>n).$$

For a fix n, define the events $A_1, A_2, \ldots, A_N$ as follows: A_j is the event that no type j coupon is contained among the first n coupons collected, $j = 1,\ldots,N$. Hence,

$$P(T>n) = P\left(\bigcup_{j=1}^{N} A_j\right).$$

Note that $\bigcup_{j=1}^{N} A_j$ means that one of the coupons has not been collected during the n collections, which is the same as $\{T>n\}$.
Via way of the inclusion-exclusion principle, we have

$$\begin{aligned} P(T>n) &= P\left(\bigcup_{j=1}^{N} A_j\right) \\ &= \sum_j P(A_j) - \sum_{j_1<j_2} P(A_{j_1}A_{j_2}) + \cdots \\ &\quad + (-1)^{k+1} \sum_{j_1<j_2<\cdots<j_k} P(A_{j_1}A_{j_2}\cdots A_{j_k}) \cdots \\ &\quad + (-1)^{N+1} P(A_1 A_2 \cdots A_N). \end{aligned}$$

A_j occurs if each of the n coupons collected is not of type j. Thus

$$P(A_j) = \left(\frac{N-1}{N}\right)^n.$$

Likewise, the event $A_{j_1}A_{j_2}$ occurs if none of the first n coupons collected is of either type j_1 or type j_2. Thus, again using independence, we see that

$$P(A_{j_1}A_{j_2}) = \left(\frac{N-2}{N}\right)^n.$$

The same reasoning gives

$$P(A_{j_1}A_{j_2}\cdots A_{j_k}) = \left(\frac{N-k}{N}\right)^n.$$

Collating, we see that for $n > 0$,

$$\begin{aligned} P(T > n) &= N\left(\frac{N-1}{N}\right)^n - \binom{N}{2}\left(\frac{N-2}{N}\right)^n + \binom{N}{3}\left(\frac{N-3}{N}\right)^n - \cdots \\ &\quad + (-1)^N \binom{N}{N-1}\left(\frac{1}{N}\right)^n \\ &= \sum_{i=1}^{N-1} \binom{N}{i}\left(\frac{N-i}{N}\right)^n (-1)^{i+1}. \end{aligned}$$

The probability that T equals n can now be obtained from the preceding formula by the use of

$$P(T = n) = P(T > n-1) - P(T > n).$$

Remark 14. One must collect at least N coupons to obtain a compete set, so $P\{T > n\} = 1$ if $n < N$. Therefore we obtain the interesting combinatorial identity that, for integers $1 \le n < N$,

$$\sum_{i=1}^{N-1} \binom{N}{i}\left(\frac{N-i}{N}\right)^n (-1)^{i+1} = 1$$

which can be written as

$$\sum_{i=0}^{N-1} \binom{N}{i}\left(\frac{N-i}{N}\right)^n (-1)^{i+1} = 0.$$

Another random variable of interest is the number of distinct types of coupons that are contained in the first n selections. For more details, refer to [Ross].

4.2 Discrete Random Variables

Definition 4.7. *A random variable is said to be **discrete** if the range of X is either finite or countably infinite.*

Definition 4.8. *Suppose that random variable X is discrete, taking values $x_1, x_2, \ldots$, then the **probability mass function** of X, denoted by p_X (or simply as p if the context is clear), is defined as*

$$p_X(x) = \begin{cases} P(X = x) & \text{if } x = x_1, x_2, \ldots \\ 0 & \text{otherwise} \end{cases}.$$

(Properties of the probability mass function)

(i) $p_X(x_i) \geq 0$; for $i = 1, 2, \ldots$;

(ii) $p_X(x) = 0$; for other values of x;

(iii) Since X must take on one of the values of x_i, $\sum_{i=1}^{\infty} p_X(x_i) = 1$.

Example 4.9. Suppose a random variable X only takes values $0, 1, 2 \ldots$. If the probability mass function of X (p.m.f. of X) is of the form:

$$p(k) = c\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

where $\lambda > 0$ is a fixed positive value and c is a suitably chosen constant.

(a) What is this suitable constant?

(b) Compute $P(X = 0)$ and

(c) $P(X > 2)$.

Solution:

(a) Since

$$\sum_{k=0}^{\infty} p(k) = 1,$$

we have that

$$1 = \sum_{k=0}^{\infty} c\frac{\lambda^k}{k!} = c\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = ce^{\lambda}$$

and so $c = e^{-\lambda}$.

Note: c is known as the **normalising constant**.

(b) It follows that

$$P(X=0) = p(0) = c\frac{\lambda^0}{0!} = c = e^{-\lambda}.$$

(c)

$$\begin{aligned} P(X>2) &= 1 - P(X \leq 2) \\ &= 1 - P(X=0) - P(X=1) - P(X=2) \\ &= 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{\lambda^2}{2} e^{-\lambda} \\ &= 1 - e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right). \end{aligned}$$

Definition 4.10. *The **cumulative distribution function** of X, abbreviated to **distribution function** (d.f.) of X, (denoted as F_X or F if context is clear) is defined as*

$$F_X : \mathbb{R} \longrightarrow \mathbb{R}$$

where

$$F_X(x) = P(X \leq x) \quad \text{for } x \in \mathbb{R}.$$

Remark 15. Suppose that X is discrete and takes values $x_1, x_2, x_3, \ldots$ where $x_1 < x_2 < x_3 < \ldots$. Note then that F is a step function, that is, F is constant in the interval $[x_{i-1}, x_i)$ (F takes value $p(x_1) + \cdots + p(x_{i-1})$), and then take a jump of size $= p(x_i)$.

For instance, if X has a probability mass function given by

$$p(1) = \frac{1}{4}, \quad p(2) = \frac{1}{2}, \quad p(3) = \frac{1}{8}, \quad p(4) = \frac{1}{8},$$

then its cumulative distribution function is

$$F(a) = \begin{cases} 0, & a < 1 \\ \frac{1}{4} & 1 \leq a < 2 \\ \frac{3}{4} & 2 \leq a < 3 \\ \frac{7}{8} & 3 \leq a < 4 \\ 1 & 4 \leq a \end{cases}.$$

Example 4.11. Three fair coins example continued.

$$F_Y(x) = \begin{cases} 0 & x < 0 \\ 1/8 & 0 \le x < 1 \\ 1/2 & 1 \le x < 2 \\ 7/8 & 2 \le x < 3 \\ 1 & 3 \le x \end{cases}.$$

4.3 Expected Value

Definition 4.12. *If X is a discrete random variable having a probability mass function p_X, the **expectation** or the **expected value** of X, denoted by $E(X)$ or μ_X, is defined by*

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

Commonly used notation:
Use U, V, X, Y, Z upper case of letters to denote random variables (for they are actually functions) and use $u, v, \ldots$ lower case of letters to denote values of random variables (values of random variables are just real numbers).

Interpretations of Expectation

See [Ross, pp. 125 – 126, 128]

(i) Weighted average of possible values that X can take on. Weights here are the probability that X assumes it.

(ii) Frequency point of views (relative frequency, in fact).

(iii) Center of gravity.

Example 4.13. Suppose X takes only two values 0 and 1 with

$$P(X = 0) = 1 - p \quad \text{and} \quad P(X = 1) = p.$$

We call this random variable, a **Bernoulli** random variable of parameter p. And we denote it by $X \sim Be(p)$.

$$E(X) = 0 \times (1 - p) + 1 \times p = p.$$

Example 4.14. Let X denote the number obtained when a fair die is rolled. Then, $E(X) = 3.5$.

Solution:
Here X takes values $1, 2, \ldots, 6$ each with probability $1/6$. Hence

$$\begin{aligned} E(X) &= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} \\ &= \frac{1+2+3+4+5+6}{6} \\ &= \frac{7}{2}. \end{aligned}$$

Example 4.15. A newly married couple decides to continue to have children until they have one of each sex. If the events of having a boy and a girl are independent and equiprobable, how many children should this couple expect?

Solution:
Let X be the number of children they should continue to have until they have one of each sex. For $i \geq 2$, $X = i$ if and only if either

(i) all of their first $i-1$ children are boys and the ith child is a girl, or

(ii) all of their first $i-1$ children are girls and the ith child is a boy.

So by independence,

$$P(X = i) = \left(\frac{1}{2}\right)^{i-1} \frac{1}{2} + \left(\frac{1}{2}\right)^{i-1} \frac{1}{2} = \left(\frac{1}{2}\right)^{i-1}, i \geq 2.$$

And

$$E(X) = \sum_{i=2}^{\infty} i \left(\frac{1}{2}\right)^{i-1} = -1 + \sum_{i=1}^{\infty} i \left(\frac{1}{2}\right)^{i-1} = -1 + \frac{1}{(1-1/2)^2} = 3.$$

Note that for $|r| < 1$, $\sum_{i=1}^{\infty} i r^{i-1} = \frac{1}{(1-r)^2}$.

Example 4.16. A contestant on a quiz show is presented with 2 questions, 1 and 2. He can answer them in any order. If he decides on the ith question, and if his answer is correct, then he can go on to the other question; if his answer to the initial question is wrong, then he is not allowed to answer the other. For right answer to question i, he receives v_i. If the probability

that he knows the answer to question i is p_i where $i = 1, 2$, what is the strategy for improving his expected return? Assume independence of answers to each questions.

Solution:

(a) Answer question 1 then question 2:

He will win

$$\begin{cases} 0 & \text{with probability } 1 - p_1 \\ v_1 & \text{with probability } p_1(1 - p_2) \\ v_1 + v_2 & \text{with probability } p_1 p_2 \end{cases}.$$

Hence his expected winnings is

$$v_1 p_1 (1 - p_2) + (v_1 + v_2) p_1 p_2.$$

(b) Answer question 2 then question 1:
He will win

$$\begin{cases} 0 & \text{with probability } 1 - p_2 \\ v_2 & \text{with probability } (1 - p_1) p_2 \\ v_1 + v_2 & \text{with probability } p_1 p_2 \end{cases}.$$

Hence his expected winnings is

$$v_2 p_2 (1 - p_1) + (v_1 + v_2) p_1 p_2.$$

Therefore it is better to answer question 1 first if

$$\begin{aligned} v_1 p_1 (1 - p_2) &\geq v_2 p_2 (1 - p_1) \\ \frac{v_1 p_1}{1 - p_1} &\geq \frac{v_2 p_2}{1 - p_2}. \end{aligned}$$

For example, if he is 60 percent certain of answering question 1, worth \$200, correctly and he is 80 percent certain of answering question 2, worth \$100, correctly, then he should attempt to answer question 2 first because

$$400 = \frac{100 \times 0.8}{0.2} > \frac{200 \times 0.6}{0.4} = 300.$$

4.4 Expectation of a Function of a Random Variable

Given X, we are often interested about $g(X)$ and $E[g(X)]$. How do we compute $E[g(X)]$? One way is to find the probability mass function of $g(X)$ first and proceed to compute $E[g(X)]$ by definition.

Example 4.17. Let X be the length (in m) of a square which we want to paint. Say X is a random variable with a certain distribution. We are interested in cX^2. Here c is the cost of painting per unit m^2. What is the expected cost of painting?

Example 4.18. Let X be a random variable that takes values $-1, 0$ and 1 with probabilities:

$$P(X=-1)=0.2, \quad P(X=0)=0.5 \quad P(X=1)=0.3.$$

We are interested to compute $E(X^2)$.

Solution:
Let $Y=X^2$, it follows that Y takes values 0 and 1 with probabilities

$$\begin{aligned} P(Y=0) &= P(X=0)=0.5 \\ P(Y=1) &= P(X=-1 \text{ or } X=1) \\ &= P(X=-1)+P(X=1)=0.5 \end{aligned}$$

Hence

$$E(X^2)=E(Y)=0\times 0.5+1\times 0.5=0.5.$$

The procedure of going through a new random variable $Y=g(X)$ and find its probability mass function is clumsy. Fortunately we have the following

Proposition 4.19. *If X is a discrete random variable that takes values x_i, $i \geq 1$, with respective probabilities $p_X(x_i)$, then for any real value function g*

$$\begin{aligned} E[g(X)] &= \sum_i g(x_i)p_X(x_i) \quad \textit{or equivalently} \\ &= \sum_x g(x)p_X(x) \end{aligned}$$

Proof. Group together all the terms in $\sum_i g(x_i)p(x_i)$ having the same value of $g(x_i)$. Suppose y_j, $j \geq 1$, represent the different values of $g(x_i)$, $i \geq 1$.

Then, grouping all the $g(x_i)$ having the same value gives

$$\begin{aligned}\sum_i g(x_i)p(x_i) &= \sum_j \sum_{i:g(x_i)=y_j} g(x_i)p(x_i) \\ &= \sum_j \sum_{i:g(x_i)=y_j} y_j p(x_i) \\ &= \sum_j y_j \sum_{i:g(x_i)=y_j} p(x_i) \\ &= \sum_i y_j P(g(X)=y_j) \\ &= E[g(X)].\end{aligned}$$

□

Example 4.20. Let's apply Proposition 4.19 to Example 4.18.

$$E(X^2) = (-1)^2(0.2) + 0^2(0.5) + 1^2(0.3) = 0.5$$

which is in agreement with the result previously computed.

Corollary 4.21. *Let a and b be constants, then*

$$E[aX+b] = aE(X)+b.$$

Proof. Apply $g(x) = ax+b$ in Proposition 4.19, we have

$$\begin{aligned}E[aX+b] &= \sum_x [ax+b]p(x) \\ &= \sum_x [axp(x)+bp(x)] \\ &= a\sum_x xp(x) + b\sum_x p(x) \\ &= aE(X)+b.\end{aligned}$$

□

Example 4.22. Take $g(x) = x^2$. Then,

$$E(X^2) = \sum_x x^2 p_X(x),$$

is called the **second moment** of X.

In general, for $k \geq 1$, $E\left(X^k\right)$ is called the **k-th moment** of X.

Example 4.23. Let $\mu = E(X)$, and take $g(x) = (x-\mu)^k$, then

$$E(X-\mu)^k$$

is called the kth **central** moment.

Remark 16. (a) The expected value of a random variable X, $E(X)$ is also referred to as the **first moment** or the **mean** of X.

(b) The first central moment is 0.

(c) The second central moment, namely,

$$E(X-\mu)^2$$

is called the **variance** of X.

Example 4.24. For an event $A \subset S$, let I_A be the indicator of A, that is,

$$I_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \in S \setminus A \end{cases}.$$

Note that I_A is a Bernoulli random variable with success probability $p = P(A)$. It then follows that

$$E(I_A) = P(A).$$

Theoretically, one can view calculating probability as calculating expectation. We will revisit this idea in Chapter 7.

Example 4.25. A product, sold seasonally, yields a net profit of b dollars for each unit sold and a net loss of l dollars for each unit left unsold when the season ends. The number of this product sold in Peter's store is a random variable, denoted by X with probability mass function $p(i)$, $i \geq 0$. Furthermore, this product has to be pre-ordered. How much should Peter pre-order in order to maximize his profit?

Solution:
Let s be the amount Peter orders. Let Y be the profit (note that, it is a function of s), then

$$Y = \begin{cases} bX - l(s-X) & \text{if } X \leq s \\ sb & \text{if } X > s \end{cases}.$$

Here

$$\begin{aligned}
\phi(s) := E(Y) &= \sum_{i=0}^{s}[bi - l(s-i)]p(i) + \sum_{i=s+1}^{\infty} sbp(i) \\
&= \sum_{i=0}^{s}[(b+l)i - ls]p(i) + \sum_{i=s+1}^{\infty} sbp(i) \\
&= (b+l)\sum_{i=0}^{s} ip(i) - ls\sum_{i=0}^{s} p(i) \\
&\quad + sb\left[1 - \sum_{i=0}^{s} p(i)\right] \\
&= (b+l)\sum_{i=0}^{s} ip(i) - (b+l)s\sum_{i=0}^{s} p(i) + sb \\
&= sb - (b+l)\sum_{i=0}^{s}(s-i)p(i).
\end{aligned}$$

Locating the optimal order:
Consider $\phi(s+1) - \phi(s)$.

$$\begin{aligned}
&\phi(s+1) - \phi(s) \\
= &\left[(s+1)b - (b+l)\sum_{i=0}^{s+1}[(s+1)-i]p(i)\right] \\
&- \left[sb - (b+l)\sum_{i=0}^{s}(s-i)p(i)\right] \\
= &\; b - (b+l)\sum_{i=0}^{s} p(i)
\end{aligned}$$

From this we notice that ϕ first increase and then decrease, let s^* is the change point, that is,

$$\phi(0) < \phi(1) < \cdots < \phi(s^*-1) < \phi(s^*) > \phi(s^*+1) > \cdots.$$

So the optimal order is s^*.

Proposition 4.26 (Tail Sum Formula for Expectation)**.**
For nonnegative integer-valued random variable X (that is, X takes values $0, 1, 2, \ldots$*),*

$$E(X) = \sum_{k=1}^{\infty} P(X \geq k) = \sum_{k=0}^{\infty} P(X > k).$$

Proof. Consider the following triangularization:

$$\begin{array}{lllllllllll}
\displaystyle\sum_{k=1}^{\infty} P(X \geq k) = P(X=1) & + & P(X=2) & + & P(X=3) & + & P(X=4) & + & \cdots \\
& + & P(X=2) & + & P(X=3) & + & P(X=4) & + & \cdots \\
& & & + & P(X=3) & + & P(X=4) & + & \cdots \\
& & & & & + & P(X=4) & + & \cdots \\
& & & & & & & + & \cdots \\
= P(X=1) & + & 2P(X=2) & + & 3P(X=3) & + & 4P(X=4) & + & \cdots \\
= E(X). & & & & & & & &
\end{array}$$

□

4.5 Variance and Standard Deviation

Definition 4.27. *If X is a random variable with mean μ, then the **variance** of X, denoted by $\mathrm{var}(X)$, is defined by*

$$\mathrm{var}(X) = E(X-\mu)^2.$$

It is a measure of scattering (or spread) of the values of X.

Definition 4.28. *The **standard deviation** of X, denoted by σ_X or $\mathrm{SD}(X)$, is defined as*

$$\sigma_X = \sqrt{\mathrm{var}(X)}.$$

An alternative formula for variance:

$$\mathrm{var}(X) = E(X^2) - [E(X)]^2.$$

Proof. (For discrete case only). Here $g(x) = (x-\mu)^2$.

$$\begin{aligned}
E(X-\mu)^2 &= \sum_x (x-\mu)^2 p_X(x) \\
&= \sum_x (x^2 - 2\mu x + \mu^2) p_X(x) \\
&= \sum_x x^2 p_X(x) - \sum_x 2\mu x p_X(x) + \sum_x \mu^2 p_X(x) \\
&= E(X^2) - 2\mu \sum_x x p_X(x) + \mu^2 \sum_x p_X(x) \\
&= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2.
\end{aligned}$$

□

Remark 17. (1) Note that $\text{var}(X) \geq 0$. (Why?)

(2) $\text{var}(X) = 0$ if and only if X is a **degenerate** random variable (that is, the random variable taking only one value, its mean).

(3) It follows from the formula that

$$E(X^2) \geq [E(X)]^2 \geq 0.$$

Example 4.29. Calculate $\text{var}(X)$ if X represents the outcome when a fair die is rolled.

Solution:
It was shown that $E[X] = \frac{7}{2}$. Now

$$E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = 91 \times \frac{1}{6}.$$

So

$$\text{var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$

(Scaling and shifting property of variance and standard deviation:)

(i) $\text{var}(aX+b) = a^2 \text{var}(X)$.

(ii) $\text{SD}(aX+b) = |a| \text{SD}(X)$.

Proof.

(i)

$$\begin{aligned}
\text{var}(aX+b) &= E[(aX+b-E(aX+b))^2] \\
&= E[(aX+b-a\mu-b)^2] \\
&= E[a^2(X-\mu)^2] \\
&= a^2 E[(X-\mu)^2] \\
&= a^2 \text{var}(X).
\end{aligned}$$

(ii) Follows immediately from above. Note that $\sqrt{a^2} = |a|$.

□

4.6 Discrete Random Variables arising from Repeated Trials

We study a mathematical model for repeated trials:

(1) Each trial results in whether a particular event occurs or doesn't. Occurrence of this event is called **success**, and non-occurrence called **failure**. Write $p := P(\text{success})$, and $q := 1 - p = P(\text{failure})$.

Examples:

nature of trial	meaning of success	meaning of failure	probabilities p and q
Flip a fair coin	head	tail	0.5 and 0.5
Roll a fair die	six	non-six	1/6 and 5/6
Roll a pair of fair dice	double six	not double six	1/36 and 35/36
Birth of a child	girl	boy	0.487 and 0.513
Pick an outcome	in A	not in A	$P(A)$ and $1-P(A)$

(2) Each trial with success probability p, failure with $q = 1 - p$;

(3) We repeat the trials independently.

Such trials are called **Bernoulli**(p) **trials**. We now introduce some random variables related to Bernoulli trials:

(a) **Bernoulli random variable**, denoted by Be(p);
We only perform the experiment once, and define

$$X = \begin{cases} 1 & \text{if it is a success} \\ 0 & \text{if it is a failure} \end{cases}.$$

Here

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

and

$$E(X) = p, \quad \text{var}(X) = p(1-p).$$

(b) **Binomial random variable**, denoted by Bin(n, p);
We perform the experiment (under identical conditions and independently) n times and define

$$X = \text{ number of successes in } n \text{ Bernoulli}(p) \text{ trials.}$$

Therefore, X takes values $0, 1, 2, \ldots, n$. In fact, for $0 \leq k \leq n$,

$$P(X = k) = \binom{n}{k} p^k q^{n-k}.$$

Here

$$E(X) = np, \quad \text{var}(X) = np(1-p).$$

Proof.

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= np \sum_{k=1}^{n} \frac{(n-1)!}{[(n-1)-(k-1)]!(k-1)!} p^{k-1} q^{(n-1)-(k-1)} \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{(n-1)-j}, \quad \text{where } j = k-1 \\ &= np. \end{aligned}$$

We will make use of the fact that

$$\text{var}(X) = E(X^2) - (EX)^2 = E(X(X-1)) + E(X) - (EX)^2.$$

Now

$$\begin{aligned} E(X(X-1)) &= \sum_{k=0}^{n} k(k-1) \binom{n}{k} p^k q^{n-k} = \sum_{k=2}^{n} k(k-1) \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= \cdots \\ &= n(n-1)p^2. \end{aligned}$$

So

$$\text{var}(X) = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

□

Computing the Binomial Distribution Function

Suppose that X is binomial with parameters (n, p). The key to computing its distribution function

$$P(X \leq i) = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k}, \quad i = 0, 1, \ldots, n$$

is to utilize the following relationship between $P(X=k+1)$ and $P(X=k)$:

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

Proof.

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

□

(c) **Geometric random variable**, denoted by Geom(p);
Define the random variable

$$\begin{aligned} X &= \text{number of Bernoulli}(p)\text{ trials required} \\ &\quad \text{to obtain the first success.} \end{aligned}$$

(Note, the trial leading to the first success is included.) Here, X takes values $1,2,3,\ldots$ and so on. In fact, for $k\geq 1$,

$$P(X=k)=pq^{k-1}.$$

And

$$E(X)=\frac{1}{p},\quad \text{var}(X)=\frac{1-p}{p^2}.$$

Another version of Geometric distribution:

$$\begin{aligned} X' &= \text{number of failures in the Bernoulli}(p)\text{ trials} \\ &\quad \text{in order to obtain the first success.} \end{aligned}$$

Here

$$X=X'+1.$$

Hence, X' takes values $0,1,2,\ldots$ and

$$P(X'=k)=pq^k,\qquad \text{for } k=0,1,\ldots.$$

And

$$E(X')=\frac{1-p}{p},\quad \text{var}(X')=\frac{1-p}{p^2}.$$

(d) **Negative Binomial random variable**, denoted by NB(r,p);
Define the random variable

$$X = \text{number of Bernoulli}(p) \text{ trials required to obtain } r \text{ success.}$$

Here, X takes values $r, r+1, \ldots$ and so on. In fact, for $k \geq r$,

$$P(X=k) = \binom{k-1}{r-1} p^r q^{k-r}.$$

And

$$E(X) = \frac{r}{p}, \quad \text{var}(X) = \frac{r(1-p)}{p^2}.$$

Remark 18. Take note that $\text{Geom}(p) = \text{NB}(1,p)$.

Example 4.30. A gambler makes a sequence of 1-dollar bets, betting each time on black at roulette at Las Vegas. Here a success is winning 1 dollar and a failure is losing 1 dollar. Since in American roulette the gambler wins if the ball stops on one of 18 out of 38 positions and loses otherwise, the probability of winning is $p = 18/38 = 0.474$.

Example 4.31. A communication system consists of n components, each of which will, independently, function with probability p. The total system will be able to operate effectively if at least one-half of its components function.

(a) For what values of p is a 5-component system more likely to operate effectively than a 3-component system?

(b) In general, when is a $(2k+1)$-component system better than a $(2k-1)$-component system?

Solution:

(a) The number of functioning components X is a binomial random variable with parameters (n,p). The probability that a 5-component system will be effective is

$$\begin{aligned} &P(\text{at least 3 components in a 5-component system functions}) \\ &= \binom{5}{3} p^3 (1-p)^2 + \binom{5}{4} p^4 (1-p) + p^5 \end{aligned}$$

whereas the corresponding probability for a 3-component system is

$$P(\text{at least 2 components in a 3-component system functions}) = \binom{3}{2} p^2(1-p) + p^3.$$

So a 5-component system more likely to operate effectively than a 3-component system when

$$10p^3(1-p)^2 + 5p^4(1-p) + p^5 > 3p^2(1-p) + p^3,$$

which reduces to

$$3(p-1)^2(2p-1) > 0 \iff p > \frac{1}{2}.$$

So p needs to be greater than $\frac{1}{2}$.

(b) Consider a system of $2k+1$-components and let X denote *the number of the first* $2k-1$ *that function*.

A $(2k+1)$-component system will be effective if either

(i) $X \geq k+1$;

(ii) $X = k$ and at least one of the remaining 2 components function; or

(iii) $X = k-1$ and both of the next 2 components function.

So the probability that a $(2k+1)$-component system will be effective is given by

$$P_{2k+1}(\text{effective}) = P(X \geq k+1) + P(X = k)(1-(1-p)^2) + P(X = k-1)p^2.$$

Now the probability that a $(2k-1)$-component system will be effective is given by

$$\begin{aligned} P_{2k-1}(\text{effective}) &= P(X \geq k) \\ &= P(X = k) + P(X \geq k+1). \end{aligned}$$

Thus we need

$$\begin{aligned}
&P_{2k+1}(\text{effective}) - P_{2k-1}(\text{effective}) \\
&\quad = P(X = k-1)p^2 - (1-p)^2 P(X = k) \\
&\quad = \binom{2k-1}{k-1} p^{k-1}(1-p)^k p^2 - (1-p)^2 \binom{2k-1}{k} p^k (1-p)^{k-1} \\
&\quad = \binom{2k-1}{k-1} p^k (1-p)^k [p-(1-p)] > 0 \Longleftrightarrow p > \frac{1}{2}.
\end{aligned}$$

The last equality follows because $\binom{2k-1}{k-1} = \binom{2k-1}{k}$.

So a $2k+1$-components system will be better than a $2k-1$-components system if (and only if) $p > \frac{1}{2}$.

Example 4.32. In a small town, out of 12 accidents that occurred in 1986, four happened on Friday the 13th. Is this a good reason for a superstitious person to argue that Friday the 13th is inauspicious?

Solution:
Suppose the probability that each accident occurs on Friday the 13th is 1/30, just as on any other day. Then the probability of at least four accidents on Friday the 13th is

$$1 - \sum_{i=0}^{3} \binom{12}{i} \left(\frac{1}{30}\right)^i \left(\frac{29}{30}\right)^{12-i} \approx 0.000493.$$

Since this probability is small, this is a good reason for a superstitious person to argue that Friday the 13th is inauspicious.

Example 4.33. The geometric distribution plays an important role in the theory of queues, or waiting lines. For example, suppose a line of customers waits for service at a counter. It is often assumed that, in each small time unit, either 0 or 1 new customers arrive at the counter. The probability that a customer arrives is p and that no customer arrives is $q = 1 - p$. Then the time T until the next arrival has a geometric distribution. It is natural to ask for the probability that no customer arrives in the next k time units, that is, for $P(T > k)$.

Solution:
This is given by

$$\begin{aligned}
P(T > k) &= \sum_{j=k+1}^{\infty} q^{j-1} p \\
&= q^k (p + qp + q^2 p + \cdots) = q^k.
\end{aligned}$$

This probability can also be found by noting that we are asking for no successes (i.e., arrivals) in a sequence of k consecutive time units, where the probability of a success in any one time unit is p. Thus, the probability is just q^k.

Example 4.34 (Banach match problem). At all times, a pipe smoking mathematician carries 2 matchboxes – 1 in his left-hand pocket and 1 in his right-hand pocket. Each time he needs a match, he is equally likely to take it from either pocket. Consider the moment when the mathematician first discovers that one of his matchboxes is empty. If it is assumed that both matchboxes initially contained N matches, what is the probability that there are exactly i matches, $i = 0, 1, 2, \ldots, N$, in the other box?

Solution:
Let E denote the event that the mathematician first discovers that the right-hand matchbox is empty and that there are i matches in the left-hand box at the time. Now, this event will occur if and only if the $(N+1)$th choice of the right-hand matchbox is made at the $(N+1+N-i)$th trial. Hence, using the Negative Binomial formulation, with $p = \frac{1}{2}$, $r = N+1$ and $k = 2N-i+1$, we see that

$$P(E) = \binom{2N-i}{N} \left(\frac{1}{2}\right)^{2N-i+1}.$$

Since there is an equal probability that it is the left-box that is first discovered to be empty and there are i matches in the right-hand box at that time, the desired result is

$$2P(E) = \binom{2N-i}{N} \left(\frac{1}{2}\right)^{2N-i}.$$

4.7 Poisson Random Variable

A random variable X is said to have a **Poisson** distribution with parameter λ if X takes values $0, 1, 2, \ldots$ with probabilities given as:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad \text{for } k = 0, 1, 2, \ldots. \tag{4.1}$$

This defines a probability mass function, since

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

Notation: $X \sim \text{Poisson}(\lambda)$.

The Poisson random variable has a tremendous range of application in diverse areas because it can be used as an approximation for a binomial random variable with parameters (n, p) when n is large and p is small enough so that np is of moderate size. To see this, suppose X is a binomial random variable with parameters (n, p) and let $\lambda = np$. Then

$$\begin{aligned} P(X = k) &= \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \\ &= \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^k} \end{aligned}$$

Note that for n large and λ moderate,

$$\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \frac{n(n-1)\cdots(n-k+1)}{n^k} \approx 1 \text{ and } \left(1 - \frac{\lambda}{n}\right)^k \approx 1.$$

Hence for n large and λ moderate,

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

Remark 19. In other words, if n independent trials, each of which results in a success with probability p are performed, then when n is large and p is small enough to make np moderate, the number of successes occurring is approximately aPoisson random variable with parameter $\lambda = np$.

Some examples of random variables that obey the Poisson probability law, that is, Equation (4.1) are:

(i) number of misprints on a page;

(ii) number of people in a community living to 100 years;

(iii) number of wrong telephone numbers that are dialed in a day;

(iv) number of people entering a store on a given day;

(v) number of particles emitted by a radioactive source;

(vi) number of car accidents in a day;

(vii) number of people having a rare kind of disease.

Each of the preceding, and numerous other random variables, are approximately Poisson for the same reason – because of the Poisson approximation to the binomial.

Example 4.35. Suppose that the number of typographical errors on a page of a book has a Poisson distribution with parameter $\lambda = \frac{1}{2}$. Calculate the probability that there is at least one error on a page.

Solution:
Let X denote the number of errors on the page, we have

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-1/2} = 0.393.$$

Example 4.36. Suppose that the probability that an item produced by a certain machine will be defective is 0.1. Find the probability that a sample of 10 items will contain at most 1 defective item.

Solution:
The desired probability is

$$\binom{10}{0}(0.1)^0(0.9)^{10} + \binom{10}{1}(0.1)^1(0.9)^9 = 0.7361,$$

whereas the Poisson approximation ($\lambda = np = 0.1 \times 10 = 1$) yields the value

$$e^{-1} + e^{-1} \approx 0.7358.$$

Proposition 4.37. *If* $X \sim \text{Poisson}(\lambda)$,

$$E(X) = \lambda, \quad \text{var}(X) = \lambda.$$

Proof. We have

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

and

$$\begin{aligned} E(X(X-1)) &= \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \\ &= \lambda^2 e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda^2. \end{aligned}$$

Note that

$$\text{var}(X) = E(X^2) - (E(X))^2 = E(X(X-1)) + E(X) - (E(X))^2,$$

and so we have $\text{var}(X) = \lambda$. □

Remark 20. The Poisson distribution with parameter np is a very good approximation to the distribution of the number of successes in n independent trials when each trial has probability p of being a success, provided that n is large and p small. In fact, it remains a good approximation even when the trials are not independent, provided that their dependence is weak.

Example 4.38. Consider again Example 2.22 of Chapter 2 (page 25) (The matching problem).
Define the events $E_i, i = 1, \ldots, n$, by

$$E_i = \{\text{person } i \text{ selects his own hat}\}.$$

It is easy to see that

$$X = \text{number of matches} = E_1 + \cdots + E_n$$

and that

$$P(E_i) = \frac{1}{n} \quad \text{and} \quad P(E_i|E_j) = \frac{1}{n-1}, j \neq i$$

showing that though the events $E_i, i = 1, \ldots, n$ are not independent, their dependence, for large n, is weak. So it is reasonable to expect that the number of matches X will approximately have a Poisson distribution with parameter $n \times 1/n = 1$, and thus

$$P(X = 0) \approx e^{-1} = 0.37$$

which agrees with what was shown in Chapter 2.

For a second illustration of the strength of the Poisson approximation when the trials are weakly dependent, let us consider again the birthday problem presented in Example 2.19 of Chapter 2 (page 22).

Example 4.39. Suppose that each of n people is equally likely to have any of the 365 days of the year as his or her birthday, the problem is to determine the probability that a set of n independent people all have different birthdays. This probability was shown to be less than $\frac{1}{2}$ when $n = 23$.

We can approximate the preceding probability by using the Poisson approximation as follows:

Imagine that we have a trial for each of the $\binom{n}{2}$ pairs of individuals i and j, $i \neq j$, and say that trial i, j is a success if persons i and j have the same birthday.

Let E_{ij} denote the event that trial i, j is a success. Then the events $E_{ij}, 1 \leq i < j \leq n$, are not independent and their dependence is weak.

Now $P(E_{ij}) = 1/365$, it is reasonable to suppose that the number of successes should approximately have a Poisson distribution with mean $\lambda = \binom{n}{2}/365 = n(n-1)/730$. Therefore,

$$\begin{aligned} P(\text{no 2 people have the same birthday}) &= P(0 \text{ successes}) \\ &\approx e^{-\lambda} \\ &= \exp\left(\frac{-n(n-1)}{730}\right). \end{aligned}$$

We want the smallest integer n so that

$$\exp\left(\frac{-n(n-1)}{730}\right) \leq \frac{1}{2}.$$

This can be solved to yield $n = 23$, in agreement with the result in Chapter 2.

Remark 21. For the number of events to occur to approximately have a Poisson distribution, it is not essential that all the events have the same probability of occurrence, but only that all of these probabilities be small.

(Poisson Paradigm)
Consider n events, with p_i equal to the probability that event i occurs, $i = 1$, ..., n. If all the p_i are "small"' and the trials are either independent or at most "weakly dependent", then the number of these events that occur approximately has a Poisson distribution with mean $\sum_{i=1}^{n} p_i$.

Another use of the Poisson distribution arises in situations where events occur at certain points in time. One example is to designate the occurrence of an earthquake as an event; another possibility would be for events to correspond to people entering a particular establishment; and a third possibility is for an event to occur whenever a war starts. Let us suppose that events are indeed occurring at certain random points of time, and let us assume that, for some positive constant λ, the following assumptions hold true:

1. The probability that exactly 1 event occurs in a given interval of length h is equal to $\lambda h + o(h)$, where $o(h)$ stands for any function $f(h)$ for which $\lim_{h\to 0} f(h)/h = 0$. [1]

2. The probability that 2 or more events occur in an interval of length h is equal to $o(h)$.

3. For any intergers n, $j_1, j_2, \ldots, j_n$ and any set of n non-overlapping intervals, if we define E_i to be the event that exactly j_i of the events under consideration occur in ith of these intervals, then the events $E_1, E_2, \ldots, E_n$ are independent.

It can be shown that under the assumptions mentioned above, the number of events occurring in any interval of length t is a Poisson random variable with parameter λt, and we say that the events occur in accordance with a Poisson process with rate λ.

Example 4.40. Suppose that earthquakes occur in the western portion of the United States in accordance with assumptions 1, 2, and 3, with $\lambda = 2$ and with 1 week as the unit of time. (That is, earthquakes occur in accordance with the three assumptions at a rate of 2 per week.)

(a) Find the probability that at least 3 earthquakes occur during the next 2 weeks.

(b) Find the probability distribution of the time, starting from now, until the next earthquake.

Solution:
Let $N(t)$ denote the number of earthquakes that occur in t weeks.

(a) Then $N(2) \sim \text{Poisson}(4)$ and

$$P(N(2) \geq 3) = 1 - P(N(2) \leq 2) = 1 - e^{-4} - 4e^{-4} - \frac{4^2}{2}e^{-4} = 1 - 13e^{-4}.$$

[1] For instance, $f(h) = h^2$ is $o(h)$, whereas $f(h) = h$ is not.

(b) Let X denote the amount of time (in weeks) until the next earthquake. Because X will be greater than t if and only if no events occur with the next t units of time, we have

$$P(X > t) = P(N(t) = 0) = e^{-\lambda t}$$

so the required probability distribution function F is given by

$$F(t) = P(X \leq t) = 1 - e^{-\lambda t} = 1 - e^{-2t}.$$

Computing the Poisson Distribution Function

If X is Poisson with parameter λ, then

$$\frac{P(X = i+1)}{P(X = i)} = \frac{e^{-\lambda}\lambda^{i+1}/(i+1)!}{e^{-\lambda}\lambda^{i}/i!} = \frac{\lambda}{i+1},$$

which means that

$$P(X = i+1) = \frac{\lambda}{i+1} P(X = i).$$

This gives a nice recursive way of computing the Poisson probabilities:

$$\begin{aligned} P(X = 0) &= e^{-\lambda} \\ P(X = 1) &= \lambda P(X = 0) = \lambda e^{-\lambda} \\ P(X = 2) &= \frac{\lambda}{2} P\{X = 1\} = \frac{\lambda^2}{2} e^{-\lambda} \\ &\vdots \end{aligned}$$

4.8 Hypergeometric Random Variable

Suppose that we have a set of N balls, of which m are red and $N - m$ are blue. We choose n of these balls, *without replacement*, and define X to be the number of red balls in our sample. Then

$$P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}},$$

for $x = 0, 1, \ldots, N$.

A random variable whose probability mass function is given as the above equation for some values of n, N, m is said to be a **hypergeometric** random variable, and is denoted by $H(n,N,m)$. Here

$$E(X) = \frac{nm}{N}, \quad \text{var}(X) = \frac{nm}{N}\left[\frac{(n-1)(m-1)}{N-1} + 1 - \frac{nm}{N}\right].$$

Example 4.41. A purchaser of electrical components buys them in lots of size 10. It is his policy to inspect 3 components randomly from a lot and to accept the lot only if all 3 are nondefective. If 30 percent of the lots have 4 defective components and 70 percent have only 1, what proportion of lots does the purchaser reject?

Solution:
Let A denote the event that the purchaser accepts a lot. Then

$$\begin{aligned} P(A) &= P(A|\text{lot has 4 defectives}) \times \frac{3}{10} + P(A|\text{lot has 1 defective}) \times \frac{7}{10} \\ &= \frac{\binom{4}{0}\binom{6}{3}}{\binom{10}{3}} \times \frac{3}{10} + \frac{\binom{1}{0}\binom{9}{3}}{\binom{10}{3}} \times \frac{7}{10} = \frac{54}{100}. \end{aligned}$$

Hence, 46 percent of the lots are rejected.

4.9 Expected Value Of Sums Of Random Variables

In this section, we will prove the result that the expected value of a sum of random variables is equal to the sum of their expectations.

For a random variable X, let $X(s)$ denote the value of X when $s \in S$ is the outcome. Now, if X and Y are both random variables, then so is their sum. That is, $Z = X + Y$ is also a random variable. Moreover, $Z(s) = X(s) + Y(s)$.

Example 4.42. Suppose that an experiment consists of flipping a coin 5 times, with the outcome being the resulting sequence of heads and tails. Let X be the number of heads in the first 3 flips and Y the number of heads in the final 2 flips. Let $Z = X + Y$. Then for the outcome $s = (h,t,h,t,h)$,

$$X(s) = 2, \quad Y(s) = 1, \quad Z(s) = X(s) + Y(s) = 3.$$

For the outcome $s = (h,h,h,t,h)$,

$$X(s) = 3, \quad Y(s) = 1, \quad Z(s) = X(s) + Y(s) = 4.$$

Let $p(s) = P(\{s\})$ be the probability that s is the outcome of the experiment.

Proposition 4.43.

$$E[X] = \sum_{s \in S} X(s)p(s).$$

Proof. Suppose that the distinct values of X are $x_j, i \geq 1$. For each i, let S_i be the event that X is equal to x_i. That is, $S_i = \{s : X(s) = x_i\}$. Then,

$$\begin{aligned} E[X] &= \sum_i x_i P\{X = x_i\} \\ &= \sum_i x_i P(S_i) \\ &= \sum_i x_i \sum_{s \in S_i} p(s) \\ &= \sum_i \sum_{s \in S_i} x_i p(s) \\ &= \sum_i \sum_{s \in S_i} X(s)p(s) \\ &= \sum_{s \in S} X(s)p(s). \end{aligned}$$

The final equality follows because $S_1, S_2, \ldots$ are mutually exclusive events whose union is S. □

Example 4.44. Suppose that two independent flips of a coin that comes up heads with probability p are made, and let X denote the number of heads obtained. Now

$$\begin{aligned} &P(X = 0) = P(t,t) = (1-p)^2, \\ &P(X = 1) = P(h,t) + P(t,h) = 2p(1-p), \\ &P(X = 2) = P(h,h) = p^2 \end{aligned}$$

It follows from the definition that

$$E[X] = 0 \times (1-p)^2 + 1 \times 2p(1-p) + 2 \times p^2 = 2p$$

which agrees with

$$\begin{aligned} E[X] &= X(h,h)p^2 + X(h,t)p(1-p) + X(t,h)(1-p)p + X(t,t)(1-p)^2 \\ &= 2p^2 + p(1-p) + (1-p)p \\ &= 2p. \end{aligned}$$

As a consequence of Proposition 4.43, we have the important and useful result that the expected value of a sum of random variables is equal to the sum of their expectations.

Proposition 4.45. *For random variables $X_1, X_2, \ldots, X_n$,*

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right].$$

Proof. Let $Z = \sum_{i=1}^{n} X_i$. Proposition 4.43 gives

$$\begin{aligned} E[Z] &= \sum_{s\in S} Z(s)p(s) \\ &= \sum_{s\in S} (X_1(s) + X_2(s) + \cdots + X_n(s))p(s) \\ &= \sum_{s\in S} X_1(s)p(s) + \sum_{s\in S} X_2(s)p(s) + \cdots + \sum_{s\in S} X_n(s)p(s) \\ &= E[X_1] + E[X_2] + \cdots + E[X_n] \end{aligned}$$

□

Remark 22. Note that this result does not require that the X_i's be independent.

Example 4.46. Find the expected total number of successes that result from n trials when trial i is a success with probability $p_i, i = 1, \ldots, n$.

Solution:
Let

$$X_i = \begin{cases} 1, & \text{if trial } i \text{ is a success} \\ 0, & \text{if trial } i \text{ is a failure} \end{cases}.$$

We have the representation

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

Consequently,

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p_i$$

Example 4.47. For $X \sim \text{Bin}(n, p)$, find $\text{var}(X)$.

Solution:
Note that from the previous example, $E(X) = np$.

Let X_i be define as in the previous example. Then

$$\begin{aligned} E[X^2] &= E\left[\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} X_j\right)\right] \\ &= E\left[\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n}\sum_{j\neq i} X_i X_j\right] \\ &= \sum_{i=1}^{n} E[X_i^2] + \sum_{i=1}^{n}\sum_{j\neq i} E[X_i X_j] \\ &= \sum_{i} p_i + \sum_{i=1}^{n}\sum_{j\neq i} E[X_i X_j]. \end{aligned}$$

Note that $X_i^2 = X_i$ and that

$$X_i X_j = \begin{cases} 1, & \text{if } X_i = 1 \text{ and } X_j = 1. \\ 0, & \text{otherwise} \end{cases}$$

Hence,

$$E[X_i X_j] = P(X_i = 1, X_j = 1) = P(\text{trials } i \text{ and } j \text{ are successes}).$$

Note that if X is binomial, then for $i \neq j$, the results of trial i and trial j are independent, with each being a success with probability p. Therefore,

$$E[X_i X_j] = p^2, i \neq j.$$

Thus

$$E[X^2] = np + n(n-1)p^2,$$

which gives

$$\text{var}(X) = E[X^2] - (E[X])^2 = np + n(n-1)p^2 - n^2p^2 = np(1-p).$$

4.10 Distribution Functions and Probability Mass Functions

Let X be a discrete random variable. Recall the definitions for the distribution function (d.f.) and the probability mass function (p.m.f.) of X:

(1) For distribution function, $F_X : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

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

(2) For probability mass function, $p_X : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$p_X(x) = P(X = x).$$

Properties of distribution function

(i) F_X is a nondecreasing function, i.e., if $a < b$, then $F_X(a) \leq F_X(b)$.

(ii) $\lim_{b\to\infty} F_X(b) = 1$.

(iii) $\lim_{b\to-\infty} F_X(b) = 0$.

(iv) F_X is right continuous. That is, for any $b \in \mathbb{R}$

$$\lim_{x\to b^+} F_X(x) = F_X(b).$$

Proof.

(i) Note that for $a < b$, $\{X \leq a\} \subset \{X \leq b\}$ and so the result follows from Proposition 2.9 (page 16).

(ii) For $b_n \uparrow \infty$, the events $\{X \leq b_n\}$, $n \geq 1$, are increasing events whose union is the event $\{X < \infty\}$. By Proposition 2.25 (page 27),

$$\lim_{n\to\infty} P(X \leq b_n) = P(X < \infty) = 1.$$

(iii) Similar to the above. Consider $b_n \downarrow -\infty$ and the events $\{X \leq b_n\}$, $n \geq 1$.

(iv) If b_n decreases to b, then $\{X \leq b_n\}$, $n \geq 1$, are decreasing events whose intersection is $\{X \leq b\}$. Again, Proposition 2.25 yields

$$\lim_{n\to\infty} P(X \leq b_n) = P(X \leq b).$$

□

Some useful calculations

Theoretically, all probability questions about X can be computed in terms of density function (or probability mass function).

(1) **Calculating probabilities from density function**

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

Proof. Note that $\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}$ so

$$P(X \leq b) = P(X \leq a) + P(a < X \leq b).$$

Rearrangement yields the result. □

(ii) $P(X < b) = \lim_{n\to\infty} F\left(b - \frac{1}{n}\right)$.

Proof.

$$\begin{aligned} P(X < b) &= P\left(\lim_{n\to\infty}\left\{X \leq b - \frac{1}{n}\right\}\right) \\ &= \lim_{n\to\infty} P\left(X \leq b - \frac{1}{n}\right) \\ &= \lim_{n\to\infty} F\left(b - \frac{1}{n}\right). \end{aligned}$$

Note that $P\{X < b\}$ does not necessarily equal $F(b)$, since $F(b)$ also includes the probability that X equals b. □

(iii) $P(X = a) = F_X(a) - F_X(a^-)$ where $F_X(a^-) = \lim_{x\to a^-} F_X(x)$.

(iv) Using the above, we can compute $P(a \leq X \leq b)$; $P(a \leq X < b)$ and $P(a < X < b)$. For example,

$$\begin{aligned} P(a \leq X \leq b) &= P(X = a) + P(a < X \leq b) \\ &= F_X(a) - F_X(a^-) + [F_X(b) - F_X(a)] \\ &= F_X(b) - F_X(a^-). \end{aligned}$$

Similarly for the other two.

(2) **Calculating probabilities from probability mass function**

$$P(A) = \sum_{x \in A} p_X(x).$$

(3) **Calculate probability mass function from density function**

$$p_X(x) = F_X(x) - F_X(x^-), \quad x \in \mathbb{R}.$$

(4) **Calculate density function from probability mass function**

$$F_X(x) = \sum_{y \leq x} p_X(y) \quad x \in \mathbb{R}.$$

Example 4.48. The distribution function of the random variable X is given by

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{x}{2}, & 0 \leq x < 1, \\ \frac{2}{3}, & 1 \leq x < 2, \\ \frac{11}{12}, & 2 \leq x < 3, \\ 1, & 3 \leq x. \end{cases}$$

Compute

(a) $P(X < 3)$,

(b) $P(X = 1)$,

(c) $P\left(X > \frac{1}{2}\right)$,

(d) $P(2 < X \leq 4)$.

Solution:

(a) $P(X < 3) = \lim_n P\left(X \leq 3 - \frac{1}{n}\right) = \lim_n F\left(3 - \frac{1}{n}\right) = \lim_n \frac{11}{12} = \frac{11}{12}$.

(b) $P(X = 1) = F(1) - \lim_n F\left(1 - \frac{1}{n}\right) = \frac{2}{3} - \lim_n \frac{1 - \frac{1}{n}}{2} = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$.

(c) $P\left(X > \frac{1}{2}\right) = 1 - P\left(X \leq \frac{1}{2}\right) = 1 - F\left(\frac{1}{2}\right) = \frac{3}{4}$.

(d) $P(2 < X \leq 4) = F(4) - F(2) = \frac{1}{12}$.

Chapter 5

Continuous Random Variables

5.1 Introduction

There are many "natural" random variables whose set of possible values is *uncountable*. For example, consider (i) random variable X which denotes the lifetime of an electrical appliance or (ii) random variable Y which denotes the amount of rainfall we get in a month.

Definition 5.1. *We say that X is a **continuous** random variable if there exists a nonnegative function f_X, defined for all real $x \in \mathbb{R}$, having the property that, for any set B of real numbers,*

$$P(X \in B) = \int_B f_X(x)\,\mathrm{dx}\,.$$

*The function f_X is called the **probability density function** (p.d.f.) of the random variable X.*

All probability statements about X can be answered in terms of f_X. For instance, letting $B = [a,b]$, we obtain

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{dx}\,. \qquad (*)$$

Definition 5.2. *We define the **distribution function** of X by*

$$F_X(x) = P(X \leq x), \quad \textit{for } x \in \mathbb{R}.$$

Remark 23. Note that the definition for distribution function is the same for discrete and continuous random variables. Therefore, in the context of continuous random variable,

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{dt}, \;\; x \in \mathbb{R} \qquad (1)$$

and, using the Fundamental Theorem of Calculus,

$$F_X'(x) = f_X(x), \;\; x \in \mathbb{R}. \tag{2}$$

That is, the density is the derivative of the cumulative distribution function. A more intuitive interpretation:

$$P\left(x - \frac{\varepsilon}{2} \leq X \leq x + \frac{\varepsilon}{2}\right) = \int_{x-\varepsilon/2}^{x+\varepsilon/2} f_X(x)\,\mathrm{dx} \approx \varepsilon f(x),$$

when ε is small and when $f(\cdot)$ is continuous at x.

The probability that X will be contained in an interval of length ε around the point x is approximately $\varepsilon f(x)$. From this result we see that $f(x)$ is a measure of how likely it is that the random variable will be near x.

(i) By taking $a \uparrow x$ and $b \downarrow x$ in (*), we see that

$$P(X = x) = 0$$

for any $x \in (-\infty, \infty)$.

(ii) The distribution function, F_X, is continuous.

(iii) For any $a, b \in (-\infty, \infty)$,

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X \leq b) \\ &= P(a \leq X < b) \\ &= P(a < X < b). \end{aligned}$$

(iv) Equations (1) and (2) (in the continuous setting) are to be compared with a discrete random variable Y.

$$F_Y(x) = \sum_{-\infty < t \leq x} p_Y(t), \;\; x \in \mathbb{R} \tag{a}$$

and

$$p_Y(x) = F_Y(x) - F_Y(x-), \;\; x \in \mathbb{R}. \tag{b}$$

Determining the constant in the probability density function

Taking $a \to -\infty$ and $b \to \infty$ in (*), we get

$$1 = P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x)\,\mathrm{dx}\,. \qquad (**)$$

Example 5.3. Suppose that X is a continuous random variable whose probability density function is of the form:

$$f_X(x) = \begin{cases} c(4x-2x^2) & 0 < x < 2 \\ 0 & \text{otherwise}\,. \end{cases}$$

(i) What is the value of c?

(ii) Compute $P(X > 1)$.

Solution:

(i)

$$1 = \int_{-\infty}^{\infty} f_X(x)\,\mathrm{dx} = c\int_0^2 (4x-2x^2)\,\mathrm{dx} = \frac{8c}{3}.$$

Therefore $c = 3/8$.

(ii)

$$P(X > 1) = \int_1^{\infty} f_X(x)\,\mathrm{dx} = \frac{3}{8}\int_1^2 (4x-2x^2)\,\mathrm{dx} = \frac{1}{2}.$$

Example 5.4. An electrical appliance will function for a random amount of time T. If the probability density function of T is given in the form (where $\lambda >$ is a fixed constant):

$$f_T(t) = \begin{cases} ce^{-\lambda t} & 0 < t < \infty \\ 0 & \text{otherwise}\,. \end{cases}$$

(i) What is the value of c?

(ii) Show that for any $s,t > 0$, then

$$P(T > s+t|T > s) = P(T > t).$$

Remark 24. (i) The distribution of T is commonly called a **exponential distribution** with parameter λ.

(ii) This particular property is called the **memoryless property**.

Solution:

(i)

$$1 = \int_{-\infty}^{\infty} f_T(t)\,\mathrm{dt} = c\int_0^{\infty} e^{-\lambda t}\,\mathrm{dt} = \frac{c}{\lambda}.$$

Hence $c = \lambda$.

(ii) First consider

$$P(T > t) = \int_t^{\infty} f_T(x)\,\mathrm{dx} = \lambda\int_t^{\infty} e^{-\lambda x}\,\mathrm{dx} = e^{-\lambda t}.$$

Therefore

$$\begin{aligned} P(T > s+t|T > s) &= \frac{P(T > s+t; T > s)}{P(T > s)} \\ &= \frac{P(T > s+t)}{P(T > s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \\ &= P(T > t). \end{aligned}$$

Remark 25. From the calculation in (ii) above, we obtain

$$F_T(t) = \begin{cases} 1 - e^{-\lambda t}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

Example 5.5. The lifetime in hours of a certain kind of radio tube is a random variable having a probability density function given by

$$f(x) = \begin{cases} 0, & x \leq 100 \\ 100/x^2, & x > 100 \end{cases}.$$

What is the probability that exactly 2 of 5 such tubes in a radio will have to be replaced within the first 150 hours of operation? Assume that the radio tubes function independently.

Solution:
For $1 \leq i \leq 5$, let E_i be the event that the ith radio tube has to be replaced within 150 hours of operation.
Let X be the number of radio tubes (out of 5) that have to be replaced within 150 hours of operation. Then

$$X \sim \text{Bin}(5, p)$$

where $p = P(E_1)$.
We are asked $P(X = 2)$. This is given as

$$\binom{5}{2} p^2 (1-p)^3.$$

Let T_i be the lifetime of the radio tube i. We need to compute

$$p = P(T_1 \leq 150) = \int_{-\infty}^{150} f(t)\,\mathrm{dt} = \int_{100}^{150} \frac{100}{t^2}\,\mathrm{dt} = \frac{1}{3}.$$

Probability asked is

$$\binom{5}{2} (1/3)^2 (1-1/3)^3 = 80/243 = 0.3292.$$

5.2 Expectation and Variance of Continuous Random Variables

Definition 5.6. *Let X be a continuous random variable with probability density function f_X, then*

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{dx} \quad \textit{and} \\ \text{var}(X) &= \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x)\,\mathrm{dx}\,. \end{aligned}$$

A number of results in the discrete case either carry over or with obvious change to the continuous case, namely:

Proposition 5.7. *If X is a continuous random variable with probability density function f_X, then for any real value function g*

(a)

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{dx}\,.$$

(b) Same linearity Property:

$$E(aX+b) = aE(X)+b.$$

(c) Same alternative formula for variance:

$$\mathrm{var}(X) = E(X^2) - [E(X)]^2.$$

Lemma 5.8 (Tail sum formula)**.** *Suppose X is a nonnegative continuous random variable, then*

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

Proof. We have

$$\int_0^{\infty} P(X > x)\,\mathrm{dx} = \int_0^{\infty} \int_x^{\infty} f(y)\,\mathrm{dy}\,\mathrm{dx}$$

where we have used the fact that $P(X > x) = \int_x^{\infty} f(y)\,\mathrm{dy}$. Interchanging the order of integration we get

$$\int_0^{\infty} P(X > x)\,\mathrm{dx} = \int_0^{\infty} \left(\int_0^{y} \mathrm{dx} \right) f(y)\,\mathrm{dy} = \int_0^{\infty} y f(y)\,\mathrm{dy} = E[X].$$

□

Example 5.9. Find the mean and the variance of the random variable, X, with the probability density function given as

$$f(x) = \begin{cases} 1/(b-a), & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}.$$

Here $a < b$ are given constants. X is said to be a **uniform distribution** over $[a,b]$.

Solution:
Mean of X:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{dx} = \int_a^b \frac{x}{b-a}\,\mathrm{dx} = \frac{a+b}{2}.$$

Variance of X:

$$\begin{aligned}
\text{var}(X) &= E(X^2) - \left(\frac{a+b}{2}\right)^2 \\
&= \int_a^b \frac{x^2}{b-a}\,\text{dx} - \frac{(a+b)^2}{4} \\
&= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} \\
&= \frac{b^2+ab+a^2}{3} - \frac{(a+b)^2}{4} \\
&= \frac{(b-a)^2}{12}.
\end{aligned}$$

Example 5.10. Find $E\left(e^X\right)$ where X has probability density function given by

$$f(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}.$$

Solution:

$$E\left(e^X\right) = \int_0^1 e^x 2x\,\text{dx} = 2\left[(x-1)e^x\right]_0^1 = 2.$$

Example 5.11. We have

(a) $\text{var}(aX+b) = a^2\,\text{var}(X)$.

(b) $\text{SD}(aX+b) = |a|\text{SD}(X)$.

Example 5.12. A stick of length 1 is broken at random. Determine the expected length of the piece that contains the point of length p from one end, call A, where $0 \le p \le 1$.

Solution:
Let U be the point (measuring from A) where the stick is broken into two pieces. Then the probability density function of U is given by

$$f_U(u) = \begin{cases} 1, & \text{for } 0 < u < 1 \\ 0, & \text{otherwise} \end{cases}.$$

Let $L_p(U)$ denote the length of the piece which contains the point p, and note that

$$L_p(U) = \begin{cases} 1-U, & \text{if } U < p \\ U, & \text{if } U > p \end{cases}.$$

Hence from Proposition 5.7

$$\begin{aligned} E[L_p(U)] &= \int_{-\infty}^{\infty} L_p(u) \cdot f_U(u) du \\ &= \int_0^1 L_p(u) \cdot 1 du \\ &= \int_0^p (1-u) du + \int_p^1 u du \\ &= 1/2 + p(1-p). \end{aligned}$$

Example 5.13. Suppose that if you are s minutes early for an appointment, then you incur a cost cs, and if you are s minutes late, then you incur a cost ks. Suppose the travelling time from where you are to the location of your appointment is a continuous random variable having probability density function f. Determine the time at which you should depart if you want to minimize your expected cost.

Solution:
Let X be your travel time. If you leave t minutes before your appointment, then the cost, call it $C_t(X)$, is given by

$$C_t(X) = \begin{cases} c(t-X), & \text{if } X \le t \\ k(X-t), & \text{if } X > t \end{cases}.$$

Therefore,

$$\begin{aligned} E[C_t(X)] &= \int_0^{\infty} C_t(x) f(x)\,\mathrm{dx} \\ &= \int_0^t c(t-x) f(x)\,\mathrm{dx} + \int_t^{\infty} k(x-t) f(x)\,\mathrm{dx} \\ &= ct \int_0^t f(x)\,\mathrm{dx} - c\int_0^t x f(x)\,\mathrm{dx} + k\int_t^{\infty} x f(x)\,\mathrm{dx} - kt\int_t^{\infty} f(x)\,\mathrm{dx} \\ &= ctF(t) - c\int_0^t x f(x)\,\mathrm{dx} + k\int_t^{\infty} x f(x)\,\mathrm{dx} - kt[1-F(t)]. \end{aligned}$$

The value of t which minimizes $E[C_t(X)]$ is obtained by calculus. Differentiation yields

$$\begin{aligned} \frac{d}{\mathrm{dt}} E[C_t(X)] &= cF(t) + ctf(t) - ctf(t) - ktf(t) - k[1-F(t)] + ktf(t) \\ &= (k+c)F(t) - k. \end{aligned}$$

Equating to 0, the minimal expected cost is obtained when you leave t^* minutes before your appointment, where t^* satisfies

$$F(t^*) = \frac{k}{k+c}$$

that is $t^* = F^{-1}(\frac{k}{k+c})$ if F^{-1} exists.
How do we know that this t^* gives us a minimum and not a maximum?

5.3 Uniform Distribution

A random variable X is said to be **uniformly** distributed over the interval $(0,1)$ if its probability density function is given by

$$f_X(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}.$$

We denote this by $X \sim U(0,1)$.
Finding F_X:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } 0 \le x < 1 \\ 1, & \text{if } 1 \le x \end{cases}.$$

In general, for $a < b$, we say that a random variable X is **uniformly** distributed over the interval (a,b) if its probability density function is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}.$$

We denote this by $X \sim U(a,b)$.
In a similar way,

$$F_X(x) \quad = \quad \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y = \begin{cases} 0, & \text{if } x < a \\ (x-a)/(b-a), & \text{if } a \le x < b \\ 1, & \text{if } b \le x \end{cases}.$$

It was shown that

$$E(X) = \frac{a+b}{2} \quad \text{and} \quad \text{var}(X) = \frac{(b-a)^2}{12}.$$

Example 5.14. Buses arrive at a specified stop at 15-minute intervals starting at 7 am. That is, they arrive at 7, 7:15, 7:30, 7:45, and so on. If a passenger arrives at the stop at a time that is uniformly distributed between 7 and 7:30, find the probability that he waits

(i) less than 5 minutes for a bus;

(ii) more than 10 minutes for a bus.

Solution:
Let X denote the arrival time of the passenger (after 7 in minutes). Then,

$$X \sim U(0,30).$$

(i) The passenger waits less than 5 minutes for a bus when and only when he arrives (a) between 7:10-7:15 or (b) 7:25-7:30. So the desired probability is

$$P(10 < X < 15) + P(25 < X < 30) = \frac{15-10}{30} + \frac{30-25}{30} = \frac{1}{3}.$$

(ii) The passenger waits more than 10 minutes for a bus when and only when he arrives (a) between 7:00-7:05 or (b) 7:15-7:20. So the desired probability is

$$P(0 < X < 5) + P(15 < X < 20) = \frac{5-0}{30} + \frac{20-15}{30} = \frac{1}{3}.$$

5.4 Normal Distribution

A random variable X is said to be **normally** distributed with parameters μ and σ^2 if its probability density function is given by

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

We denote this by $X \sim N(\mu, \sigma^2)$.
Note that, this density function is **bell-shaped**, always positive, symmetric at μ and attains its maximum at $x = \mu$.

Let's verify that $f(x)$ is indeed a probability density function, that is,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \, \mathrm{dx} = 1.$$

To do so, make the substitution $y = (x-\mu)/\sigma$ to get

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}\,\mathrm{dx} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{dy}\,.$$

Thus we must show that

$$\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{dy} = \sqrt{2\pi}.$$

Towards this end, let $I = \int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{dy}$. Then

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{dx}\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{dy} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,\mathrm{dy}\,\mathrm{dx}\,. \end{aligned}$$

Perform a change of variable: $x = r\cos\theta$, $y = r\sin\theta$,

$$\begin{aligned} I^2 &= \int_0^{\infty}\int_0^{2\pi} e^{-r^2/2} r\,\mathrm{d}\theta\,\mathrm{dr} \\ &= 2\pi\int_0^{\infty} r e^{-r^2/2}\,\mathrm{dr} \\ &= 2\pi. \end{aligned}$$

Thus, $I = \sqrt{2\pi}$ and we have shown that $f(x)$ is indeed a probability density function.

A normal random variable is called a **standard normal** random variable when $\mu = 0$ and $\sigma = 1$ and is denoted by Z. That is $Z \sim N(0,1)$. Its probability density function is usually denoted by ϕ and its distribution function by Φ. That is,

$$\begin{aligned} \phi(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2}; \\ \Phi(x) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,\mathrm{dy}\,. \end{aligned}$$

An observation:
Let $Y \sim N(\mu,\sigma^2)$ and $Z \sim N(0,1)$, then

$$\begin{aligned} P(a < Y \le b) &= P\left(\frac{a-\mu}{\sigma} < Z \le \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

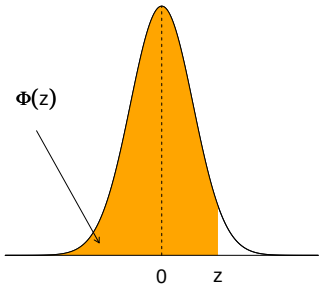
This is because

$$\begin{aligned} P(a < Y \leq b) &= \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, \mathrm{dy} \\ &= \frac{1}{\sqrt{2\pi}} \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} e^{-t^2/2} \, \mathrm{dt} \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right), \end{aligned}$$

where we let $t = (y-\mu)/\sigma$.

The values of $\Phi(x)$ for nonnegative x are given in the next page:

Distribution Function of the Normal Distribution



The function tabulated is $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}u^2}\,\mathrm{du}$.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999822	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

(Properties of the Standard Normal)

(i) $P(Z \geq 0) = P(Z \leq 0) = 0.5$.

(ii) $-Z \sim N(0,1)$.

(iii) $P(Z \leq x) = 1 - P(Z > x)$ for $-\infty < x < \infty$.

(iv) $P(Z \leq -x) = P(Z \geq x)$ for $-\infty < x < \infty$.

(v) If $Y \sim N(\mu, \sigma^2)$, then $X = \frac{Y-\mu}{\sigma} \sim N(0,1)$.

(vi) If $X \sim N(0,1)$, then $Y = aX + b \sim N(b, a^2)$ for $a, b \in \mathbb{R}$.

Example 5.15. When $X \sim N(65, 5^2)$, compute $P(47.5 < X \leq 80)$.

Solution:

$$\begin{aligned} P(47.5 < X \leq 80) &= P\left(\frac{47.5-65}{5} < Z \leq \frac{80-65}{5}\right) \\ &= P(-3.5 < Z \leq 3) \\ &= P(Z \leq 3) - P(Z \leq -3.5) \\ &= P(Z \leq 3) - P(Z \geq 3.5) \\ &= P(Z \leq 3) - (1 - P(Z < 3.5)) \\ &= 0.99865 - 1 + 0.999767 = 0.998417. \end{aligned}$$

(Important facts)

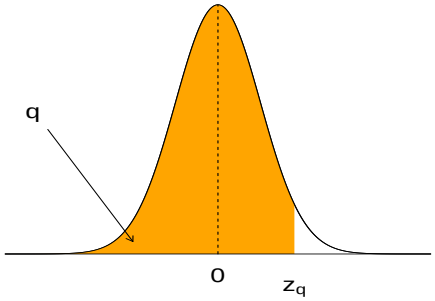
(a) If $Y \sim N(\mu, \sigma^2)$, then $E(Y) = \mu$ and $\text{var}(Y) = \sigma^2$.

(b) If $Z \sim N(0,1)$, then $E(Z) = 0$ and $\text{var}(Z) = 1$.

Definition 5.16. *The qth quantile of a random variable X is defined as a number z_q so that $P(X \leq z_q) = q$.*

It is sometimes of interest to look for the quantiles of the Normal distribution. The next table gives the quantiles of the Normal distribution.

QUANTILES OF THE NORMAL DISTRIBUTION



For a given q, this table gives z_q such that $\Phi(z_q) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{z_q} e^{-\frac{1}{2}u^2}\,\mathrm{du} = q$.

q	z_q	q	z_q	q	z_q	q	z_q	q	z_q	q	z_q	q	z_q
0.50	0.000	0.950	1.645	0.960	1.751	0.970	1.881	0.980	2.054	0.990	2.326	0.95	1.645
0.55	0.126	0.951	1.655	0.961	1.762	0.971	1.896	0.981	2.075	0.991	2.366	0.99	2.326
0.60	0.253	0.952	1.665	0.962	1.774	0.972	1.911	0.982	2.097	0.992	2.409	0.999	3.090
0.65	0.385	0.953	1.675	0.963	1.787	0.973	1.927	0.983	2.120	0.993	2.457	0.9999	3.719
0.70	0.524	0.954	1.685	0.964	1.799	0.974	1.943	0.984	2.144	0.994	2.512	0.99999	4.265
0.75	0.674	0.955	1.695	0.965	1.812	0.975	1.960	0.985	2.170	0.995	2.576	0.975	1.960
0.80	0.842	0.956	1.706	0.966	1.825	0.976	1.977	0.986	2.197	0.996	2.652	0.995	2.576
0.85	1.036	0.957	1.717	0.967	1.838	0.977	1.995	0.987	2.226	0.997	2.748	0.9995	3.291
0.90	1.282	0.958	1.728	0.968	1.852	0.978	2.014	0.988	2.257	0.998	2.878	0.99995	3.891
0.95	1.645	0.959	1.739	0.969	1.866	0.979	2.034	0.989	2.290	0.999	3.090	0.999995	4.417

Example 5.17. The width of a slot of a duralumin in forging is (in inches) normally distributed with $\mu = 0.9000$ and $\sigma = 0.0030$. The specification limits were given as 0.9000 ± 0.0050.

(a) What percentage of forgings will be defective?

(b) What is the maximum allowable value of σ that will permit no more than 1 in 100 defectives when the widths are normally distributed with $\mu = 0.9000$ and σ?

Solution:

(a) Let X be the width of our normally distributed slot. The probability that a forging is acceptable is given by

$$\begin{aligned} P(0.895 < X < 0.905) &= P\left(\frac{0.905-0.9}{0.003} < Z < \frac{0.895-0.9}{0.003}\right) \\ &= P(-1.67 < Z < 1.67) \\ &= 2\Phi(1.67) - 1 = 0.905. \end{aligned}$$

So that the probability that a forging is defective is $1 - 0.905 = 0.095$. Thus 9.5 percent of forgings are defective.

(b) We need to find the value of σ such that

$$P(0.895 < X < 0.905) = \frac{99}{100}.$$

Now

$$P(0.895 < X < 0.905) = 2P\left(Z < \frac{0.905-0.9}{\sigma}\right) - 1.$$

We thus have to solve for σ so that

$$2P\left(Z < \frac{0.005}{\sigma}\right) - 1 = 0.99.$$

or

$$P\left(Z < \frac{0.005}{\sigma}\right) = (1+0.99)/2 = 0.995.$$

The normal quantile table shows that $P(Z < 2.576) = 0.995$ so we can use $\frac{0.005}{\sigma} = 2.576$ which gives $\sigma = 0.0019$.

Example 5.18. An expert witness in a paternity suit testifies that the length (in days) of pregnancy is approximately normally distributed with parameters $\mu = 270$ and $\sigma = 10$. The defendant in the suit is able to prove that he was out of the country during a period that began 290 days before the birth of the child and ended 240 days before the birth. If the defendant was, in fact, the father of the child, what is the probability that the mother could have had a very long or a very short pregnancy indicated by the testimony?

Solution:
Let X denote the length of the pregnancy and assume that the defendant is the father. Then the probability of the birth could occur within the indicated duration is

$$\begin{aligned} & P(X > 290 \text{ or } X < 240) \\ = \; & P(X > 290) + P(X < 240) \\ = \; & P\left(Y > \frac{290-270}{10}\right) + P\left(Y < \frac{240-270}{10}\right) \\ = \; & 1 - \Phi(2) + \Phi(-3) \\ = \; & 1 - \Phi(2) + [1 - \Phi(3)] = 0.0241. \end{aligned}$$

Example 5.19 (The 3-σ Rule). Let $X \sim N(\mu, \sigma^2)$. Compute

$$P(|X - \mu| > 3\sigma).$$

Solution:

$$\begin{aligned} P(|X - \mu| > 3\sigma) &= P(|Z| > 3) \\ &= 2[1 - \Phi(3)] \\ &= 2(1 - 0.9987) \\ &= 0.0026 = 0.26\%. \end{aligned}$$

This says that for a normal distribution, nearly all (99.74%) of the values lie within 3 standard deviations of the mean.

5.5 Exponential Distribution

A random variable X is said to be **exponentially** distributed with parameter $\lambda > 0$ if its probability density function is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}.$$

The distribution function of X is given by

$$F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda x}, & x > 0 \end{cases}.$$

Recall the memoryless property of exponential distribution (Example 5.4, page 91):

$$P(X > s+t | X > s) = P(X > t), \quad \text{for } s,t > 0.$$

Mean and variance of $X \sim \text{Exp}(\lambda)$:

$$E(X) = \frac{1}{\lambda} \quad \text{var}(X) = \frac{1}{\lambda^2}.$$

Example 5.20. Suppose that the length of a phone call in minutes is an exponential random variable with parameter $\alpha = 1/10$. Someone arrives immediately ahead of you at a public phone booth, find the probability that you will have to wait

(i) more than 10 minutes;

(ii) between 10 to 20 minutes.

Solution:
Let X denote the duration of his call.

(i)
$$P(X > 10) = 1 - F_X(10) = e^{-10 \times 1/10} = e^{-1} = 0.368.$$

(ii)
$$P(10 < X < 20) = F(20) - F(10) = (1 - e^{-2}) - (1 - e^{-1}) = 0.233.$$

Example 5.21. Consider a post office that is staffed by two clerks. Suppose that when Mr. Smith enters the system, he discovers that Ms. Jones is being served by one of the clerks and Mr. Brown by the other. Suppose also that Mr. Smith is told that his service will begin as soon as either Ms. Jones or Mr. Brown leaves. If the amount of time that a clerk spends with a customer is exponentially distributed with parameter λ, what is the probability that, of the three customers, Mr. Smith is the last to leave the post office?

Solution:
The answer is obtained by reasoning as follows: Consider the time at

which Mr. Smith first finds a free clerk. At this point, either Ms. Jones or Mr. Brown would have just left, and the other one would still be in service. However, because the exponential is memoryless, it follows that the additional amount of time that this other person (either Ms. Jones or Mr. Brown) would still have to spend in the post office is exponentially distributed with parameter λ. That is, it is the same as if service for that person were just starting at this point. Hence, by symmetry, the probability that the remaining person finishes before Smith leaves must equal $\frac{1}{2}$.

Example 5.22. Jones figures that the total number of thousands of miles that an auto can be driven before it would need to be junked is an exponential random variable with parameter $1/20$. Smith has a used car that he claims has been driven only $10,000$ miles. If Jones purchases the car, what is the probability that she would get at least 20,000 additional miles out of it? Repeat under the assumption that the lifetime mileage of the car is not exponentially distributed but rather is (in thousands of miles) uniformly distributed over $(0,40)$.

Solution:
Let T be the lifetime mileage of the car in thousands of miles. Since the exponential random variable has no memory, the fact that the car has been driven $10,000$ miles makes no difference. The probability we are looking for is

$$P(T > 20) = e^{-\frac{1}{20}(20)} = e^{-1}.$$

If the lifetime distribution is not exponential but is uniform over $(0,40)$ then the desired probability is given by

$$P(T > 30|T > 10) = \frac{P(T > 30)}{P(T > 10)} = \frac{(1/4)}{(3/4)} = \frac{1}{3}.$$

5.6 Gamma Distribution

A random variable X is said to have a **gamma distribution** with parameters (α,λ), denoted by $X \sim \Gamma(\alpha,\lambda)$, if its probability density function is given by

$$f_X(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & x \geq 0 \\ 0, & x < 0 \end{cases},$$

where $\lambda > 0$, $\alpha > 0$ and $\Gamma(\alpha)$, called the **gamma function**, is defined by

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

Remark 26. (a) $\Gamma(1) = \int_0^\infty e^{-y}\,\mathrm{d}y = 1$.

(b) It can be shown, via integration by parts, that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$.

(c) For integral values of α, say, $\alpha = n$,

$$\begin{aligned}\Gamma(n) &= (n-1)\Gamma(n-1)\\ &= (n-1)(n-2)\Gamma(n-2)\\ &= \dots\\ &= (n-1)(n-2)\cdots 3\cdot 2\cdot \Gamma(1)\\ &= (n-1)!\end{aligned}$$

(d) Note that $\Gamma(1,\lambda) = \mathrm{Exp}(\lambda)$.

(e) $\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-y} y^{-\frac{1}{2}}\,\mathrm{d}y = \sqrt{\pi}$.

Proposition 5.23. *If events are occurring randomly and in accordance with the three axioms of page* **??**, *then it turns out that the amount of time one has to wait until a total of n events has occurred will be a gamma random variable with parameters* (n,λ).

Proof. Let T_n denote the time at which the nth event occurs, and $N(t)$ equal to the number of events in $[0,t]$. Note that $N(t) \sim \mathrm{Poisson}(\lambda t)$ according to the three axioms.

It is easy to see that

$$\{T_n \le t\} = \{N(t) \ge n\}.$$

Therefore

$$P(T_n \le t) = P(N(t) \ge n) = \sum_{j=n}^{\infty} P(N(t) = j) = \sum_{j=n}^{\infty} \frac{e^{-\lambda t}(\lambda t)^j}{j!}.$$

To get the density function of T_n, take derivatives on both sides with respect to t:

$$\begin{aligned}f(t) &= \sum_{j=n}^{\infty} \frac{e^{-\lambda t} j(\lambda t)^{j-1}\lambda}{j!} - \sum_{j=n}^{\infty} \frac{\lambda e^{-\lambda t}(\lambda t)^j}{j!}\\ &= \sum_{j=n}^{\infty} \frac{\lambda e^{-\lambda t}(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=n}^{\infty} \frac{\lambda e^{-\lambda t}(\lambda t)^j}{j!}\\ &= \sum_{j=n-1}^{\infty} \frac{\lambda e^{-\lambda t}(\lambda t)^j}{j!} - \sum_{j=n}^{\infty} \frac{\lambda e^{-\lambda t}(\lambda t)^j}{j!}\\ &= \frac{\lambda e^{-\lambda t}(\lambda t)^{n-1}}{(n-1)!}\end{aligned}$$

Hence, T_n has the gamma distribution with parameters (n, λ). □

5.7 Beta Distribution

A random variable X is said to have a **beta distribution** with parameters (a, b), denoted by $X \sim \text{Beta}(a, b)$, if its density is given by

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,\text{dx}$$

is known as the **beta function**.

It can be shown that

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Using this, it is easy to show that if X is $\text{Beta}(a, b)$, then

$$E[X] = \frac{a}{a+b} \quad \text{and} \quad \text{var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

5.8 Cauchy Distribution

A random variable X is said to have a **Cauchy distribution** with parameter θ, $-\infty < \theta < \infty$, denoted by $X \sim \text{Cauchy}(\theta)$, if its density is given by

$$f(x) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}, \quad -\infty < x < \infty.$$

5.9 Approximations of Binomial Random Variables

Let $X \sim \text{Bin}(n, p)$. Here n is assumed (which is natural) to be large (say, ≥ 30).

There are two commonly used approximations of the binomial distributions:

(a) Normal approximation; and

(b) Poisson approximation.

Normal approximation of binomial random variable

This method can be viewed as a precursor of a very important theorem in probability: Central Limit Theorem.

Theorem 5.24 (De Moivre-Laplace Limit Theorem)**.** *Suppose that* $X \sim \text{Bin}(n,p)$. *Then for any* $a < b$,

$$P\left(a < \frac{X-np}{\sqrt{npq}} \leq b\right) \to \Phi(b) - \Phi(a)$$

as $n \to \infty$, *where* $q = 1-p$.

That is,

$$\text{Bin}(n,p) \approx N(np, npq).$$

Equivalently,

$$\frac{X-np}{\sqrt{npq}} \approx Z$$

where $Z \sim N(0,1)$.

Remark 27. The normal approximation will be generally quite good for values of n satisfying $np(1-p) \geq 10$.

Approximation is further improved if we incorporate

(Continuity-correction) If $X \sim \text{Bin}(n,p)$, then

$$P(X = k) = P\left(k - \frac{1}{2} < X < k + \frac{1}{2}\right)$$
$$P(X \geq k) = P\left(X \geq k - \frac{1}{2}\right)$$
$$P(X \leq k) = P\left(X \leq k + \frac{1}{2}\right)$$

Example 5.25. Let X be the number of times that a fair coin, flipped 40 times, lands heads. Find the probability that $X = 20$. Use the normal approximation to compute this probability as well and compare the two answers.

Solution:
Exact answer:

$$P(X = 20) = \binom{40}{20}\left(\frac{1}{2}\right)^{40} = 0.1254.$$

Approximate answer:

$$\begin{aligned} P(X=20) &= P(19.5 \le X \le 20.5) \\ &= P\left(\frac{19.5-20}{\sqrt{10}} \le \frac{X-20}{\sqrt{10}} \le \frac{20.5-20}{\sqrt{10}}\right) \\ &\approx P(-0.16 \le Z \le 0.16) \\ &= 0.1272. \end{aligned}$$

Example 5.26. Let $X \sim \text{Bin}(1000000, 0.01)$. We are interested in

$$P(8000 \le X \le 120000).$$

This is a non-trivial computation. A normal approximation will be helpful in this case.

Example 5.27. The ideal size of a first-year class at a particular college is 150 students. The college, knowing from the past experience that on the average only 30% of those accepted for admission to this class will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 students attend this class.
For the same question above, if this college desires that the probability of more than 150 students will attend this college should be at most 0.01. What is the largest number of students should this college admit?

Solution:
Let X denote the number of students that attend, then

$$X \sim \text{Bin}(450, 0.3).$$

Applying normal approximation with continuity correction:
First

$$E(X) = 450 \times 0.3 = 135; \quad \text{SD}(X) = \sqrt{450(0.3)(0.7)} = 9.721,$$

then

$$\begin{aligned} P(X > 150) &= P(X \ge 150.5) \\ &\approx P\left(Z \ge \frac{150.5-135}{9.721}\right) \\ &= P(Z \ge 1.59) \\ &= 0.0559 = 5.59\%. \end{aligned}$$

Let n be the number of students to be admitted. then

$$Y \sim \text{Bin}(n, 0.3).$$

Applying normal approximation with continuity correction:
First

$$E(X) = 3n/10 \quad \text{and} \quad \text{SD}(X) = \sqrt{21n}/10,$$

then

$$\begin{aligned} P(Y > 150) &= P(Y \geq 150.5) \\ &\approx P\left(Z \geq \frac{150.5 - 3n/10}{\sqrt{21n}/10}\right) \\ &\leq 0.01. \end{aligned}$$

The normal quantile table gives

$$P(Z < 2.326) = 0.99 \Longleftrightarrow P(Z > 2.326) = 0.01.$$

Accordingly, we can pick

$$\frac{1505 - 3n}{\sqrt{21n}} \geq 2.326$$

and this works out to

$$n \leq 428.$$

Example 5.28 (Continuity Correction).
An example on how to apply Continuity Correction:
Let $X \sim \text{Bin}(60, 0.3)$ and $Z \sim N(0,1)$. Then

(i) $P(12 \leq X \leq 26) \approx P\left(\frac{11.5-18}{\sqrt{12.6}} \leq Z \leq \frac{26.5-18}{\sqrt{12.6}}\right)$.

(ii) $P(12 < X \leq 26) \approx P\left(\frac{12.5-18}{\sqrt{12.6}} \leq Z \leq \frac{26.5-18}{\sqrt{12.6}}\right)$.

(iii) $P(12 \leq X < 26) \approx P\left(\frac{11.5-18}{\sqrt{12.6}} \leq Z \leq \frac{25.5-18}{\sqrt{12.6}}\right)$.

(iv) $P(12 < X < 26) \approx P\left(\frac{12.5-18}{\sqrt{12.6}} \leq Z \leq \frac{25.5-18}{\sqrt{12.6}}\right)$.

Poisson approximation of binomial random variable

The Poisson distribution is used as an approximation to the binomial distribution when the parameters n and p are large and small, respectively and that np is moderate.

As a working rule, use the Poisson approximation if $p < 0.1$ and put $\lambda = np$.

Example 5.29. Let $X \sim \text{Bin}(400, 0.01)$, then $X \approx Z$ where $Z \sim \text{Poisson}(\lambda)$ and $\lambda = np = 4$. Therefore,

$$0.890375 = P(X \leq 6) \approx P(Z \leq 6) = 0.889326.$$

Remark 28. If $p > 0.9$, put $\lambda = n(1-p)$ and work in terms of "failure".

5.10 Distribution of a Function of a Random Variable

Example 5.30. Let $X \sim N(0,1)$. What are the distribution function and probability density function of Y, where $Y = X^2$?

Solution:
First note that Y takes nonnegative values. Therefore for $y > 0$, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-x^2/2}\,\text{dx}\,. \end{aligned}$$

Hence

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

This gives

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-x^2/2}\,\text{dx}, & y > 0 \end{cases}$$

and

$$f_Y(y) = \begin{cases} 0, & y \leq 0 \\ \frac{y^{-1/2} e^{-y/2}}{\sqrt{2\pi}}, & y > 0 \end{cases}.$$

Y is known in the literature as a **chi-square (χ^2) random variable** of degree 1, denoted by χ_1^2. Note that χ_1^2 is $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$.

Example 5.31. Let $X \sim N(0,1)$. Define $Y = e^X$, commonly known as the **lognormal random variable**. Find the probability density function f_Y .

Solution:
First note that Y takes nonnegative values. Therefore for $y > 0$, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(e^X \leq y) \\ &= P(X \leq \ln y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\ln y} e^{-x^2/2}\,\text{dx}\,. \end{aligned}$$

Hence

$$f_Y(y) = F_Y'(y) = \frac{1}{\sqrt{2\pi}}\frac{1}{y}e^{-(\ln y)^2/2}.$$

This gives

$$f_Y(y) = \begin{cases} 0, & y \le 0 \\ \frac{1}{y\sqrt{2\pi}}e^{-(\ln y)^2/2}, & y > 0 \end{cases}.$$

Example 5.32. Let X be a continuous random variable with probability density function f_X. And let $Y = X^n$ where n is odd. Find the probability density function f_Y.

Solution:
Let $y \in \mathbb{R}$, then

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(X^n \le y) \\ &= P(X \le y^{1/n}) = F_X(y^{1/n}). \end{aligned}$$

Hence

$$f_Y(y) = F_Y'(y) = \frac{1}{n}y^{1/n-1}f_X(y^{1/n}).$$

The method employed in the previous examples can be used to prove

Theorem 5.33. *Let X be a continuous random variable having probability density function f_X. Suppose that $g(x)$ is a strictly monotonic (increasing or decreasing), differentiable (and thus continuous) function of x. Then the random variable Y defined by $Y = g(X)$ has a probability density function given by*

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|, & \text{if } y = g(x) \text{ for some } x \\ 0, & \text{if } y \ne g(x) \text{ for all } x \end{cases}$$

where $g^{-1}(y)$ is defined to be equal that value of x such that $g(x) = y$.

Proof. We shall assume that $g(x)$ is an increasing function. Suppose $y = g(x)$ for some x. Then, with $Y = g(X)$,

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

Differentiation gives

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y),$$

which agrees with the form given in the theorem, since $g^{-1}(y)$ is nondecreasing, so its derivative is non-negative.
When $y \ne g(x)$ for any x, $F_Y(y)$ is either 0 or 1. In either case $f_Y(y) = 0$. □

Example 5.34. Let Y be a continuous random variable with distribution function F. We assume further that F is a strictly increasing function. Define the random variable X by $X = F(Y)$.

(i) What are the possible values of X?

(ii) Find the probability density function of X. Can you identify X?

Solution:

(i) X takes values in $[0,1]$.

(ii) For $x \in (0,1)$,

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(F(Y) \leq x) \\ &= P(Y \leq F^{-1}(x)) \\ &= F(F^{-1}(x)) \\ &= x. \end{aligned}$$

$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases},$$

that is, $X \sim \text{uniform}(0,1)$.

Remark 29. There are two things worth taking note of in this example.

(a) If X is a continuous random variable with distribution function F, then $F(X)$ is the uniform distribution $U(0,1)$.

(b) If U is the uniform distribution $U(0,1)$, then $F^{-1}(U)$ will have the distribution function $F(x)$. This result, known as the *inverse transformation method*, is often made used of to generate continuous random variables having distribution function F in computer packages.

Example 5.35 (Generating an Exponential Random Variable). Let

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

be the distribution function of an exponential random variable with parameter λ. Then $F^{-1}(u)$ is that value x such that

$$u = F(x) = 1 - e^{-\lambda x}$$

or, equivalently,

$$x = -\frac{1}{\lambda}\log(1-u)$$

So we can generate an exponential random variable X with parameter λ by generating a uniform $(0,1)$ random variable U and setting

$$X = -\frac{1}{\lambda}\log(1-U).$$

Because $1-U$ is also a uniform $(0,1)$ random variable, it follows that $-\frac{1}{\lambda}\log(1-U)$ and $-\frac{1}{\lambda}\log(U)$ have the same distribution, thus showing that

$$X = -\frac{1}{\lambda}\log(U)$$

is also exponential with parameter λ.

Chapter 6

Jointly Distributed Random Variables

6.1 Joint Distribution Functions

Very often we are interested in two or more random variables at the same time. For example,

(1) In the population of HKUST students (our sample space), we are interested in the following characteristics of a student, his/her age (A), gender (G), major (M), and year (Y) of studies in HKUST.

Here, for each student, (A, G, M, Y) denotes a student's age, ..., etc.

(2) For any particular day, we are interested in the number of vehicle accidents, how many deaths in these accidents, and how many major injuries on the road.

Here we can use (A, D, I) to denote the 3 quantities we are interested.

Definition 6.1. *For any two random variables X and Y defined on the same sample space, we define the* ***joint distribution function of X and Y*** *(we abbreviate it to joint d.f. and denote it by $F_{X,Y}$) by*

$$F_{X,Y}(x,y) = P(X \le x, Y \le y) \quad \text{for } x, y \in \mathbb{R}.$$

Remark 30. Notation

$$\begin{aligned} \{X \le x; Y \le y\} &= \{s \in S : X(s) \le x \text{ and } Y(s) \le y\} \\ &= \{s \in S : X(s) \le x\} \cap \{s \in S : Y(s) \le y\} \\ &= \{X \le x\} \cap \{Y \le y\}. \end{aligned}$$

The distribution function of X can be obtained from the joint density function of X and Y in the following way:

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y).$$

We call F_X the **marginal distribution function of X.**
Similarly,

$$F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y)$$

and F_Y is called the **marginal distribution function of Y.**

Some useful calculations

These formulas are not only useful in some calculations, but the derivations are just as important. Let $a, b, a_1 < a_2, b_1 < b_2$ be real numbers, then

$$P(X > a, Y > b) = 1 - F_X(a) - F_Y(b) + F_{X,Y}(a,b), \tag{6.1}$$

$$\begin{aligned} P(a_1 < X \leq a_2,\ b_1 < Y \leq b_2) = F_{X,Y}(a_2,b_2) - F_{X,Y}(a_1,b_2) \\ + F_{X,Y}(a_1,b_1) - F_{X,Y}(a_2,b_1). \end{aligned} \tag{6.2}$$

Proof.
Equation (6.1) – Write $A = \{X \leq a\}$ and $B = \{Y \leq b\}$. Then

$$\{X > a, Y > b\} = A^c B^c = (A \cup B)^c.$$

Therefore,

$$\begin{aligned} P(X > a,\ Y > b) &= P((A \cup B)^c) \\ &= 1 - P(A \cup B) \\ &= 1 - P(A) - P(B) + P(AB) \\ &= 1 - F_X(a) - F_Y(b) + F_{X,Y}(a,b). \end{aligned}$$

Equation (6.2) – Left as an exercise. □

6.1.1 Jointly Discrete Random Variables

In the case when both X and Y are discrete random variables, we define the **joint probability mass function of X and Y** (abbreviated to joint p.m.f. and denoted by $p_{X,Y}$) as :

$$p_{X,Y}(x,y) = P(X = x, Y = y).$$

As in the distribution functions situation, we can recover the probability mass function of X and Y in the following manner:

$$p_X(x) = P(X = x) = \sum_{y \in \mathbb{R}} p_{X,Y}(x,y),$$
$$p_Y(y) = P(Y = y) = \sum_{x \in \mathbb{R}} p_{X,Y}(x,y).$$

We call p_X the **marginal probability mass function of** X and p_Y the **marginal probability mass function of** Y.

Example 6.2. Suppose that 3 balls are randomly selected from an urn containing 3 red, 4 white and 5 blue balls. If we let R and W denote the number of red and white balls chosen, then the joint probability mass function of R and W is

w / r	0	1	2	3	$P(R = r)$
0	10/220	40/220	30/220	4/220	84/220
1	30/220	60/220	18/220	0	108/220
2	15/220	12/220	0	0	27/220
3	1/220	0	0	0	1/220
$P(W = w)$	56/220	112/220	48/220	4/220	

We illustrate how some of the entries are computed:

(i) $p(0,0) = \binom{5}{3} / \binom{12}{3} = \frac{10}{220}$.

(ii) $p(2,2) = 0$ as the event of getting 2 red balls and 2 white balls are impossible.

Some useful formulas

(i)
$$P(a_1 < X \leq a_2,\ b_1 < Y \leq b_2) = \sum_{a_1 < x \leq a_2} \sum_{b_1 < y \leq b_2} p_{X,Y}(x,y),$$

(ii)
$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = \sum_{x \leq a} \sum_{y \leq b} p_{X,Y}(x,y),$$

(iii)
$$P(X > a, Y > b) = \sum_{x > a} \sum_{y > b} p_{X,Y}(x,y).$$

Example 6.3. Suppose that 15% of the families in a certain community have no children 20% have 1, 35% have 2, and 30% have 3. We suppose further that, in each family, each child is equally likely (independently) to be a boy or a girl. If a family is chosen at random from this community, then B, the number of boys; and G, the number of girls, in this family will have the joint probability mass function as shown in the following table.

Table for $P(B = i, G = j)$.

j / i	0	1	2	3	$P(B = i)$
0	0.1500	0.1000	0.0875	0.0375	0.3750
1	0.1000	0.1750	0.1125	0	0.3875
2	0.0875	0.1125	0	0	0.2000
3	0.0375	0	0	0	0.0375
$P(G = j)$	0.3750	0.3875	0.2000	0.0375	

Some of these probabilities are calculated as follows:

$$P(B = 0, G = 0) = P(\text{no children}) = 0.15.$$

$$\begin{aligned} P(B = 0, G = 1) &= P(1 \text{ girl and total of children is } 1) \\ &= P(1 \text{ child})P(1 \text{ girl}|1 \text{ child}) \\ &= 0.20 \times \frac{1}{2} = 0.10. \end{aligned}$$

$$\begin{aligned} P(B = 0, G = 2) &= P(2 \text{ girls and total of children is } 2) \\ &= P(2 \text{ children})P(2 \text{ girls}|2 \text{ children}) \\ &= 0.35 \times \frac{1}{2^2} = 0.0875. \end{aligned}$$

$$\begin{aligned} P(B = 1, G = 2) &= P(2 \text{ girls and total of children is } 3) \\ &= P(3 \text{ children})P(2 \text{ girls}|3 \text{ children}) \\ &= 0.30 \times \binom{3}{2} \times \frac{1}{2^3} = 0.1125. \end{aligned}$$

6.1.2 Jointly Continuous Random Variables

We say that X and Y are **jointly continuous random variables** if there exists a function (which is denoted by $f_{X,Y}$, called the **joint probability**

density function of X and Y) if for every set $C \subset \mathbb{R}^2$, we have

$$P((X,Y) \in C) = \int\int_{(x,y)\in C} f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy}\,.$$

Some useful formulas

(i) Let $A, B \subset \mathbb{R}$, take $C = A \times B$ above

$$P(X \in A, Y \in B) = \int_A \int_B f_{X,Y}(x,y)\,\mathrm{dy}\,\mathrm{dx}\,.$$

(ii) In particular, Let $a_1, a_2, b_1, b_2 \in \mathbb{R}$ where $a_1 < a_2$ and $b_1 < b_2$, we have

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)\,\mathrm{dy}\,\mathrm{dx}\,.$$

(iii) Let $a, b \in \mathbb{R}$, we have

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x,y)\,\mathrm{dy}\,\mathrm{dx}\,.$$

As a result of this,

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

Definition 6.4. *The marginal probability density function of X is given by*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{dy}\,.$$

Similarly, the marginal probability density function of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{dx}\,.$$

Example 6.5. The joint probability density function of X and Y is given by

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x}e^{-2y}, & 0 < x < \infty, 0 < y < \infty \\ 0, & \text{otherwise} \end{cases}.$$

Compute

(i) $P(X > 1, Y < 1)$,

(ii) $P(X < Y)$,

(iii) $P(X \leq x)$,

(iv) the marginal probability density function of X,

(v) the marginal distribution function of Y.

Solution:

(i)

$$\begin{aligned} P(X>1, Y<1) &= \int_0^1 \int_1^\infty f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy} \\ &= \int_0^1 \int_1^\infty 2e^{-x}e^{-2y}\,\mathrm{dx}\,\mathrm{dy} \\ &= e^{-1}(1-e^{-2}). \end{aligned}$$

(ii)

$$\begin{aligned} P(X<Y) &= \int\int_{x<y} f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy} \\ &= \int_0^\infty \int_0^y 2e^{-x}e^{-2y}\,\mathrm{dx}\,\mathrm{dy} \\ &= \int_0^\infty [2e^{-2y}-2e^{-3y}]\,\mathrm{dy} \\ &= 1/3. \end{aligned}$$

(iv) We will work on (iv) first.
For $x \le 0$, $f_X(x) = 0$. For $x > 0$,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^\infty f_{X,Y}(x,y)\,\mathrm{dy} \\ &= \int_0^\infty 2e^{-x}e^{-2y}\,\mathrm{dy} \\ &= e^{-x}, \end{aligned}$$

Hence

$$f_X(x) = \begin{cases} e^{-x}, & x>0 \\ 0, & x \le 0 \end{cases}.$$

(iii) Marginal distribution function of X is: For $x > 0$,

$$F_X(x) = \int_{-\infty}^x f_X(t)\,\mathrm{dt} = \int_0^x e^{-t}\,\mathrm{dt} = 1-e^{-x}.$$

And for $x \le 0$, $F_X(x) = 0$. Hence, marginal distribution function of X is

$$F_X(x) = \begin{cases} 1-e^{-x}, & x>0 \\ 0, & x \le 0 \end{cases}.$$

(v) Marginal distribution function of Y.
For $y \leq 0$, $F_Y(y) = 0$. For $y > 0$,

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{y} f_Y(t)\,\mathrm{dt} \\ &= \int_0^y \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,\mathrm{dx}\,\mathrm{dt} \\ &= \int_0^y \int_0^{\infty} 2e^{-x}e^{-2t}\,\mathrm{dx}\,\mathrm{dt} \\ &= 1 - e^{-2y}. \end{aligned}$$

Hence, marginal distribution function of Y is

$$F_Y(y) = \begin{cases} 1 - e^{-2y}, & y \geq 0 \\ 0, & y < 0 \end{cases}.$$

Example 6.6. The joint density of X and Y is given by

$$f(x,y) = \begin{cases} e^{-(x+y)}, & 0 < x < \infty, 0 < y < \infty \\ 0, & \text{otherwise} \end{cases}.$$

Find the density function of the random variable X/Y.

Solution:
We start by computing the distribution function of X/Y. For $a > 0$,

$$\begin{aligned} F_{X/Y}(a) &= P\left(\frac{X}{Y} \leq a\right) \\ &= \iint_{x/y \leq a} e^{-(x+y)}\,\mathrm{dx}\,\mathrm{dy} \\ &= \int_0^{\infty} \int_0^{ay} e^{-(x+y)}\,\mathrm{dx}\,\mathrm{dy} \\ &= \int_0^{\infty} (1 - e^{-ay})e^{-y}\,\mathrm{dy} \\ &= \left[-e^{-y} + \frac{e^{-(a+1)y}}{a+1}\right]_0^{\infty} \\ &= 1 - \frac{1}{a+1}. \end{aligned}$$

Differentiation yields that the density function of X/Y is given by

$$f_{X/Y}(a) = 1/(a+1)^2, \quad 0 < a < \infty.$$

6.2 Independent Random Variables

Two random variables X and Y are said to be **independent** if

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \quad \text{for any } A, B \subset \mathbb{R}.$$

Random variables that are **not** independent are said to be **dependent.**

Proposition 6.7 (For jointly discrete random variables)**.** *The following three statements are equivalent:*

(i) Random variables X and Y are independent.

(ii) For all $x, y \in \mathbb{R}$, we have

$$p_{X,Y}(x,y) = p_X(x)p_Y(y).$$

(iii) For all $x, y \in \mathbb{R}$, we have

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

Proposition 6.8 (For jointly continuous random variables)**.** *The following three statements are equivalent:*

(i) Random variables X and Y are independent.

(ii) For all $x, y \in \mathbb{R}$, we have

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

(iii) For all $x, y \in \mathbb{R}$, we have

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

Example 6.9. Refer to the Example 6.2. Note that

$$P(R = 2, W = 2) = 0 \neq P(R = 2)P(W = 2).$$

Hence, R and W are dependent.
Any obvious reason (I mean without any calculation) why they are dependent?

Example 6.10. Refer to the Example 6.5. One can verify that

$$f_X(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \quad ; \quad f_Y(y) = \begin{cases} 2e^{-2y}, & y > 0 \\ 0, & \text{otherwise} \end{cases} .$$

Then, we can check that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for every $x, y \in \mathbb{R}$. Hence, X and Y are independent.

Example 6.11. Suppose that $n+m$ independent trials, having a common success probability p, are performed. If X is the number of successes in the first n trials, and Y is the number of successes in the final m trials, then X and Y are independent, since knowing the number of successes in the first n trials does not affect the distribution of the number of successes in the final m trials (by the assumption of independent trials). In fact, for integral x and y,

$$\begin{aligned} P(X = x, Y = y) &= \binom{n}{x} p^x (1-p)^{n-x} \binom{m}{y} p^y (1-p)^{m-y} \\ &= P(X = x)P(Y = y), \quad 0 \le x \le n, 0 \le y \le m. \end{aligned}$$

On the other hand, X and Z will be dependent, where $Z = X + Y$ is the total number of successes in the $n+m$ trials. To see this, consider the fact that

$$P(Z = z) = \binom{m+n}{z} p^z (1-p)^{m+n-z}, \quad 0 \le z \le m+n.$$

However,

$$\begin{aligned} P(X = x, Z = z) &= P(X = x, X + Y = z) \\ &= P(X = x, Y = z - x) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \binom{m}{z-x} p^{z-x} (1-p)^{m-z+x}, \end{aligned}$$

for $0 \le x \le n, x \le z \le m + x$.
So we see that

$$P(X = x, Z = z) \neq P(X = x)P(Z = z),$$

indicating that X and Z are not independent.

Example 6.12. A man and a woman decide to meet at a certain location. If each person independently arrives at a time uniformly distributed between 12 noon and 1 pm, find the probability that the first to arrive has to wait longer than 10 minutes.

Solution:
If we let X and Y denote, respectively, the time past 12 that the man and the woman arrive, then X and Y are independent random variables, each of which is uniformly distributed over $(0,60)$. The desired probability, $P(X+10<Y)+P(Y+10<X)$, which by symmetry equals $2P(X+10<Y)$, is obtained as follows:

$$\begin{aligned}
2P(X+10<Y) &= 2\iint_{x+10<y} f(x,y)\,\mathrm{dx}\,\mathrm{dy} \\
&= 2\iint_{x+10<y} f_X(x)f_Y(y)\,\mathrm{dx}\,\mathrm{dy} \\
&= 2\int_{10}^{60}\int_{0}^{y-10}\left(\frac{1}{60}\right)^2 \mathrm{dx}\,\mathrm{dy} \\
&= \frac{2}{60^2}\int_{10}^{60}(y-10)\,\mathrm{dy} \\
&= \frac{25}{36}.
\end{aligned}$$

Our next example presents the oldest problem dealing with geometrical probabilities. It was first considered and solved by Buffon, a French naturalist of the eighteenth century, and is usually referred to as Buffon's needle problem.

Example 6.13 (Buffon's needle problem). A table is ruled with equidistant parallel lines a distance D apart. A needle of length L, where $L \leq D$, is randomly thrown on the table. What is the probability that the needle will intersect one of the lines (the other possibility being that the needle will be completely contained in the strip between two lines)?

Solution:
Let us determine the position of the needle by specifying the distance X from the middle point of the needle to the nearest parallel line, and the angle θ between the needle and the projected line of length X. The needle will intersect a line if the hypotenuse of the right triangle is less than $L/2$, that is, if

$$\frac{X}{\cos\theta} < \frac{L}{2} \text{ or } X < \frac{L}{2}\cos\theta.$$

As X varies between 0 and $D/2$ and θ between 0 and $\pi/2$, it is reasonable to assume that they are independent, uniformly distributed random variables over these respective ranges. Hence

$$
\begin{aligned}
P\left(X < \frac{L}{2}\cos\theta\right) &= \iint_{x<L/2\cos\theta} f_X(x) f_\theta(\theta)\,\mathrm{dx}\,\mathrm{d}\theta \\
&= \frac{4}{\pi D}\int_0^{\pi/2}\int_0^{L/2\cos\theta} \mathrm{dx}\,\mathrm{d}\theta \\
&= \frac{4}{\pi D}\int_0^{\pi/2}\frac{L}{2}\cos\theta\,\mathrm{d}\theta \\
&= \frac{2L}{\pi D}.
\end{aligned}
$$

Proposition 6.14. *Random variables X and Y are independent if and only if there exist functions $g, h : \mathbb{R} \longrightarrow \mathbb{R}$ such that for all $x, y \in \mathbb{R}$, we have*

$$f_{X,Y}(x,y) = g(x)h(y).$$

Proof. Let us give the proof in the continuous case. First note that independence implies that the joint density is the product of the marginal densities of X and Y, so the preceding factorization will hold when the random variables are independent. Now, suppose that

$$f_{X,Y}(x,y) = h(x)g(y).$$

Then

$$
\begin{aligned}
1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy} \\
&= \int_{-\infty}^{\infty} h(x)\,\mathrm{dx}\int_{-\infty}^{\infty} g(y)\,\mathrm{dy} \\
&= C_1 C_2.
\end{aligned}
$$

where $C_1 = \int_{-\infty}^{\infty} h(x)\,\mathrm{dx}$ and $C_2 = \int_{-\infty}^{\infty} g(y)\,\mathrm{dy}$. Also,

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{dy} = C_2 h(x), \\
f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{dx} = C_1 g(y).
\end{aligned}
$$

Since $C_1 C_2 = 1$, we thus see that

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

and the proof is complete. □

Example 6.15. In a previous example, we have

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x}e^{-2y}, & 0 < x < \infty, 0 < y < \infty \\ 0, & \text{otherwise} \end{cases}.$$

We can take

$$g(x) = \begin{cases} 4e^{-x}, & 0 < x < \infty \\ 0, & \text{elsewhere} \end{cases}$$

and

$$h(y) = \begin{cases} 1/2e^{-2y}, & 0 < y < \infty \\ 0, & \text{elsewhere} \end{cases}.$$

Then it is easily verified that

$$f_{X,Y}(x,y) = g(x)h(y) \quad \text{for all } x, y \in \mathbb{R}.$$

Therefore, X and Y are independent.

Example 6.16. If the joint density function of X and Y is

$$f(x,y) = 24xy, \quad 0 < x < 1, 0 < y < 1, 0 < x + y < 1$$

and is equal to 0 otherwise, are the random variables independent?

Solution:
The region in which the joint density is nonzero cannot be expressed in the form $x \in A, y \in .B$. This means that the joint density does not factor and so the random variables are not independent.
More formally, we can define

$$I(x,y) = \begin{cases} 1, & \text{if } 0 < x < 1, 0 < y < 1, 0 < x + y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Then $f(x,y)$ can be written as

$$f(x,y) = 24xyI(x,y),$$

which clearly does not factor into a part depending only on x and another depending only on y.

Example 6.17. Suppose there are N traffic accidents in a day, where N is assumed to Poisson distributed with parameter λ. Each traffic accident is classified as major and minor. Suppose that given a traffic accident, it is a major accident with probability $p \in (0,1)$. Let X and Y denote the numbers of major and minor accidents respectively.

(i) Find the joint probability mass function of X and Y.

(ii) Are X and Y independent?

(iii) Can you identify the distributions of X and Y?

Solution:

(i) For $0 \leq i, j < \infty$,

$$\begin{aligned} P(X=i,Y=j) &= P(X=i,N=i+j) \\ &= P(N=i+j)P(X=i|N=i+j) \\ &= \frac{e^{-\lambda}\lambda^{i+j}}{(i+j)!}\binom{i+j}{i}p^i q^j \\ &= \frac{e^{-\lambda}\lambda^{i+j}p^i q^j}{i!j!} \\ &= \frac{e^{-\lambda}(\lambda p)^i}{i!}\frac{(\lambda q)^j}{j!} \\ &= \frac{e^{-\lambda p}(\lambda p)^i}{i!}\frac{e^{-\lambda q}(\lambda q)^j}{j!}. \end{aligned}$$

(ii) Define

$$g(x) = \begin{cases} \frac{e^{-\lambda p}(\lambda p)^x}{x!}, & \text{for } x \text{ integer} \geq 0 \\ 0, & \text{otherwise} \end{cases};$$

and

$$h(y) = \begin{cases} \frac{e^{-\lambda q}(\lambda q)^y}{y!}, & \text{for } y \text{ integer} \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

Hence, $p_{X,Y}(x,y) = g(x)h(y)$ for all $x, y \in \mathbb{R}$. And so X and Y are independent.

(iii) From the previous part, we see that

$$X \sim \text{Poisson}(\lambda p) \text{ and } Y \sim \text{Poisson}(\lambda q).$$

Remark 31. In a lot of applications, we either know or assume that X and Y are independent, then the joint probability density function (or probability mass function) of X and Y can be readily obtained by multiplying the probability density functions (or probability mass functions of X and Y).

Example 6.18. Suppose that X and Y are independent standard normal distribution. Find the joint probability density function of X and Y.

Solution:
Recall that X and Y have the common probability density function given by

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

Therefore the joint probability density function of X, Y is

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x) f_Y(y) \quad \text{(by independence)} \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \\ &= \frac{1}{2\pi} e^{-[x^2+y^2]/2}, \quad \text{for } -\infty < x < \infty, -\infty < y < \infty. \end{aligned}$$

Example 6.19. Suppose that X and Y are both discrete and independent. Then

$$p_{X,Y}(x,y) = p_X(x) p_Y(y).$$

That is, the entries of the table can be obtained from the margins.

x / y	x_1	x_2	$\cdots$	x_n	$p_Y(y)$
y_1					$p_Y(y_1)$
y_2		*			$p_Y(y_2)$
$\vdots$					$\vdots$
y_m					$p_Y(y_m)$
$p_X(x)$	$p_X(x_1)$	$p_X(x_2)$	$\cdots$	$p_X(x_n)$	

Example 6.20. Let X, Y, Z be independent and uniformly distributed over $(0,1)$. Compute $P(X \geq YZ)$.

Solution:
Since

$$f_{X,Y,Z}(x,y,z) = f_X(x) f_Y(y) f_Z(z) = 1, \quad 0 \leq x, y, z \leq 1,$$

we have

$$
\begin{aligned}
P(X \geq YZ) &= \iiint_{x \geq yz} f_{X,Y,Z}(x,y,z)\,\mathrm{dx}\,\mathrm{dy}\,\mathrm{dz} \\
&= \int_0^1 \int_0^1 \int_{yz}^1 \mathrm{dx}\,\mathrm{dy}\,\mathrm{dz} \\
&= \int_0^1 \int_0^1 (1-yz)\,\mathrm{dy}\,\mathrm{dz} \\
&= \int_0^1 (1-z/2)\,\mathrm{dz} \\
&= \frac{3}{4}.
\end{aligned}
$$

Remark 32 (Independence is a symmetric relation). To say that X is independent of Y is equivalent to saying that Y is independent of X, or just that X and Y are independent.

As a result, in considering whether X is independent of Y in situations where it is not at all intuitive that knowing the value of Y will not change the probabilities concerning X, it can be beneficial to interchange the roles of X and Y and ask instead whether Y is independent of X. The next example illustrates this point.

Example 6.21. If the initial throw of the dice in the game of craps results in the sum of the dice equaling 4, then the player will continue to throw the dice until the sum is either 4 or 7. If this sum is 4, then the player wins, and if it is 7, then the player loses. Let N denote the number of throws needed until either 4 or 7 appears, and let X denote the value (either 4 or 7) of the final throw. Is N independent of X?

Solution:
The answer to this question is not intuitively obvious. However, suppose that we tum it around and ask whether X is independent of N. That is, does knowing how many throws it takes to obtain a sum of either 4 or 7 affect the probability that that sum is equal to 4?

Clearly not, since the fact that none of the first $n-1$ throws were either 4 or 7 does not change the probabilities for the nth throw. Thus, we can conclude that X is independent of N, or equivalently, that N is independent of X.

6.3 Sums of Independent Random Variables

Very often we are interested in the sum of independent random variables. For example,

(1) Two dice are rolled, we are interested (as in many games) in the sum of the two numbers.

(2) In a data set, each datum collected is rounded off to the nearest integer. Let X_i denote the error in rounding the ith datum. Suppose we want to compute the total of this data set.

One quantity that we are interested is: sum of the errors due to rounding off.

6.3.1 X and Y are continuous and independent

Under the assumption of independence of X and Y, we have

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \text{for } x,y \in \mathbb{R}.$$

Then it follows that

$$\begin{aligned} F_{X+Y}(x) &= P(X+Y \le x) \\ &= \int\int_{s+t\le x} f_{X,Y}(s,t)\,\mathrm{ds}\,\mathrm{dt} \\ &= \int\int_{s+t\le x} f_X(s)f_Y(t)\,\mathrm{ds}\,\mathrm{dt} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{x-t} f_X(s)f_Y(t)\,\mathrm{ds}\,\mathrm{dt} \\ &= \int_{-\infty}^{\infty} F_X(x-t)f_Y(t)\,\mathrm{dt}\,. \end{aligned}$$

Similarly, we can show that

$$F_{X+Y}(x) = \int_{-\infty}^{\infty} F_Y(x-t)f_X(t)\,\mathrm{dt}\,.$$

It then also follows that

$$\begin{aligned} f_{X+Y}(x) &= \frac{d}{\mathrm{dx}}F_{X+Y}(x) \\ &= \int_{-\infty}^{\infty} \frac{d}{\mathrm{dx}}F_X(x-t)f_Y(t)\,\mathrm{dt} \\ &= \int_{-\infty}^{\infty} f_X(x-t)f_Y(t)\,\mathrm{dt}\,. \end{aligned}$$

(Summary)

$$F_{X+Y}(x) = \int_{-\infty}^{\infty} F_X(x-t) f_Y(t)\,\mathrm{dt} = \int_{-\infty}^{\infty} F_Y(x-t) f_X(t)\,\mathrm{dt}.$$

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(x-t) f_Y(t)\,\mathrm{dt} = \int_{-\infty}^{\infty} f_X(t) f_Y(x-t)\,\mathrm{dt}.$$

Example 6.22 (Sum of 2 Independent Uniform Random Variables). Suppose that X and Y are independent with a common uniform distribution over $(0,1)$. Find the probability density function of $X+Y$.

Solution:
Recall that

$$f_X(t) = f_Y(t) = \begin{cases} 1, & \text{if } 0 \le t \le 1 \\ 0, & \text{otherwise} \end{cases}.$$

Also note that $X+Y$ takes values in $(0,2)$. For $x \le 0$ or $x \ge 2$, it then follows that $f_{X+Y}(x) = 0$.
Next, for $0 < x < 2$,

$$\begin{aligned} f_{X+Y}(x) &= \int_{-\infty}^{\infty} f_X(x-t) f_Y(t)\,\mathrm{dt} \\ &= \int_0^1 f_X(x-t) \times 1\,\mathrm{dt} \\ &= \int_0^1 f_X(x-t)\,\mathrm{dt}. \end{aligned}$$

We see that $f_X(x-t) > 0$ (in this case, it is 1) if and only if $0 < x-t < 1$ (remember that x is fixed, and t varies in $(0,1)$). To proceed, we need to separate the range of x into 2 cases:
Case 1: for $0 < x \le 1$.

$$\begin{aligned} f_{X+Y}(x) &= \int_0^1 f_X(x-t)\,\mathrm{dt} \\ &= \int_0^x f_X(x-t)\,\mathrm{dt} + \int_x^1 f_X(x-t)\,\mathrm{dt} \\ &= \int_0^x 1\,\mathrm{dt} + 0 \\ &= x. \end{aligned}$$

Case 2: for $1 < x < 2$.

$$\begin{aligned} f_{X+Y}(x) &= \int_0^1 f_X(x-t)\,\mathrm{dt} \\ &= \int_0^{x-1} f_X(x-t)\,\mathrm{dt} + \int_{x-1}^1 f_X(x-t)\,\mathrm{dt} \\ &= \int_{x-1}^1 1\,\mathrm{dt} \\ &= 2-x. \end{aligned}$$

In the second last equality, we use the fact that $0 < x-t < 1$ is equivalent to $x-1 < t < x$; also we need $0 < t < 1$.
Summing up:

$$f_{X+Y}(x) = \begin{cases} x, & 0 < x \leq 1 \\ 2-x, & 1 < x < 2 \\ 0, & \text{elsewhere} \end{cases}.$$

The density function has the shape of a triangle, and so the random variable $X+Y$ is sometimes known as the **triangular distribution**.

Proposition 6.23 (Sum of 2 Independent Gamma Random Variables)**.** *Assume that $X \sim \Gamma(\alpha,\lambda)$ and $Y \sim \Gamma(\beta,\lambda)$, and X and Y are mutually independent. Then,*

$$X+Y \sim \Gamma(\alpha+\beta,\ \lambda).$$

Note that both X and Y must have the same second parameter.

Proof. Note that for $w > 0$, $f_{X+Y}(w) = \int_{-\infty}^{\infty} f_X(w-y) f_Y(y)\,\mathrm{dy}$.
And so, for $w > 0$,

$$\begin{aligned} f_{X+Y}(w) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\lambda^\beta}{\Gamma(\beta)} \int_0^w (w-y)^{\alpha-1} e^{-\lambda(w-y)} y^{\beta-1} e^{-\lambda y}\,\mathrm{dy} \\ &= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda w} \int_0^w (w-y)^{\alpha-1} y^{\beta-1}\,\mathrm{dy} \\ &= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda w} w^{\alpha+\beta-1} \int_0^1 (1-u)^{\alpha-1} u^{\beta-1}\,\mathrm{du} \quad (\text{letting } u = y/w) \\ &= \left(\frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 (1-u)^{\alpha-1} u^{\beta-1}\,\mathrm{du} \right) w^{\alpha+\beta-1} e^{-\lambda w}. \end{aligned}$$

Let

$$K = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 (1-u)^{\alpha-1} u^{\beta-1}\,\mathrm{du}$$

and note that K is a constant that does not depend on w. As the above is a density function and must integrate to 1, the value of K is hence determined, and we have

$$1 = K \int_0^\infty w^{\alpha+\beta-1} e^{-\lambda w} dw = K \times \frac{\Gamma(\alpha+\beta)}{\lambda^{\alpha+\beta}}.$$

Not only do we get the probability density function of $X+Y$, which is given by

$$f_{X+Y}(w) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha+\beta)} w^{\alpha+\beta-1} e^{-\lambda w}, \quad w > 0,$$

but also we obtain a by-product, which is the following identity

$$\int_0^1 (1-u)^{\alpha-1} u^{\beta-1} \, \mathrm{du} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The integral on the left is also known as the **Beta function** with parameters $\alpha > 0$ and $\beta > 0$. Indeed, the beta function, denoted by $B(\alpha, \beta)$, where for $\alpha > 0$ and $\beta > 0$, is defined to be

$$B(\alpha,\ \beta) = \int_0^1 (1-u)^{\alpha-1} u^{\beta-1} \, \mathrm{du}\,.$$

□

Example 6.24. Let $X_1, X_2, \ldots, X_n$ be n independent exponential random variables each having parameter λ. Then, as an exponential random variable with parameter λ is the same as a Gamma random variable with parameters $(1,\ \lambda)$, we see from Proposition 6.23 that $X_1 + X_2 + \cdots + X_n$ is a Gamma random variable with parameters $(n,\ \lambda)$.

Proposition 6.25 (Sum of Independent Normal Random Variables)**.** *If X_i, $i = 1, \ldots, n$ are independent random variables that are normally distributed with respective parameters μ_i, σ_i^2, $i = 1, \ldots, n$, then $\sum_{i=1}^n X_i$ is normally distributed with parameters $\sum_{i=1}^n \mu_i$ and $\sum_{i=1}^n \sigma_i^2$.*

Proof. To begin, let X and Y be independent normal random variables, with X having mean 0 and variance σ^2, and Y having mean 0 and variance 1. We will determine the density function of $X+Y$. Now, with

$$c = \frac{1}{2\sigma^2} + \frac{1}{2} = \frac{1+\sigma^2}{2\sigma^2}$$

we have

$$f_X(a-y)f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(a-y)^2}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)$$
$$= \frac{1}{2\pi\sigma}\exp\left(-\frac{a^2}{2\sigma^2}\right)\exp\left(-c\left(y^2-2y\frac{a}{1+\sigma^2}\right)\right).$$

Hence,

$$f_{X+Y}(a) = \frac{1}{2\pi\sigma}\exp\left(-\frac{a^2}{2\sigma^2}\right)\exp\left(\frac{a^2}{2\sigma^2(1+\sigma^2)}\right)\int_{-\infty}^{\infty}\exp\left(-c\left(y-\frac{a}{1+\sigma^2}\right)^2\right)\mathrm{dy}$$
$$= \frac{1}{2\pi\sigma}\exp\left(-\frac{a^2}{2(1+\sigma^2)}\right)\int_{-\infty}^{\infty}\exp\left(-cx^2\right)\mathrm{dx}$$
$$= C\exp\left(-\frac{a^2}{2(1+\sigma^2)}\right)$$

where C doesn't depend on a. But this implies that $X+Y$ is normal with mean 0 and variance $1+\sigma^2$.

Now, suppose that X_1 and X_2 are independent normal random variables, with X_i having mean μ_i and variance σ_i^2, $i = 1,2$. Then

$$X_1 + X_2 = \sigma_2\left(\frac{X_1-\mu_1}{\sigma_2} + \frac{X_2-\mu_2}{\sigma_2}\right) + \mu_1 + \mu_2.$$

But since $(X_1-\mu_1)/\sigma_2$ is normal with mean 0 and variance σ_1^2/σ_2^2, and $(X_2-\mu_2)/\sigma_2$ is normal with mean 0 and variance 1, it follows from our previous result that $(X_1-\mu_1)/\sigma_2 + (X_2-\mu_2)/\sigma_2$ is normal with mean 0 and variance $1+\sigma_1^2/\sigma_2^2$, implying that X_1+X_2 is normal with mean $\mu_1+\mu_2$ and variance $\sigma_2^2(1+\sigma_1^2/\sigma_2^2) = \sigma_1^2+\sigma_2^2$.

Thus, this proposition is established when $n = 2$. The general case now follows by induction. That is, assume that it is true when there are $n-1$ random variables. Now consider the case of n, and write

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n-1} X_i + X_n.$$

By the induction hypothesis, $\sum_{i=1}^{n-1} X_i$ is normal with mean $\sum_{i=1}^{n-1}\mu_i$ and variance $\sum_{i=1}^{n-1}\sigma_i^2$. Therefore, by the result for $n = 2$, we can conclude that $\sum_{i=1}^{n} X_i$ is normal with mean $\sum_{i=1}^{n}\mu_i$ and variance $\sum_{i=1}^{n}\sigma_i^2$. □

Example 6.26. A basketball team will play a 44-game season. Twenty six of these games are against class A teams and 18 are against class B teams. Suppose that the team will win each game against a class A team with probability 0.4 and will win each game against a class B team with probability 0.7. Suppose also that the results of the different games are independent. Approximate the probability that

(a) the team wins 25 games or more;

(b) the team wins more games against class A teams that it does against class B teams.

Solution:

(a) Let X_A and X_B denote respectively the number of games the team wins against class A and against class B teams. Note that X_A and X_B are independent binomial random variables and

$$E(X_A) = 26(0.4) = 10.4, \quad \text{var}(X_A) = 26(0.4)(0.6) = 6.24,$$
$$E(X_B) = 18(0.7) = 12.6, \quad \text{var}(X_B) = 18(0.7)(0.3) = 3.78.$$

By the normal approximation to the binomial, X_A and X_B will have approximately the same distribution as would independent normal random variables with the given parameters. So by one of the preceding examples, $X_A + X_B$ will have approximately a normal distribution with mean 23 and variance 10.02. So we have

$$\begin{aligned} P(X_A + X_B \geq 25) &= P(X_A + X_B \geq 24.5) \\ &= P\left(\frac{X_A + X_B - 23}{\sqrt{10.2}} \geq \frac{24.5 - 23}{\sqrt{10.2}}\right) \\ &\approx P\left(Z \geq \frac{1.5}{\sqrt{10.2}}\right) \\ &\approx 1 - P(Z < 0.4739) \\ &\approx 0.3178. \end{aligned}$$

(b) We note that $X_A - X_B$ will have approximately a normal distribution with mean -2.2 and variance 10.02. So we have

$$\begin{aligned} P(X_A - X_B \geq 1) &= P(X_A - X_B \geq 0.5) \\ &= P\left(\frac{X_A - X_B + 2.2}{\sqrt{10.2}} \geq \frac{0.5+2.2}{\sqrt{10.2}}\right) \\ &\approx P\left(Z \geq \frac{2.7}{\sqrt{10.2}}\right) \\ &\approx 1 - P(Z < 0.8530) \\ &\approx 0.1968. \end{aligned}$$

So there is approximately a 31.78 percent chance that the team will win at least 25 games and approximately a 19.68 percent chance that the team will win more games against class A teams than against class B teams.

6.3.2 X and Y are discrete and independent

Example 6.27 (Sum of 2 Independent Poisson Random Variables).

$$X \sim \text{Poisson}(\lambda), \quad Y \sim \text{Poisson}(\mu), \quad X, Y \text{ independent.}$$

Find probability mass function of $X+Y$.

Solution:
First note that $X+Y$ takes values $0, 1, 2, \ldots$.
For $n = 0, 1, \ldots,$

$$\begin{aligned} P(X+Y=n) &= \sum_{k=0}^{n} P(X=k,\, Y=n-k) \\ &= \sum_{k=0}^{n} P(X=k)P(Y=n-k) \quad \text{by independence} \\ &= \sum_{k=0}^{n} \frac{e^{-\lambda}\lambda^k}{k!} \times \frac{e^{-\mu}\mu^{n-k}}{(n-k)!} \\ &= e^{-(\lambda+\mu)} \sum_{k=0}^{n} \frac{\lambda^k \mu^{n-k}}{(n-k)!k!} \\ &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda^k \mu^{n-k} \\ &= \frac{e^{-(\lambda+\mu)}(\lambda+\mu)^n}{n!}. \end{aligned}$$

Conclusion: Sum of 2 independent Poisson random variables is still Poisson. The mean is the sum of the means.

Example 6.28 (Sum of 2 Independent Binomial Random Variables).

$$X \sim \text{Bin}(n,p), \quad Y \sim \text{Bin}(m,p), \quad X,Y \text{ independent.}$$

Find probability mass function of $X+Y$.
Note: p is the same in X and in Y.

Solution:
First note that $X+Y$ takes values $0,1,2,\ldots,n+m$.
For $k=0,1,\ldots,n+m$,

$$\begin{aligned}
P(X+Y=k) &= \sum_{i=0}^{k} P(X=i,\, Y=k-i) \\
&= \sum_{i=0}^{k} P(X=i)P(Y=k-i) \quad \text{by independence} \\
&= \sum_{i=0}^{k} \binom{n}{i} p^i q^{n-i} \times \binom{m}{k-i} p^{k-i} q^{m-(k-i)} \\
&= p^k q^{n+m-k} \sum_{i=0}^{k} \binom{n}{i}\binom{m}{k-i} \\
&= \binom{n+m}{k} p^k q^{n+m-k}.
\end{aligned}$$

Conclusion: Sum of 2 independent Binomial random variables with the same success probability is still Binomial with parameters $n+m$ and p.

Example 6.29 (Sum of 2 Independent Geometric Random Variables). Let $X \sim \text{Geom}(p)$ and $Y \sim \text{Geom}(p)$, and assume that X and Y are independent. Determine the distribution of $X+Y$.

Solution:

First of all, observe that $X+Y$ takes values $2,3,\ldots$. For $k \geq 2$,

$$\begin{aligned} f_{X+Y}(k) &= \sum_{i=1}^{k-1} P(X=i,\, Y=k-i) \\ &= \sum_{i=1}^{k-1} P(X=i)P(Y=k-i) \quad \text{by independence} \\ &= \sum_{i=1}^{k-1} pq^{i-1} \cdot pq^{k-i-1} \\ &= \sum_{i=1}^{k-1} p^2 q^{k-2} = (k-1)p^2 q^{k-2} \\ &= \binom{k-1}{1} p^2 q^{k-2}, \end{aligned}$$

which is the probability mass function of a negative binomial random variable with parameters $(2,p)$.
Conclusion: For independent $X \sim \text{Geom}(p), Y \sim \text{Geom}(p)$,

$$X+Y \sim NB(2,p).$$

6.4 Conditional distributions: Discrete Case

Recall that

$$P(B|A) := \frac{P(AB)}{P(A)} \quad \text{if } P(A) > 0.$$

The **conditional probability mass function** of X given that $Y=y$ is defined by

$$\begin{aligned} p_{X|Y}(x|y) &:= P(X=x|Y=y) \\ &= \frac{P(X=x,\, Y=y)}{P(Y=y)} \\ &= \frac{p_{X,Y}(x,y)}{p_Y(y)} \end{aligned}$$

for all values of y such that $p_Y(y) > 0$.
Similarly, the **conditional distribution function** of X given that $Y=y$ is defined by

$$F_{X|Y}(x|y) = P(X \leq x|Y=y) \quad \text{for } y \text{ such that } p_Y(y) > 0.$$

It now follows that

$$\begin{aligned} F_{X|Y}(x|y) &:= \frac{P(X \leq x,\, Y = y)}{P(Y = y)} \\ &= \frac{\sum_{a \leq x} p_{X,Y}(a,y)}{p_Y(y)} \\ &= \sum_{a \leq x} \frac{p_{X,Y}(a,y)}{p_Y(y)} \\ &= \sum_{a \leq x} p_{X|Y}(a|y). \end{aligned}$$

Note that the definitions are exactly parallel to the unconditional case.

Proposition 6.30. *If X is independent of Y, then the conditional probability mass function of X given $Y = y$ is the same as the marginal probability mass function of X for every y such that $p_Y(y) > 0$.*

Proof. For y such that $p_Y(y) > 0$,

$$\begin{aligned} p_{X|Y}(x|y) &:= \frac{p_{X,Y}(x,y)}{p_Y(y)} \\ &= \frac{p_X(x)p_Y(y)}{p_Y(y)} \\ &= p_X(x). \end{aligned}$$

□

Example 6.31. Suppose that $p(x,y)$, the joint probability mass function of X and Y, is given by

$$p(0,0) = 0.4,\ p(0,1) = 0.2,\ p(1,0) = 0.1,\ p(1,1) = 0.3.$$

(i) Calculate the conditional probability mass function of X given that $Y = 1$.

(ii) Calculate the conditional probability mass function of X given that $Y = 0$.

Solution:
First note that

$$p_Y(1) = p(0,1) + p(1,1) = 0.5$$

and

$$p_Y(0) = p(0,0) + p(1,0) = 0.5.$$

(i)

$$\begin{aligned} p_{X|Y}(0|1) &= \frac{p(0,1)}{p_Y(1)} = \frac{0.2}{0.5} = \frac{2}{5}; \\ p_{X|Y}(1|1) &= \frac{p(1,1)}{p_Y(1)} = \frac{0.3}{0.5} = \frac{3}{5}. \end{aligned}$$

(ii)

$$\begin{aligned} p_{X|Y}(0|0) &= \frac{p(0,0)}{p_Y(0)} = \frac{0.4}{0.5} = \frac{4}{5}; \\ p_{X|Y}(1|0) &= \frac{p(1,0)}{p_Y(0)} = \frac{0.1}{0.5} = \frac{1}{5}. \end{aligned}$$

Example 6.32. If X and Y are independent Poisson random variables with respective parameters λ and μ, calculate the conditional distribution of X given that $X+Y=n$. That is, find

$$P(X \le k | X+Y=n) \quad \text{for all values of } k.$$

Solution:
As

$$P(X \le k | X+Y=n) = \sum_{j=0}^{k} P(X=j | X+Y=n),$$

we first calculate

$$\begin{aligned} P(X=j|X+Y=n) &= \frac{P(X=j,\, X+Y=n)}{P(X+Y=n)} \\ &= \frac{P(X=j,\, Y=n-j)}{P(X+Y=n)} \\ &= \frac{P(X=j)P(Y=n-j)}{P(X+Y=n)} \\ &= \frac{\frac{e^{-\lambda}\lambda^j}{j!} \times \frac{e^{-\mu}\mu^{n-j}}{(n-j)!}}{\frac{e^{-(\lambda+\mu)}(\lambda+\mu)^n}{n!}} \\ &= \binom{n}{j} \left(\frac{\lambda}{\lambda+\mu}\right)^j \left(1-\frac{\lambda}{\lambda+\mu}\right)^{n-j}. \end{aligned}$$

This indicates that X conditioned on $X+Y=n$ is $\text{Bin}(n, \lambda/(\lambda+\mu))$. Hence

$$\begin{aligned} &P(X \le k | X+Y=n) \\ &= \sum_{j=0}^{k} \binom{n}{j} \left(\frac{\lambda}{\lambda+\mu}\right)^j \left(1-\frac{\lambda}{\lambda+\mu}\right)^{n-j}. \end{aligned}$$

6.5 Conditional distributions: Continuous Case

Suppose that X and Y are jointly continuous random variables. We define the **conditional probability density function** of X given that $Y = y$ as

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

for all y such that $f_Y(y) > 0$.
For motivation of this definition, see Ross p.266.
The use of conditional densities allows us to define conditional probabilities of events associated with one random variable when we are given the value of a second random variable.
That is, for $A \subset \mathbb{R}$ and y such that $f_Y(y) > 0$,

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)\,\mathrm{dx}\,.$$

In particular, the **conditional distribution function** of X given that $Y = y$ is defined by

$$F_{X|Y}(x|y) = P(X \leq x|Y = y) = \int_{-\infty}^{x} f_{X|Y}(t|y)\,\mathrm{dt}\,.$$

Proposition 6.33. *If X is independent of Y, then the conditional probability density function of X given $Y = y$ is the same as the marginal probability density function of X for every y such that $f_Y(y) > 0$.*

Proof. For y such that $f_Y(y) > 0$,

$$\begin{aligned} f_{X|Y}(x|y) &:= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{f_X(x)f_Y(y)}{f_Y(y)} \\ &= f_X(x). \end{aligned}$$

□

Example 6.34. Suppose that the joint probability density function of X and Y is given by

$$f(x,y) = \begin{cases} \frac{15}{2}x(2-x-y), & 0 < x < 1,\ 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}.$$

Compute the conditional probability density function of X given that $Y = y$ where $0 < y < 1$.

Solution:
For $0 < y < 1$, we have

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{dx} \\ &= \int_0^1 \frac{15}{2} x(2-x-y)\,\mathrm{dx} \\ &= \frac{15}{2}\left[\frac{2}{3} - \frac{y}{2}\right]. \end{aligned}$$

Therefore, for $0 < y < 1$

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \begin{cases} \frac{6x(2-x-y)}{4-3y}, & \text{for } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}.$$

Example 6.35. Suppose that the joint probability density function of X and Y is given by

$$f(x,y) = \begin{cases} \frac{e^{-x/y}e^{-y}}{y}, & 0 < x < \infty,\ 0 < y < \infty \\ 0, & \text{otherwise} \end{cases}.$$

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

Solution:
For $y \leq 0$, $f_Y(y) = 0$ and $P(X > 1|Y = y)$ is not defined.
For $y > 0$,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{dx} \\ &= \int_0^{\infty} \frac{e^{-x/y}e^{-y}}{y}\,\mathrm{dx} \\ &= e^{-y}\int_0^{\infty} \frac{1}{y} e^{-x/y}\,\mathrm{dx} \\ &= e^{-y}. \end{aligned}$$

Therefore, for $y > 0$,

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{y}e^{-x/y}, & x > 0 \\ 0, & x \leq 0 \end{cases}.$$

Hence,

$$\begin{aligned} P(X > 1|Y = y) &= \int_1^\infty f_{X|Y}(x|y)\,\mathrm{dx} \\ &= \frac{1}{y}\int_1^\infty e^{-x/y}\,\mathrm{dx} \\ &= e^{-1/y}. \end{aligned}$$

Example 6.36 (The Bivariate Normal Distribution). One of the most important joint distributions is the bivariate normal distribution. We say that the random variables X, Y have a bivariate normal distribution if, for constants μ_x, μ_y, $\sigma_x > 0$, $\sigma_y > 0$, $-1 < \rho < 1$, their joint density function is given, for all $-\infty < x, y < \infty$, by

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \quad \times$$
$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right).$$

We now determine the conditional density of X given that $Y = y$. In doing so, we will continually collect all factors that do not depend on x and represent them by the constants C_j. The final constant will then be found by using that $\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,\mathrm{dx} = 1$. We have

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x,y)}{f_Y(y)} \\ &= C_1 f(x,y) \\ &= C_2 \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x(y-\mu_y)}{\sigma_x\sigma_y}\right]\right) \\ &= C_3 \exp\left(-\frac{1}{2\sigma_x^2(1-\rho^2)}\left[x^2 - 2x\left(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)\right)\right]\right) \\ &= C_4 \exp\left(-\frac{1}{2\sigma_x^2(1-\rho^2)}\left[x - \left(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)\right)\right]^2\right). \end{aligned}$$

Recognizing the preceding equation as a normal density, we can conclude that, given $Y = y$, the random variable X is normally distributed with mean $\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)$ and variance $\sigma_x^2(1-\rho^2)$. Also, because the joint density of Y, X is exactly they same as that of X, Y, except that μ_x, σ_x are interchanged

with μ_y, σ_y, it similarly follows that the conditional distribution of Y given $X = x$ is the normal distribution with mean $\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_X)$ and variance $\sigma_y^2(1-\rho^2)$. It follows from these results that the necessary and sufficient condition for the bivariate normal random variables X and Y to be independent is that $\rho = 0$ (a result that also follows directly from their joint density, because it is only when $\rho = 0$ that the joint density factors into two terms, one depending only on x and the other only on y).
With $C = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$, the marginal density of X can be obtained from

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{dy} \\ &= C\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right)\mathrm{dy}\,. \end{aligned}$$

Making the change of variables $w = \dfrac{y-\mu_y}{\sigma_y}$ gives

$$\begin{aligned} f_X(x) &= C\sigma_y \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_x}{\sigma_x^2}\right)^2\right) \\ &\quad \times \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}\left[w^2 - 2\rho\frac{x-\mu_x}{\sigma_x}w\right]\right)\mathrm{dw} \\ &= C\sigma_y \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_x}{\sigma_x}\right)^2(1-\rho^2)\right) \\ &\quad \times \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}\left[w - \rho\frac{x-\mu_x}{\sigma_x}\right]^2\right)\mathrm{dw}\,. \end{aligned}$$

Because

$$\frac{1}{\sqrt{2\pi(1-\rho^2)}}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}\left[w - \frac{\rho}{\sigma_x}(x-\mu_x)\right]^2\right)\mathrm{dw} = 1$$

we see that

$$\begin{aligned} f_X(x) &= C\sigma_y\sqrt{2\pi(1-\rho^2)}\,e^{-(x-\mu_x)^2/2\sigma_x^2} \\ &= \frac{1}{\sqrt{2\pi}\sigma_x}e^{-(x-\mu_x)^2/2\sigma_x^2}. \end{aligned}$$

That is, X is normal with mean μ_x and variance σ_x^2. Similarly, Y is normal with mean μ_y and variance σ_y^2.

6.6 Joint Probability Distribution Function of Functions of Random Variables

Let X and Y be jointly distributed random variables with joint probability density function $f_{X,Y}$. It is sometimes necessary to obtain the joint distribution of the random variables U and V, which arise as functions of X and Y.
Specifically, suppose that

$$U = g(X,Y) \quad \text{and} \quad V = h(X,Y),$$

for some functions g and h.

We want to find the joint probability density function of U and V in terms of the joint probability density function $f_{X,Y}$, g and h.
For example, X and Y are independent exponentially distributed random variables, and we are interested to know the joint probability density function of $U = X+Y$ and $V = X/(X+Y)$. In this case,

$$g(x,y) = x+y \quad \text{and} \quad h(x,y) = x/(x+y).$$

Assume that the following conditions are satisfied:

(i) Let X and Y be jointly continuously distributed random variables with known joint probability density function.

(ii) Let U and V be given functions of X and Y in the form:

$$U = g(X,Y), \quad V = h(X,Y).$$

And we can uniquely solve X and Y in terms of U and V, say $x = a(u,v)$ and $y = b(u,v)$.

(iii) The functions g and h have continuous partial derivatives and

$$J(x,y) := \begin{vmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{vmatrix} = \frac{\partial g}{\partial x}\frac{\partial h}{\partial y} - \frac{\partial g}{\partial y}\frac{\partial h}{\partial x} \neq 0.$$

Conclusion: The joint probability density function of U and V is given by

$$f_{U,V}(u,v) = f_{X,Y}(x,y)|J(x,y)|^{-1},$$

where $x = a(u,v)$ and $y = b(u,v)$.

Remark 33. $J(x,y)$ is known as the **Jacobian determinant** of g and h.

Example 6.37. Let X_1 and X_2 be jointly continuous random variables with probability density function f_{X_1,X_2}. Let $Y_1 = X_1 + X_2, Y_2 = X_1 - X_2$. Find the joint density function of Y_1 and Y_2 in terms of f_{X_1,X_2}.

Solution:
Let $g_1(x_1,x_2) = x_1 + x_2$ and $g_2(x_1,x_2) = x_1 - x_2$. Then

$$J(x_1,\ x_2) = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2.$$

Also, as the equations $y_1 = x_1 + x_2$ and $y_2 = x_1 - x_2$ have as their solution $x_1 = (y_1 + y_2)/2$, $x_2 = (y_1 - y_2)/2$, it follows that the desired density is

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2} f_{X_1,X_2}\left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right).$$

For instance, if X_1 and X_2 are independent, uniform $(0,1)$ random variables, then

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \frac{1}{2}, & 0 \le y_1 + y_2 \le 2, 0 \le y_1 - y_2 \le 2 \\ 0, & \text{otherwise} \end{cases}.$$

If X_1 and X_2 were independent, exponential random variables with respective parameters λ_1 and λ_2, then

$$f_{Y_1,Y_2}(y_1,y_2)$$
$$= \begin{cases} \frac{\lambda_1\lambda_2}{2} \exp\left(-\lambda_1\left(\frac{y_1+y_2}{2}\right) - \lambda_2\left(\frac{y_1-y_2}{2}\right)\right), & y_1 + y_2 \ge 0, y_1 - y_2 \ge 0 \\ 0, & \text{otherwise} \end{cases}.$$

Finally, if X_1 and X_2 are independent standard normal random variables,

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= \frac{1}{4\pi} e^{-\left[(y_1+y_2)^2/8 + (y_1-y_2)^2/8\right]} \\ &= \frac{1}{4\pi} e^{-(y_1^2+y_2^2)/4} \\ &= \frac{1}{\sqrt{4\pi}} e^{-y_1^2/4} \frac{1}{\sqrt{4\pi}} e^{-y_2^2/4}. \end{aligned}$$

Thus, not only do we obtain (in agreement with Proposition 6.25) that both $X_1 + X_2$ and $X_1 - X_2$ are normal with mean 0 and variance 2 we also obtain the interesting result that these two random variables are independent.

Example 6.38. Let X and Y be jointly distributed with joint probability density function given as

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

That is, X and Y are independent standard normal.
Let random variables R and Θ denote the polar coordinates of the point (x,y). That is,

$$R = \sqrt{X^2+Y^2} \quad ; \quad \Theta = \tan^{-1}\left(\frac{Y}{X}\right).$$

(i) Find the joint probability density function of R and Θ.

(ii) Show that R and Θ are independent.

Solution:

(i) To find the joint probability density function of R and Θ, consider

Step 1. Random variables R and Θ takes values in $(0,\infty)$ and in $(0,2\pi)$.

Step 2. Now $r = g(x,y) := \sqrt{x^2+y^2}$ and $\theta = h(x,y) := \tan^{-1}\left(\frac{y}{x}\right)$, so

$$x = r\cos\theta \text{ and } y = r\sin\theta.$$

The transformed region is

$$0 < r < \infty \text{ and } 0 < \theta < 2\pi.$$

Step 3.

$$\begin{aligned} J(x,y) &:= \begin{vmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{vmatrix} \\ &= \begin{vmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{vmatrix} \\ &= (x^2+y^2)^{-1/2}. \end{aligned}$$

Step 4. Therefore, for $0 < r < \infty$ and $0 < \theta < 2\pi$,

$$\begin{aligned} f_{R,\Theta}(r,\theta) &= f_{X,Y}(x,y)/|J(x,y)| \\ &= \sqrt{x^2+y^2} f_{X,Y}(x,y) \\ &= \sqrt{x^2+y^2} \frac{1}{2\pi} e^{-(x^2+y^2)/2} \\ &= r\frac{1}{2\pi} e^{-r^2/2}. \end{aligned}$$

This solves (i).

(ii) Notice that

$$f_{R,\Theta}(r,\theta) = \left[re^{-r^2/2}\right] \times \frac{1}{2\pi}$$

for all r and θ, by Proposition 6.14, R and Θ are independent.

Indeed,

$$f_R(r) = re^{-r^2/2}, \quad \text{for } 0 < r < \infty$$

and

$$f_\Theta(\theta) = \frac{1}{2\pi}, \quad \text{for } 0 < \theta < 2\pi.$$

Remark 34. R is said to have the **Rayleigh** distribution.

Example 6.39. If X and Y are independent Gamma random variables with parameters (α,λ) and (β,λ), respectively, compute the joint density of $U = X+Y$ and $V = X/(X+Y)$.

Solution:
The joint density of X and Y is given by

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \times \frac{\lambda e^{-\lambda y}(\lambda y)^{\beta-1}}{\Gamma(\beta)} \\ &= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda(x+y)} x^{\alpha-1} y^{\beta-1}. \end{aligned}$$

Now, if $g_1(x,y) = x+y$, $g_2(x,y) = x/(x+y)$, then

$$J(x,y) = \begin{vmatrix} 1 & 1 \\ \frac{y}{(x+y)^2} & \frac{-x}{(x+y)^2} \end{vmatrix} = -\frac{1}{x+y}.$$

Finally, as the equations $u = x+y, v = x/(x+y)$ have as their solutions $x = uv$, $y = u(1-v)$, we see that

$$f_{U,V}(u,v) = f_{X,Y}[uv, u(1-v)] \times u$$
$$= \frac{\lambda e^{-\lambda u}(\lambda u)^{\alpha+\beta-1}}{\Gamma(\alpha+\beta)} \times \frac{v^{\alpha-1}(1-v)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

Hence $X+Y$ and $X/(X+Y)$ are independent, with $X+Y$ having a Gamma distribution with parameters $(\alpha+\beta,\lambda)$ and $X/(X+Y)$ is said to have a Beta distribution with parameters (α,β). The above also shows that $B(\alpha,\beta)$, the normalizing factor in the Beta density, is such that

$$B(\alpha,\ \beta) = \int_0^1 v^{\alpha-1}(1-v)^{\beta-1}\,\mathrm{dv} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Remark 35. The result above is quite interesting. For suppose there are $n+m$ jobs to be performed, with each (independently) taking an exponential amount of time with rate λ for performance, and suppose that we have two workers to perform these jobs. Worker I will do jobs $1,\ldots,n$, and worker II will do the remaining m jobs. If we let X and Y denote the total working times of workers I and II, respectively, then X and Y will be independent Gamma random variables having parameters (n,λ) and (m,λ), respectively. Then the above result yields that independently of the working time needed to complete $n+m$ jobs (that is, of $X+Y$), the proportion of this work that will be performed by worker I has a Beta distribution with parameters (n,m).

When the joint density function of the n random variables $X_1,X_2,\ldots,X_n$ is given and we want to compute the joint density function of $Y_1,Y_2,\ldots,Y_n$, where

$$Y_1 = g_1(X_1,\ldots,X_n),\ Y_2 = g_2(X_1,\ldots,X_n),\ \ldots,\ Y_n = g_n(X_1,\ldots,X_n),$$

the approach is the same.
Namely, we assume that the functions g_j have continuous partial derivatives and that the Jacobian determinant $J(x_1,\ldots,x_n) \neq 0$ at all points $(x_1,\ldots,x_n)$, where

$$J(x_1,\ldots,x_n) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_n} \end{vmatrix}.$$

Furthermore, we suppose that the equations $y_1 = g_1(x_1,\ldots,x_n)$, $y_2 = g_2(x_1,\ldots,x_n)$, …, $y_n = g_n(x_1,\ldots,x_n)$ have a unique solution, say, $x_1 = h_1(y_1,\ldots,y_n)$, …, $x_n = h_n(y_1,\ldots,y_n)$. Under these assumptions, the joint density function of the random variables Y_i is given by

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)|J(x_1,\ldots,x_n|^{-1},$$

where $x_i = h_i(y_1,\ldots,y_n)$, $i = 1,2,\ldots,n$.

Example 6.40. Let X_1, X_2, and X_3 be independent standard normal random variables. If $Y_1 = X_1 + X_2 + X_3$, $Y_2 = X_1 - X_2$, $Y_3 = X_1 - X_3$, compute the joint density function of Y_1, Y_2, Y_3.

Solution:
Letting $Y_1 = X_1 + X_2 + X_3$, $Y_2 = X_1 - X_2$, $Y_3 = X_1 - X_3$, the Jacobian of these transformations is given by

$$J = \begin{vmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{vmatrix} = 3.$$

As the transformations above yield that $X_1 = \frac{Y_1+Y_2+Y_3}{3}$, $X_2 = \frac{Y_1-2Y_2+Y_3}{3}$, $X_3 = \frac{Y_1+Y_2-2Y_3}{3}$, we see that

$$\begin{aligned} &f_{Y_1,Y_2,Y_3}(y_1,y_2,y_3) \\ &= \frac{1}{3} f_{X_1,X_2,X_3}\left(\frac{y_1+y_2+y_3}{3}, \frac{y_1-2y_2+y_3}{3}, \frac{y_1+y_2-2y_3}{3}\right). \end{aligned}$$

Hence, as

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = \frac{1}{(2\pi)^{3/2}} e^{-\Sigma_{i=1}^3 x_i^2/2}$$

we see that

$$f_{Y_1,Y_2,Y_3}(y_1,y_2,y_3) = \frac{1}{3(2\pi)^{3/2}} e^{-Q(y_1,y_2,y_3)/2}$$

where

$$\begin{aligned} Q(y_1,y_2,y_3) &= \left(\frac{y_1+y_2+y_3}{3}\right)^2 + \left(\frac{y_1-2y_2+y_3}{3}\right)^2 + \left(\frac{y_1+y_2-2y_3}{3}\right)^2 \\ &= \frac{y_1^2}{3} + \frac{2}{3}y_2^2 + \frac{2}{3}y_3^2 - \frac{2}{3}y_2y_3. \end{aligned}$$

6.7 Jointly Distributed Random Variables: $n \geq 3$

We will illustrate the results for 3 jointly distributed random variables, called X, Y and Z. We will assume they are jointly continuous random variables.

$$F_{X,Y,Z}(x,y,z) := P(X \leq x,\ Y \leq y,\ Z \leq z).$$

There are a number of marginal distribution functions, namely

$$\begin{aligned}
F_{X,Y}(x,y) &:= \lim_{z\to\infty} F_{X,Y,Z}(x,y,z);\\
F_{X,Z}(x,z) &:= \lim_{y\to\infty} F_{X,Y,Z}(x,y,z);\\
F_{Y,Z}(y,z) &:= \lim_{x\to\infty} F_{X,Y,Z}(x,y,z);\\
F_X(x) &:= \lim_{y\to\infty,\ z\to\infty} F_{X,Y,Z}(x,y,z);\\
F_Y(y) &:= \lim_{x\to\infty,\ z\to\infty} F_{X,Y,Z}(x,y,z);\\
F_Z(z) &:= \lim_{x\to\infty,\ y\to\infty} F_{X,Y,Z}(x,y,z).
\end{aligned}$$

Joint probability density function of X, Y and Z: $f_{X,Y,Z}(x,y,z)$

For any $D \subset \mathbb{R}^3$, we have

$$P((X,Y,Z) \in D) = \int\int\int_{(x,y,z)\in D} f_{X,Y,Z}(x,y,z)\,\mathrm{dx}\,\mathrm{dy}\,\mathrm{dz}\,.$$

Let $A, B, C \subset \mathbb{R}$, take $D = A \times B \times C$ above

$$P(X \in A, Y \in B, Z \in C) = \int_C \int_B \int_A f_{X,Y,Z}(x,y,z)\,\mathrm{dx}\,\mathrm{dy}\,\mathrm{dz}\,.$$

Marginal probability density function of X, Y and Z

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,\mathrm{dy}\,\mathrm{dz};\\
f_Y(y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,\mathrm{dx}\,\mathrm{dz};\\
f_Z(z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,\mathrm{dx}\,\mathrm{dy};\\
f_{X,Y}(x,y) &= \int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,\mathrm{dz};\\
f_{Y,Z}(y,z) &= \int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,\mathrm{dx};\\
f_{X,Z}(x,z) &= \int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,\mathrm{dy}\,.
\end{aligned}$$

Independent random variables

Proposition 6.41. *For jointly continuous random variables, the following three statements are equivalent:*

(i) Random variables X, Y and Z are independent.

(ii) For all $x, y, z \in \mathbb{R}$, we have

$$f_{X,Y,Z}(x,y,z) = f_X(x) f_Y(y) f_Z(z).$$

(iii) For all $x, y, z \in \mathbb{R}$, we have

$$F_{X,Y,Z}(x,y,z) = F_X(x) F_Y(y) F_Z(z).$$

Checking for independence

Proposition 6.42. *Random variables X, Y and Z are independent if and only if there exist functions $g_1, g_2, g_3 : \mathbb{R} \longrightarrow \mathbb{R}$ such that for all $x, y, z \in \mathbb{R}$, we have*

$$f_{X,Y,Z}(x,y,z) = g_1(x) g_2(y) g_3(z).$$

Conditional distributions

There are a number of conditional probability density functions, to name a few,

$$\begin{aligned} f_{X,Y|Z}(x,y|z) &:= f_{X,Y,Z}(x,y,z)/f_Z(z); \\ f_{X|Y,Z}(x|y,z) &:= f_{X,Y,Z}(x,y,z)/f_{Y,Z}(y,z) \end{aligned}$$

and so on ...

Chapter 7

Properties of Expectation

We start off with some elementary properties of expected values, and will develop into more elaborate and useful properties/calculations.

If $a \leq X \leq b$, then $a \leq E(X) \leq b$.

Proof. We will prove the discrete case. First note that

$$E(X) = \sum xp(x) \geq \sum ap(x) = a.$$

In the same manner,

$$E(X) = \sum xp(x) \leq \sum bp(x) = b.$$

The proof in the continuous case is similar. □

7.1 Expectation of Sums of Random Variables

Proposition 7.1.

(a) If X and Y are jointly discrete with joint probability mass function $p_{X,Y}$, then

$$E[g(X,Y)] = \sum_y \sum_x g(x,y) p_{X,Y}(x,y).$$

(b) If X and Y are jointly continuous with joint probability density function $f_{X,Y}$, then

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy}\,.$$

Example 7.2. An accident occurs at a point X that is uniformly distributed on a road of length L. At the time of the accident an ambulance is at a location Y that is also uniformly distributed on the road. Assuming that X and Y are independent, find the expected distance between the ambulance and the point of the accident.

Solution:
We need to compute $E[|X-Y|]$. Since the joint density function of X and Y is

$$f(x,y) = \frac{1}{L^2}, \quad 0 < x,y < L,$$

we obtain from Proposition 7.1 that

$$E[|X-Y|] = \frac{1}{L^2}\int_0^L \int_0^L |x-y|\,\mathrm{dy}\,\mathrm{dx}\,.$$

Now,

$$\begin{aligned}\int_0^L |x-y|\,\mathrm{dy} &= \int_0^x (x-y)\,\mathrm{dy} + \int_x^L (y-x)\,\mathrm{dy} \\ &= \frac{x^2}{2} + \frac{L^2}{2} - \frac{x^2}{2} - x(L-x) \\ &= \frac{L^2}{2} + x^2 - xL.\end{aligned}$$

Therefore,

$$E[|X-Y|] = \frac{1}{L^2}\int_0^L \left(\frac{L^2}{2} + x^2 - xL\right)\mathrm{dx} = \frac{L}{3}.$$

Some important consequences of Proposition 7.1 are:

(1) If $g(x,y) \geq 0$ whenever $p_{X,Y}(x,y) > 0$, then $E[g(X,Y)] \geq 0$.

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

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

(4) **Monotone Property**
If jointly distributed random variables X and Y satisfy $X \leq Y$. Then,

$$E(X) \leq E(Y).$$

Proof. From (1), consider

$$E(Y) - E(X) = E[Y - X] \geq 0.$$

□

Important special case:
(Mean of sum = sum of means)

$$E(X + Y) = E(X) + E(Y).$$

This of course can be easily extended to

$$E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n). \tag{7.1}$$

Applications of Equation (7.1)

Example 7.3 (The Sample Mean).
Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having distribution function F and expected value μ. Such a sequence of random variables is said to constitute a sample from the distribution F. Define the sample mean $\overline{X}$, as follows:

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Find $E(\bar{X})$.

Solution:

$$\begin{aligned} E(\overline{X}) &= E\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] \\ &= \frac{1}{n}\sum_{k=1}^{n} E(X_k) \\ &= \frac{1}{n}\sum_{k=1}^{n} \mu \\ &= \mu. \end{aligned}$$

That is, the expected value of the sample mean is μ, the mean of the distribution. When the distribution mean μ is unknown, the sample mean is often used in statistics to estimate it.

Example 7.4 (Boole's Inequality).
Let $A_1, \ldots, A_n$ be events in a probability space. Define the indicator variables I_k, $k = 1, \ldots, n$ by

$$I_k = \begin{cases} 1, & \text{if } A_k \text{ occurs} \\ 0, & \text{otherwise} \end{cases}.$$

Let

$$X = \sum_{k=1}^{n} I_k$$

and

$$\begin{aligned} Y &= \begin{cases} 1, & \text{if one of } A_k \text{ occurs} \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} 1, & \text{if } \cup_{k=1}^{n} A_k \text{ occurs} \\ 0, & \text{otherwise} \end{cases} \\ &= I_{\cup_{k=1}^{n} A_k}. \end{aligned}$$

We have

$$E(X) = \sum_{k=1}^{n} E(I_k) = \sum_{k=1}^{n} P(I_k = 1) = \sum_{k=1}^{n} P(A_k),$$

and

$$E(Y) = E\left[I_{\cup_{k=1}^{n} A_k}\right] = P\left(I_{\cup_{k=1}^{n} A_k} = 1\right) = P\left(\cup_{k=1}^{n} A_k\right).$$

Furthermore, it is easy to see that $Y \leq X$, therefore

$$E(Y) \leq E(X).$$

And this is equivalent to

$$P\left(\cup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} P(A_k),$$

commonly known as **Boole's Inequality**.

Example 7.5 (Mean of Binomial).
Let $X \sim$Binomial(n, p) and

$$I_k = \begin{cases} 1, & \text{if the kth trial is a success} \\ 0, & \text{if the kth trial is a failure} \end{cases}.$$

Important observation:

$$X = I_1 + I_2 + \cdots + I_n. \tag{7.2}$$

Then,

$$E(X) = E(I_1) + \cdots + E(I_n) = p + \cdots + p = np.$$

Example 7.6 (Mean of Negative Binomial).
If independent trials, having a constant probability p of being successes, are performed, determine the expected number of trials required to amass a total of r successes.

Solution:
If X denotes the number of trials needed to amass a total of r successes, then X is a negative binomial random variable. It can be represented by

$$X = X_1 + X_2 + \cdots + X_r,$$

where X_1 is the number of trials required to obtain the first success, X_2 the number of additional trials until the second success is obtained, X_3 the number of additional trials until the third success is obtained, and so on. That is, X_i represents the number of additional trials required, after the $(i-1)$st success, until a total of i successes is amassed. A little thought reveals that each of the random variable is a geometric random variable with parameter p. Hence, $E[X_i] = 1/p$, $i = 1, 2, \ldots, r$; and thus

$$E(X) = E(X_1) + \cdots + E(X_r) = \frac{r}{p}.$$

Example 7.7 (Mean of Hypergeometric).
If n balls are randomly selected from an urn containing N balls of which m are white, find the expected number of white balls selected.

Solution:
Let X denote the number of white balls selected, and represent X as

$$X = X_1 + \cdots + X_m$$

where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th white ball is selected} \\ 0, & \text{otherwise} \end{cases}.$$

Now,

$$\begin{aligned} E(X_i) &= P(X_i = 1) \\ &= P(i\text{th white ball is selected}) \\ &= \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} \\ &= \frac{n}{N}. \end{aligned}$$

Hence

$$E(X) = E(X_1) + \cdots + E(X_m) = \frac{nm}{N}.$$

We could also have obtained the above result by using the alternative representation

$$X = Y_1 + \cdots + Y_n$$

where

$$Y_i = \begin{cases} 1, & \text{if the } i\text{th ball selected is white} \\ 0, & \text{otherwise} \end{cases}.$$

Since the ith ball selected is equally likely to be any of the N balls, it follows that

$$E[Y_i] = \frac{m}{N}$$

so

$$E(X) = E(Y_1) + \cdots + E(Y_n) = \frac{nm}{N}.$$

Example 7.8 (Expected Number of Matches).
A group of n people throw their hats into the center of a room. The hats are mixed up, each person randomly selects one. Find the expected number of people who get back their own hats.

Solution:
For $1 \leq k \leq n$, define

$$I_k = \begin{cases} 1, & \text{if the kth person gets back his hat} \\ 0, & \text{otherwise} \end{cases}.$$

Then

$$X := I_1 + I_2 + \cdots + I_n$$

denotes the number of people who get back their hats. Now

$$P(I_k = 1) = \frac{1}{n}.$$

Then

$$\begin{aligned} E(X) &= E[I_1 + I_2 + \cdots + I_n] \\ &= P(I_1 = 1) + P(I_2 = 1) + \cdots + P(I_n = 1) \\ &= \frac{1}{n} + \cdots + \frac{1}{n} \\ &= 1. \end{aligned}$$

Example 7.9 (Coupon-collecting problems).
Suppose that there are N different types of coupons and each time one obtains a coupon it is equally likely to be any one of the N types.

(a) Find the expected number of different types of coupons that are contained in a set of n coupons.

(b) Find the expected number of coupons one need amass before obtaining a complete set of at least one of each type.

Solution:

(a) Let X denote the number of different types of coupons in the set of n coupons. We compute $E[X]$ by using the representation

$$X = X_1 + \cdots + X_N$$

where

$$X_i = \begin{cases} 1, & \text{if at least one type } i \text{ coupon is contained in the set of } n \\ 0, & \text{otherwise} \end{cases}.$$

Now,

$$\begin{aligned} E[X_i] &= P(X_i = 1) \\ &= 1 - P(\text{no type } i \text{ coupons are contained in the set of } n) \\ &= 1 - \left(\frac{N-1}{N}\right)^n. \end{aligned}$$

Hence

$$E[X] = E[X_1] + \cdots + E[X_N] = N\left[1 - \left(\frac{N-1}{N}\right)^n\right].$$

(b) Let Y denote the number of coupons collected before a complete set is attained. We compute $E[Y]$ by using the same technique as we used in computing the mean of a negative binomial random variable. That is, define $Y_i, i = 0, 1, \ldots, N-1$ to be the number of additional coupons that need be obtained after i distinct types have been collected in order to obtain another distinct type, and note that

$$Y = Y_0 + Y_1 + \cdots + Y_{N-1}.$$

When i distinct types of coupons have already been collected, it follows that a new coupon obtained will be of a distinct type with probability $(N-i)/N$. Therefore,

$$P(Y_i = k) = \frac{N-i}{N}\left(\frac{i}{N}\right)^{k-1}, \quad k \geq 1,$$

or in other words, Y_i is a geometric random variable with parameter $(N-i)/N$.

Hence

$$E[Y_i] = \frac{N}{N-i}$$

implying that

$$E[Y] = 1 + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + \frac{N}{1} = N\left[1 + \cdots + \frac{1}{N-1} + \frac{1}{N}\right].$$

7.2 Covariance, Variance of Sums, and Correlations

Definition 7.10. *The **covariance** of jointly distributed random variables X and Y, denoted by $\operatorname{cov}(X,Y)$, is defined by*

$$\operatorname{cov}(X,Y) = E(X-\mu_X)(Y-\mu_Y)$$

where μ_X, μ_Y denote the means of X and Y respectively.

Remark 36. If $\operatorname{cov}(X,Y) \neq 0$, we say that X and Y are **correlated**.
If $\operatorname{cov}(X,Y) = 0$, we say that X and Y are **uncorrelated**.

Warning: Correlation does NOT imply causation.

Example 7.11.

- The Japanese eat little fat and suffer fewer heart attacks than the British or Americans.
- The French eat a lot of fat and also suffer fewer heart attacks than the British or Americans.

- The Italians drink a lot of red wine and also suffer fewer heart attacks than the British or Americans.

Conclusions:

- Eat and drink what you like.
- Speaking English is apparently what kills you.

Irwin Knopp
Reader's Digest (Feb. 2003)

(An alternative formula of covariance:)

$$\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y).$$

Proof. Let μ_X, μ_Y denote respectively $E(X)$ and $E(Y)$.

$$\begin{aligned} \operatorname{cov}(X,Y) &= E[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] \\ &= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X\mu_Y \\ &= E(XY) - E(X)E(Y). \end{aligned}$$

□

We start with a useful result for independent random variables.

Proposition 7.12. *If X and Y are independent, then for any functions $g, h : \mathbb{R} \to \mathbb{R}$, we have*

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

Proof. Ross [p.322] proves this in the continuous case. Let's prove the discrete case.

$$\begin{aligned} E[g(X)h(Y)] &= \sum_{x,\,y} g(x)h(y)p_{X,Y}(x,y) \\ &= \sum_x \sum_y g(x)h(y)p_X(x)p_Y(y) \\ &= \left[\sum_x g(x)p_X(x)\right]\left[\sum_y h(y)p_Y(y)\right] \\ &= E[g(X)]E[h(Y)]. \end{aligned}$$

□

Corollary 7.13. *If X and Y are independent, then* $\operatorname{cov}(X,Y) = 0$.

Proof.

$$\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0.$$

□

However, the converse is not true.

Example 7.14. A simple example of two dependent random variables X and Y having zero covariance is obtained by letting X be a random variable such that

$$P(X=0) = P(X=1) = P(X=-1) = \frac{1}{3}$$

and define

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

Now, $XY = 0$, so $E[XY] = 0$. Also, $E[X] = 0$ and thus

$$\operatorname{cov}(X,Y) = E[XY] - E[X]E[Y] = 0.$$

However, X and Y are clearly not independent.

(Some properties of covariance)

(i) $\operatorname{var}(X) = \operatorname{cov}(X,X)$.

(ii) $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$.

(iii) $\operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \operatorname{cov}(X_i, Y_j)$.

Note: these applies for general situation – No independence is assumed.

Proof.

(i) This part is obvious.

(ii) This is also obvious.

(iii) We first prove

$$\operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, Z\right) = \sum_{i=1}^{n} a_i \operatorname{cov}(X_i, Z). \tag{7.3}$$

Note first that

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mu_{X_i}.$$

Then

$$\begin{aligned}
\operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, Z\right) &= E\left[\left(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} a_i \mu_{X_i}\right)(Z-\mu_Z)\right] \\
&= E\left[\sum_{i=1}^{n} a_i (X_i - \mu_{X_i})(Z-\mu_Z)\right] \\
&= \sum_{i=1}^{n} a_i E[(X_i - \mu_{X_i})(Z-\mu_Z)] \\
&= \sum_{i=1}^{n} a_i \operatorname{cov}(X_i, Z).
\end{aligned}$$

The general case follows from

$$\begin{aligned}
\operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) &= \sum_{i=1}^{n} a_i \operatorname{cov}\left(X_i, \sum_{j=1}^{m} b_j Y_j\right) \\
&= \sum_{i=1}^{n} a_i \sum_{j=1}^{m} b_j \operatorname{cov}\left(X_i, Y_j\right).
\end{aligned}$$

□

(Variance of a Sum)

$$\operatorname{var}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \operatorname{var}(X_k) + 2 \sum_{1 \le i < j \le n} \operatorname{cov}(X_i, X_j). \tag{7.4}$$

Proof.

$$\begin{aligned}
\operatorname{var}\left(\sum_{k=1}^{n} X_k\right) &= \operatorname{cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{cov}(X_i, X_j) \\
&= \sum_{i=1} \operatorname{cov}(X_i, X_i) + \sum_{1 \le i \ne j \le n} \operatorname{cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{1 \le i < j \le n} \operatorname{cov}(X_i, X_j).
\end{aligned}$$

□

(Variance of a Sum under Independence)
Let $X_1, \ldots, X_n$ be **independent** random variables, then

$$\operatorname{var}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \operatorname{var}(X_k).$$

In other words, **under independence, variance of sum = sum of variances.**

Example 7.15 (Sample Variance).
Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having expected value μ and variance σ^2, and let $\overline{X} = \sum_{i=1}^{n} X_i/n$ be the sample mean. The quantities $X_i - \overline{X}$, $i = 1, \ldots, n$, are called **deviations**, as they equal the differences between the individual data and the sample mean. The random variable

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$$

is called the **sample variance**. Find (a) $\operatorname{var}(\overline{X})$ and (b) $E[S^2]$.

Solution:
(a)

$$\begin{aligned}
\operatorname{var}(\overline{X}) &= \left(\frac{1}{n}\right)^2 \operatorname{var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \operatorname{var}(X_i) \quad \text{by independence} \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

(b) We start with the following algebraic identity

$$\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n} (X_i - \mu + \mu - \overline{X})^2 \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 + \sum_{i=1}^{n} (\overline{X} - \mu)^2 - 2(\overline{X} - \mu) \sum_{i=1}^{n} (X_i - \mu) \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 + n(\overline{X} - \mu)^2 - 2(\overline{X} - \mu) n (\overline{X} - \mu) \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X} - \mu)^2.
\end{aligned}$$

Taking expectations of the above yields that

$$\begin{aligned}(n-1)E[S^2] &= \sum_{i=1}^{n} E[(X_i-\mu)^2] - nE[(\overline{X}-\mu)^2] \\ &= n\sigma^2 - n\,\mathrm{var}(\overline{X}) \\ &= (n-1)\sigma^2\end{aligned}$$

where the final equality is due to part (a) and preceding one uses the the result that $E[\overline{X}] = \mu$. Dividing through by $n-1$ shows that

$$E[S^2] = \sigma^2.$$

Remark 37. This explains why

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$$

is used as a estimator of σ^2 instead of the more "natural" choice of

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n}.$$

Example 7.16 (Variance of Binomial).
Recall the representation of $X \sim \mathrm{Bin}(n,p)$ given by Equation (7.2). We notice further that the I_k's are independent. Then from Equation (7.4),

$$\mathrm{var}(X) = \mathrm{var}\left(\sum_{k=1}^{n} I_k\right) = \sum_{k=1}^{n} \mathrm{var}(I_k) = \sum_{k=1}^{n} p(1-p) = np(1-p).$$

Example 7.17 (Variance of Number of Matches).
Recall the number of matches, X, is given as

$$X = \sum_{k=1}^{n} I_k.$$

Hence,

$$\begin{aligned}\mathrm{var}(X) &= \sum_{i=1}^{n} \mathrm{var}(I_i) + 2 \sum_{1\le i<j\le n} \mathrm{cov}(I_i, I_j) \\ &= \sum_{i=1}^{n} \frac{1}{n}\left(1-\frac{1}{n}\right) + 2 \sum_{1\le i<j\le n} \mathrm{cov}(I_i, I_j) \\ &= n \cdot \frac{1}{n}(1-1/n) + n(n-1)\,\mathrm{cov}(I_1, I_2),\end{aligned}$$

where

$$\begin{aligned}\operatorname{cov}(I_1,I_2) &= E(I_1I_2)-E(I_1)E(I_2)\\ &= P(I_1=I_2=1)-1/n^2\\ &= P(I_1=1)P(I_2=1|I_1=1)-1/n^2\\ &= 1/n\times 1/(n-1)-1/n^2\\ &= \frac{1}{n^2(n-1)}.\end{aligned}$$

Therefore

$$\operatorname{var}(X)=(1-1/n)+1/n=1.$$

Correlation/Correlation coefficient

Definition 7.18. *The **correlation (coefficient)** of random variables X and Y, denoted by $\rho(X,Y)$, is defined by*

$$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}.$$

Proposition 7.19. *We have,*

$$-1\le\rho(X,Y)\le 1.$$

Proof. Suppose that X and Y have variances given by σ_X^2 and σ_Y^2, respectively. Then

$$\begin{aligned}0 &\le \operatorname{var}\left(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\right)\\ &= \frac{\operatorname{var}(X)}{\sigma_X^2}+\frac{\operatorname{var}(Y)}{\sigma_Y^2}+\frac{2\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}\\ &= 2[1+\rho(X,Y)]\end{aligned}$$

implying that

$$-1\le\rho(X,Y).$$

On the other hand,

$$\begin{aligned}0 &\le \operatorname{var}\left(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\right)\\ &= \frac{\operatorname{var}(X)}{\sigma_X^2}+\frac{\operatorname{var}(Y)}{(-\sigma_Y)^2}-\frac{2\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}\\ &= 2[1-\rho(X,Y)]\end{aligned}$$

implying that

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

which completes the proof. □

Remark 38. (1) The correlation coefficient is a measure of the degree of linearity between X and Y. A value of $\rho(X,Y)$ near $+1$ or -1 indicates a high degree of linearity between X and Y, whereas a value near 0 indicates a lack of such linearity. A positive value of $\rho(X,Y)$ indicates that Y tends to increase when X does, whereas a negative value indicates that Y tends to decrease when X increases. If $\rho(X,Y) = 0$, then X and Y are said to be **uncorrelated**.

(2) $\rho(X,Y) = 1$ if and only if $Y = aX + b$ where $a = \sigma_Y/\sigma_X > 0$.

(3) $\rho(X,Y) = -1$ if and only if $Y = aX + b$ where $a = -\sigma_Y/\sigma_X < 0$.

(4) $\rho(X,Y)$ is dimensionless.

(5) Similar to covariance, if X and Y are independent, then $\rho(X,Y) = 0$.

Note that the converse is not true. In other words: If $\rho(X,Y) = 0$, meaning $\text{cov}(X,Y) = 0$, then X and Y **may not be independent.**

Example 7.20. Let I_A and I_B be indicator variables for the events A and B. That is,

$$I_A = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{otherwise} \end{cases},$$

$$I_B = \begin{cases} 1, & \text{if } B \text{ occurs} \\ 0, & \text{otherwise} \end{cases}.$$

Then

$$E[I_A] = P(A),$$
$$E[I_B] = P(B),$$
$$E[I_A I_B] = P(AB).$$

Therefore

$$\text{cov}(I_A, I_B) = P(AB) - P(A)P(B) = P(B)[P(A|B) - P(A)].$$

Thus we obtain the quite intuitive result that the indicator variables for A and B are either positively correlated, uncorrelated, or negatively correlated depending on whether $P(A|B)$ is, respectively, greater than, equal to, or less than $P(A)$.

Our next example shows that the sample mean and a deviation from the sample mean are uncorrelated.

Example 7.21. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having variance σ^2. Show that

$$\operatorname{cov}(X_i - \overline{X}, \overline{X}) = 0.$$

Solution:
We have

$$\begin{aligned}
\operatorname{cov}(X_i - \overline{X}, \overline{X}) &= \operatorname{cov}(X_i, \overline{X}) - \operatorname{cov}(\overline{X}, \overline{X}) \\
&= \operatorname{cov}\left(X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\right) - \operatorname{var}(\overline{X}) \\
&= \frac{1}{n}\sum_{j=1}^{n} \operatorname{cov}(X_i, X_j) - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0
\end{aligned}$$

where the next-to-last equality uses the result that $\operatorname{var}(\overline{X}) = \frac{\sigma^2}{n}$, and the final equality follows since

$$\operatorname{cov}(X_i, X_j) = \begin{cases} 0, & \text{if } j \neq i \text{ by independence} \\ \sigma^2, & \text{if } j = i \text{ since } \operatorname{var}(X_i) = \sigma^2 \end{cases}.$$

7.3 Conditional Expectation

Definition 7.22.

(1) If X and Y are jointly distributed discrete random variables, then

$$E[X|Y = y] := \sum_x x p_{X|Y}(x|y), \qquad \text{if } p_Y(y) > 0.$$

(2) If X and Y are jointly distributed continuous random variables, then

$$E[X|Y = y] := \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\ \mathrm{dx}, \qquad \text{if } f_Y(y) > 0.$$

Example 7.23. If X and Y are independent binomial random variables with identical parameters n and p, calculate the conditional expected value of X given that $X+Y=m$.

Solution:
Let us first calculate the conditional probability mass function of X, given that $X+Y=m$. For $k \leq \min(n,m)$,

$$\begin{aligned} P(X=k|X+Y=m) &= \frac{P(X=k, X+Y=m)}{P(X+Y=m)} \\ &= \frac{P(X=k, Y=m-k)}{P(X+Y=m)} \\ &= \frac{P(X=k)P(Y=m-k)}{P(X+Y=m)} \\ &= \frac{\binom{n}{k}p^k(1-p)^{n-k}\binom{n}{m-k}p^{m-k}(1-p)^{n-m+k}}{\binom{2n}{m}p^m(1-p)^{2n-m}} \\ &= \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}} \end{aligned}$$

where we have used the fact that $X+Y$ is a binomial random variable with parameters $(2n,p)$. Hence the conditional distribution of X, given that $X+Y=m$, is the hypergeometric distribution; thus, we obtain

$$E[X|X+Y=m] = \frac{m}{2}.$$

Example 7.24. Suppose that X and Y have the joint probability density function

$$f(x,y) = \begin{cases} \frac{e^{-x/y}e^{-y}}{y}, & 0<x<\infty,\ 0<y<\infty \\ 0, & \text{otherwise} \end{cases}.$$

Compute $E[X|Y=y]$.

Solution:
It can be shown that

$$f_{X|Y}(x|y) = \frac{1}{y}e^{-x/y}.$$

Hence,

$$E[X|Y=y] = \int_0^\infty x f_{X|Y}(x|y)\,\mathrm{dx} = \int_0^\infty \frac{x}{y}e^{-x/y}\,\mathrm{dx} = y.$$

(Some important formulas)

$$E[g(X)|Y=y] = \begin{cases} \sum_x g(x) p_{X|Y}(x|y), & \text{for discrete case} \\ \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\,\mathrm{dx}, & \text{for continuous case} \end{cases},$$

and hence

$$E\left[\sum_{k=1}^{n} X_k | Y=y\right] = \sum_{k=1}^{n} E[X_k|Y=y].$$

7.3.1 Computing expectation by conditioning

Notation: Let us denote by $E[X|Y]$ that function of the random variable Y whose value at $Y=y$ is $E[X|Y=y]$.

Example 7.25.

(1) Suppose that $E[X|Y=y] = y/2-5$, then $E[X|Y] = Y/2-5$.

(2) Suppose that $E[X|Y=y] = e^{y/2}-5y$, then $E[X|Y] = e^{Y/2}-5Y$.

Proposition 7.26.

$$E[X] = E[E(X|Y)] = \begin{cases} \sum_y E(X|Y=y)P(Y=y), & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} E(X|Y=y) f_Y(y)\,\mathrm{dy}, & \text{if } Y \text{ is continuous} \end{cases}.$$

Proof. We prove the result when X and Y are both continuous. For the case where X and Y are both discrete, see Ross p.333.
Write $g(y) = E(X|Y=y)$, then

$$\begin{aligned} E[E(X|Y)] &= E[g(Y)] = \int_{-\infty}^{\infty} g(y) f_Y(y)\ \mathrm{dy} \\ &= \int_{-\infty}^{\infty} E(X|Y=y) f_Y(y)\ \mathrm{dy} \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\ \mathrm{dx} \right) f_Y(y)\ \mathrm{dy} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y)\ \mathrm{dx}\,\mathrm{dy} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y)\ \mathrm{dx}\,\mathrm{dy} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y)\ \mathrm{dy}\,\mathrm{dx} \\ &= \int_{-\infty}^{\infty} x f_X(x)\ \mathrm{dx} = E[X]. \end{aligned}$$

□

Example 7.27. A miner is trapped in a mine containing 3 doors. The first door leads to a tunnel that will take him to safety after 3 hours of travel. The second door leads to a tunnel that will return him to the mine after 5 hours of travel. The third door leads to a tunnel that will return him to the mine after 7 hours of travel. If we assume that the miner, on each choice (he "forgets his previous choices"), is equally likely to choose any one of the three doors, what is his expected length of time until he reaches for safety?

Solution:
Let X denote amount of time until he reaches for safety; Y denote the door he initially chooses.
Now

$$\begin{aligned} E[X] &= E[X|Y=1]P(Y=1)+E[X|Y=2]P(Y=2)+E[X|Y=3]P(Y=3) \\ &= \frac{1}{3}\left(E[X|Y=1]+E[X|Y=2]+E[X|Y=3]\right). \end{aligned}$$

Now,

$$\begin{aligned} E[X|Y=1] &= 3, \\ E[X|Y=2] &= 5+E[X], \\ E[X|Y=3] &= 7+E[X], \end{aligned}$$

hence

$$E[X] = \frac{1}{3}\left(3+5+E[X]+7+E[X]\right).$$

Solving this equation, we obtain

$$E[X] = 15.$$

Example 7.28 (Expectation of a Random Sum).
Suppose $X_1, X_2, \ldots$ are independent and identically distributed with common mean μ. Suppose that N is a nonnegative integer valued random variable, independent of the X_k's. We are interested to find the mean of

$$T = \sum_{k=1}^{N} X_k,$$

where T is taken to be zero when $N = 0$.
Interpretations of the problem:

(a) N denotes the number of customers entering a department store during a period of time; X_k's amount spent by the kth customer; T total revenue.

(b) N denotes the number of claims against an insurance company; X_k's size of the kth claim; T total pay off.

Solution:

$$\begin{aligned} E[T] &= E\left[\sum_{k=1}^{N} X_k\right] \\ &= \sum_{n=0}^{\infty} E\left[\sum_{k=1}^{N} X_k | N=n\right] P(N=n) \\ &= \sum_{n=1}^{\infty} E\left[\sum_{k=1}^{n} X_k | N=n\right] P(N=n) \\ &= \sum_{n=1}^{\infty} \left[\sum_{k=1}^{n} E(X_k)\right] P(N=n) \\ &= \sum_{n=1}^{\infty} n\mu P(N=n) \\ &= \mu E[N]. \end{aligned}$$

Example 7.29. Suppose that the number of people entering a department store on a given day is a random variable with mean 50. Suppose further that the amounts of money spent by those customers are independent random variables having a common mean of 8. Finally, suppose also that the amount of money spent by a customer is also independent of the total number of customers who enter the store. What is the expected amount of money spent in the store on a given day?

Solution:
Using the previous example, the expected amount of money spent in the store is $50 \times \$8 = \$400$.

Example 7.30 (Mean and variance of Geometric).
Let N be a geometrically distributed random variable with probability of success p. Let X_1 be the outcome of first trial.

$$\begin{aligned} E[N] &= E[N|X_1=0]P(X_1=0) + E[N|X_1=1]P(X_1=1) \\ &= (1+E[N])q + 1 \times p. \end{aligned}$$

Solving for $E[N]$, we have

$$E[N] = \frac{1}{p}.$$

Similarly,

$$\begin{aligned} E[N^2] &= E[N^2|X_1 = 0]P(X_1 = 0) + E[N^2|X_1 = 1]P(X_1 = 1) \\ &= E[(1+N)^2]q + 1 \times p \\ &= 1 + 2qE[N] + qE[N^2] \\ &= 1 + \frac{2q}{p} + qE[N^2]. \end{aligned}$$

Solving for $E[N^2]$, we get

$$E[N^2] = \frac{1}{p} + \frac{2q}{p^2} = \frac{1+q}{p^2}.$$

And hence,

$$\text{var}(N) = E[N^2] - (EN)^2 = \frac{q}{p^2}.$$

7.3.2 Computing probabilities by conditioning

Not only can we obtain expectations by first conditioning on an appropriate random variable, but we may also use this approach to compute probabilities. To see this, let $X = I_A$ where A is an event. Then, we have

$$E(I_A) = P(A), \quad \text{and} \quad E(I_A|Y = y) = P(A|Y = y),$$

and by Proposition 7.26,

$$\begin{aligned} P(A) = E(I_A) &= E[E(I_A|Y)] \\ &= \begin{cases} \sum_y E(I_A|Y = y)P(Y = y), & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} E(I_A|Y = y)f_Y(y)\,\mathrm{d}y, & \text{if } Y \text{ is continuous} \end{cases} \\ &= \begin{cases} \sum_y P(A|Y = y)P(Y = y), & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} P(A|Y = y)f_Y(y)\,\mathrm{d}y, & \text{if } Y \text{ is continuous} \end{cases}. \end{aligned}$$

This can be compared with the Bayes first formula.

Example 7.31. Suppose that X and Y are independent continuous random variables having probability density functions f_X and f_Y respectively. Compute $P(X < Y)$.

Solution:
Conditioning on the value of Y yields

$$\begin{aligned} P(X<Y) &= \int_{-\infty}^{\infty} P(X<Y|Y=y) f_Y(y)\,\mathrm{dy} \\ &= \int_{-\infty}^{\infty} P(X<y|Y=y) f_Y(y)\,\mathrm{dy} \\ &= \int_{-\infty}^{\infty} P(X<y) f_Y(y)\,\mathrm{dy} \quad \text{by independence} \\ &= \int_{-\infty}^{\infty} F_X(y) f_Y(y)\,\mathrm{dy}\,. \end{aligned}$$

Example 7.32. Consider Example 6.5 on page 121. Note that $X \sim \mathrm{Exp}(1)$ and $Y \sim \mathrm{Exp}(2)$ and X, Y are independent. We can also compute $P(X<Y)$ using conditioning:

$$\begin{aligned} P(X<Y) &= \int_{-\infty}^{\infty} P(X<y) f_Y(y)\,\mathrm{dy} \\ &= \int_0^{\infty} (1-e^{-y}) 2e^{-2y}\,\mathrm{dy} \\ &= \int_0^{\infty} 2e^{-2y}\,\mathrm{dy} - \int_0^{\infty} 2e^{-3y}\,\mathrm{dy} \\ &= \int_0^{\infty} 2e^{-2y}\,\mathrm{dy} - \frac{2}{3}\int_0^{\infty} 3e^{-3y}\,\mathrm{dy} \\ &= 1 - 2/3 = 1/3. \end{aligned}$$

Example 7.33. Suppose that X and Y are independent continuous random variables having probability density functions f_X and f_Y respectively. Find the distribution function of $X+Y$.

Solution:
By conditioning on the value of Y yields

$$\begin{aligned} P(X+Y\le a) &= \int_{-\infty}^{\infty} P(X+Y\le a|Y=y) f_Y(y)\,\mathrm{dy} \\ &= \int_{-\infty}^{\infty} P(X+y\le a|Y=y) f_Y(y)\,\mathrm{dy} \\ &= \int_{-\infty}^{\infty} P(X\le a-y) f_Y(y)\,\mathrm{dy} \quad \text{by independence} \\ &= \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\,\mathrm{dy}\,. \end{aligned}$$

Example 7.34. Let U be a uniform random variable on (0,1), and suppose that the conditional density function of X, given that $U=p$, is binomial with parameters n and p. Find the probability mass function of X.

Solution:
First note that X takes values $0,1,\ldots,n$. For each $0 \le k \le n$, by conditioning on the value of U,

$$\begin{aligned} P(X=k) &= \int_0^1 P(X=k|U=p) f_U(p) \mathrm{dp} \\ &= \int_0^1 P(X=k|U=p) \mathrm{dp} \\ &= \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \mathrm{dp} \\ &= \binom{n}{k} \frac{k!(n-k)!}{(n+1)!} \quad \text{(claim below).} \end{aligned}$$

Claim.

$$B(k,n-k) := \int_0^1 p^k (1-p)^{n-k} \mathrm{dp} = \frac{k!(n-k)!}{(n+1)!}.$$

Proof of claim:

$$\begin{aligned} B(k,n-k) &= \int_0^1 (1-p)^{n-k} \mathrm{d}[p^{k+1}/(k+1)] \\ &= \frac{(n-k)}{(k+1)} \int_0^1 p^{k+1} (1-p)^{n-k-1} \mathrm{dp} \\ &= \frac{(n-k)}{(k+1)} B(k+1,n-k-1) \\ &= \frac{(n-k)(n-k-1)}{(k+1)(k+2)} B(k+2,n-k-2) \\ &\vdots \quad \vdots \\ &= \frac{(n-k)!}{(k+1)(k+2)\cdots n} B(n,0) \\ &= \frac{k!(n-k)!}{(n+1)!} \end{aligned}$$

as $B(n,0) = \int_0^1 p^n dp = 1/(n+1)$.

So we obtain

$$P(X=k) = \frac{1}{n+1}, \quad k = 0,1,\ldots,n.$$

That is, if a coin whose probability of coming up head is uniformly distributed over $(0,1)$ is flipped n times, then the number of heads occurring is equally likely to be any of the values $0,1,\ldots,n$.

7.4 Conditional Variance

Definition 7.35. *The **conditional variance** of X given that $Y = y$ is defined as*

$$\text{var}(X|Y) \equiv E[(X - E[X|Y])^2|Y]$$

A very useful relationship between $\text{var}(X)$ and $\text{var}(X|Y)$ is given by

Proposition 7.36.

$$\text{var}(X) = E[\text{var}(X|Y)] + \text{var}(E[X|Y])$$

Proof. Note that

$$\text{var}(X|Y) = E[X^2|Y] - (E[X|Y])^2.$$

Taking expectation on both sides yields

$$E[\text{var}(X|Y)] = E[E[X^2|Y]] - E[(E[X|Y])^2] = E[X^2] - E[(E[X|Y])^2].$$

Since $E[E[X|Y]] = E[X]$, we have

$$\text{var}(E[X|Y]) = E[(E[X|Y])^2] - (E[X])^2.$$

Summing the two expressions above yields the required result. □

Example 7.37. Suppose that by any time t the number of people that have arrived at a train depot is a Poisson random variable with mean λt. If the initial train arrives at the depot at a time (independent of when the passengers arrive) that is uniformly distributed over $(0,T)$, what are the mean and variance of the number of passengers who enter the train?

Solution:
For $t \geq 0$, let $N(t)$ denote the number of arrivals during the interval $(0,t)$, and Y the time at which the train arrives.

We are interested to compute the mean and variance of the random variable $N(Y)$.

We condition on Y to get

$$\begin{aligned} E[N(Y)|Y = t] &= E[N(t)|Y = t] \\ &= E[N(t)] \quad \text{by the independence of } Y \text{ and } N(t) \\ &= \lambda t \end{aligned}$$

since $N(t)$ is Poisson(λt).

Hence,

$$E[N(Y)|Y] = \lambda Y.$$

Taking expectations gives

$$E[N(Y)] = E[E[N(Y)|Y]] = \lambda E[Y] = \frac{\lambda T}{2}$$

since $Y \sim U(0,T)$.

To obtain $\text{var}(N(Y))$, note that

$$\begin{aligned}\text{var}(N(Y)|Y=t) &= \text{var}(N(t)|Y=t)\\ &= \text{var}(N(t)) \quad \text{by independence}\\ &= \lambda t\end{aligned}$$

So

$$\text{var}(N(Y)|Y) = \lambda Y.$$

Using the conditional variance formula, and the fact that $E[N(Y)|Y] = \lambda Y$ and $\text{var}(N(Y)|Y) = \lambda Y$, we obtain

$$\begin{aligned}\text{var}(N(Y)) &= E[\text{var}(N(Y)|Y)] + \text{var}(E[N(Y)|Y])\\ &= E[\lambda Y] + \text{var}(\lambda Y)\\ &= \lambda \frac{T}{2} + \lambda^2 \frac{T^2}{12}\end{aligned}$$

where we have used the fact that $\text{var}(Y) = T^2/12$.

7.5 Moment Generating Functions

Definition 7.38. *The **moment generating function** of random variable X, denoted by M_X, is defined as*

$$\begin{aligned}M_X(t) &= E\left[e^{tX}\right]\\ &= \begin{cases}\sum_x e^{tx} p_X(x), & \text{if } X \text{ is discrete with probability mass function } p_X\\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\,\mathrm{dx}, & \text{if } X \text{ is continuous with probability density function } f_X\end{cases}.\end{aligned}$$

Why do we call such a function a moment generating function?

Because this function generates all the moments of this random variable X. Indeed, for $n \geq 0$,

$$E(X^n) = M_X^{(n)}(0)$$

where

$$M_X^{(n)}(0) := \frac{d^n}{\mathrm{dt}^n} M_X(t)\,|_{t=0}\,.$$

Proof. By Taylor series expansion, we have

$$\begin{aligned} E\left[e^{tX}\right] &= E\left[\sum_{k=0}^{\infty}\frac{(tX)^k}{k!}\right] \\ &= \sum_{k=0}^{\infty}\frac{E(tX)^k}{k!} \\ &= \sum_{k=0}^{\infty}\frac{E(X^k)}{k!}t^k \end{aligned}$$

and

$$M_X(t) = \sum_{k=0}^{\infty}\frac{M_X^{(k)}(0)}{k!}t^k.$$

Equating coefficient of t^n, we get

$$M_X^{(n)}(0) = E(X^n).$$

□

Proposition 7.39 (Multiplicative Property)**.**
If X and Y are independent, then

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

Proof.

$$\begin{aligned} E\left[e^{t(X+Y)}\right] &= E\left[e^{tX}e^{tY}\right] \\ &= E\left[e^{tX}\right]E\left[e^{tY}\right] \\ &= M_X(t)M_Y(t). \end{aligned}$$

□

Proposition 7.40 (Uniqueness Property)**.**
Let X and Y be random variables with their moment generating functions M_X and M_Y respectively. Suppose that there exists an $h > 0$ such that

$$M_X(t) = M_Y(t), \quad \forall t \in (-h,h),$$

then X and Y have the same distribution (i.e., $F_X = F_Y$; or $f_X = f_Y$.)

Example 7.41.

1. When $X \sim \text{Be}(p)$, $M(t) = 1 - p + pe^t$.

   Proof.

   $$E\left[e^{tX}\right] = e^{t\cdot 0}P(X=0) + e^{t\cdot 1}P(X=1) = (1-p) + pe^t.$$

   □

2. When $X \sim \text{Bin}(n,p)$, $M(t) = (1-p+pe^t)^n$.

   Proof. Use the representation Equation (7.2) of the Binomial,

   $$\begin{aligned} E\left[e^{tX}\right] &= E\left[e^{t(I_1+\cdots+I_n)}\right] \\ &= E\left[e^{tI_1}\cdots e^{tI_n}\right] \\ &= E\left[e^{tI_1}\right]\cdots E\left[e^{tI_n}\right] \\ &= (1-p+pe^t)^n. \end{aligned}$$

   □

3. When $X \sim \text{Geom}(p)$, $M(t) = \frac{pe^t}{1-(1-p)e^t}$.

4. When $X \sim \text{Poisson}(\lambda)$, $M(t) = \exp(\lambda(e^t - 1))$.

   Proof.

   $$\begin{aligned} E\left[e^{tX}\right] &= \sum_{k=0}^{\infty} \frac{e^{tk}e^{-\lambda}\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} \\ &= e^{-\lambda}e^{\lambda e^t} = \exp\left(\lambda\left(e^t-1\right)\right). \end{aligned}$$

   □

5. When $X \sim U(\alpha,\beta)$, $M(t) = \frac{e^{\beta t}-e^{\alpha t}}{(\beta-\alpha)t}$.

6. When $X \sim \text{Exp}(\lambda)$, $M(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$.

Proof.

$$\begin{aligned} E\left[e^{tX}\right] &= \int_0^\infty e^{tx}\lambda e^{-\lambda x}\,\mathrm{dx} \\ &= \lambda\int_0^\infty e^{-(\lambda-t)x}\,\mathrm{dx} \\ &= \frac{\lambda}{\lambda-t}, \quad \text{for } t<\lambda. \end{aligned}$$

Note that $M(t)$ is only defined for $t<\lambda$. □

7. When $X\sim N(\mu,\sigma^2)$, $M(t)=\exp(\mu t+\sigma^2t^2/2)$.

Proof. Suppose that Y is standard normal.

$$\begin{aligned} E\left[e^{tY}\right] &= \int_{-\infty}^\infty e^{ty}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\,\mathrm{dy} \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-[y^2-2ty]/2}\,\mathrm{dy} \\ &= e^{t^2/2}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-(y-t)^2/2}\,\mathrm{dy} \\ &= e^{t^2/2}. \end{aligned}$$

For $X\sim N(\mu,\sigma^2)$, we write $X=\sigma Y+\mu$ where $Y\sim N(0,1)$.

$$\begin{aligned} E\left[e^{tX}\right] &= E\left[e^{t\mu+t\sigma Y}\right] \\ &= e^{\mu t}E\left[e^{(\sigma t)Y}\right] \\ &= e^{\mu t+\sigma^2t^2/2}. \end{aligned}$$

□

Example 7.42. Suppose that $M_X(t)=e^{3(e^t-1)}$. Find $P(X=0)$.

Solution:
The given moment generating function is from that of a Poisson and the parameter of the Poisson is λ. Hence, $X\sim\text{Poisson}(\lambda)$. Therefore,

$$P(X=0)=e^{-3}.$$

Example 7.43. Sum of independent Binomial random variables with the same success probability is Binomial.

Solution:
Let $X \sim \text{Bin}(n,p)$; $Y \sim \text{Bin}(m,p)$; and X and Y are independent.

$$\begin{aligned} E\left[e^{t(X+Y)}\right] &= E\left[e^{tX}\right] E\left[e^{tY}\right] \\ &= \left[1-p+pe^t\right]^n \left[1-p+pe^t\right]^m \\ &= \left[1-p+pe^t\right]^{n+m} \end{aligned}$$

which is the moment generating function of a $\text{Bin}(n+m,p)$. Hence, by uniqueness theorem,

$$X+Y \sim \text{Bin}(n+m,p).$$

Example 7.44. Sum of independent Poisson random variables is Poisson.

Solution:
Proceed as in the binomial case.

Example 7.45. Sum of independent normal random variables is normal.

Solution:
Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. We assume that X and Y are independent.

$$\begin{aligned} E\left[e^{t(X+Y)}\right] &= E\left[e^{tX}\right] E\left[e^{tY}\right] \\ &= \exp(\mu_1 t + \sigma_1^2 t^2/2)\ \exp(\mu_2 t + \sigma_2^2 t^2/2) \\ &= \exp([\mu_1+\mu_2]t + [\sigma_1^2+\sigma_2^2]t^2/2), \end{aligned}$$

which is the moment generating function of a normal distribution with mean $\mu_1+\mu_2$ and $\sigma_1^2+\sigma_2^2$.

Example 7.46. A chi-squared random variable with n degrees of freedom, χ_n^2, is given as

$$Z_1^2 + \cdots + Z_n^2$$

where $Z_1, \ldots, Z_n$ are independent standard normal random variables. Compute the moment generating function of a chi-squared random variable with n degrees of freedom.

Solution:
Let $M(t)$ be its moment generating function. By the above form,

$$M(t) = \left(E\left[e^{tZ^2}\right]\right)^n$$

where Z is a standard normal.
Now,

$$\begin{aligned} E\left[e^{tZ^2}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx^2}e^{-x^2/2}\,\mathrm{dx} \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2\sigma^2}\,\mathrm{dx}, \quad \text{where } \sigma^2 = (1-2t)^{-1}. \\ &= \sigma \\ &= (1-2t)^{-1/2}, \end{aligned}$$

where the next-to-last equality uses the fact that the normal density with mean 0 and variance σ^2 integrates to 1.
Therefore,

$$M(t) = (1-2t)^{-n/2}.$$

Example 7.47. Find all the moments of the exponential distribution of parameter $\lambda > 0$.

Solution:
Let $X \sim \mathrm{Exp}(\lambda)$. Recall that

$$\begin{aligned} E\left[e^{tX}\right] = \frac{\lambda}{\lambda - t} &= \frac{1}{1-t/\lambda} \\ &= \sum_{k=0}^{\infty} \frac{t^k}{\lambda^k} \\ &= \sum_{k=0}^{\infty} \frac{k!/\lambda^k}{k!}\, t^k. \end{aligned}$$

Hence

$$E\left[X^k\right] = \frac{k!}{\lambda^k}.$$

7.6 Joint Moment Generating Functions

It is also possible to define the joint moment generating function of two or more random variables. This is done as follows. For any n random variables $X_1, \ldots, X_n$, the joint moment generating function, $M(t_1, \ldots, t_n)$, is defined for all real values of $t_1, \ldots, t_n$ by

$$M(t_1, \ldots, t_n) = E\left[e^{t_1X_1 + \cdots + t_nX_n}\right].$$

The individual moment generating functions can be obtained from $M(t_1,\ldots,t_n)$ by letting all but one of the t_j's be 0. That is,

$$M_{X_i}(t) = E\left[e^{tX_i}\right] = M(0,\ldots,0,t,0,\ldots,0),$$

where the t is in the ith place.
It can be proved (although the proof is too advanced for this course) that $M(t_1,\ldots,t_n)$ uniquely determines the joint distribution of $X_1, \ldots, X_n$. This result can then be used to prove that the n random variables $X_1, \ldots, X_n$ are independent if and only if

$$M(t_1,\ldots,t_n) = M_{X_1}(t_1)\cdots M_{X_n}(t_n). \tag{7.5}$$

This follows because, if the n random variables are independent, then

$$\begin{aligned} M(t_1,\ldots,t_n) &= E\left[e^{t_1X_1+\cdots+t_nX_n}\right] \\ &= E\left[e^{t_1X_1}\cdots e^{t_nX_n}\right] \\ &= E\left[e^{t_1X_1}\right]\cdots E\left[e^{t_nX_n}\right] \quad \text{by independence} \\ &= M_{X_1}(t_1)\cdots M_{X_n}(t_n). \end{aligned}$$

On the other hand, if Equation (7.5) is satisfied, then the joint moment generating function $M(t_1,\ldots,t_n)$ is the same as the joint moment generating function of n independent random variables, the ith of which has the same distribution as X_i. As the joint moment generating function uniquely determines the joint distribution, this must be the joint distribution; hence the random variables are independent.

Example 7.48. Let X and Y be independent normal random variables, each with mean μ and variance σ^2. In Example 6.37 of Chapter 6 we showed that $X+Y$ and $X-Y$ are independent. Let us now establish this result by computing their joint moment generating function.

Solution:

$$\begin{aligned} E\left[e^{t(X+Y)+s(X-Y)}\right] &= E\left[e^{(t+s)X+(t-s)Y}\right] \\ &= E\left[e^{(t+s)X}\right]E\left[e^{(t-s)Y}\right] \\ &= e^{\mu(t+s)+\sigma^2(t+s)^2/2}e^{\mu(t-s)+\sigma^2(t-s)^2/2} \\ &= e^{2\mu t+\sigma^2t^2}e^{\sigma^2s^2}. \end{aligned}$$

But we recognize the preceding as the joint moment generating function of the sum of a normal random variable with mean 2μ and variance $2\sigma^2$

and an independent normal random variable with mean 0 and variance $2\sigma^2$. As the joint moment generating function uniquely determines the joint distribution, it thus follows that $X+Y$ and $X-Y$ are independent normal random variables.

Example 7.49. Suppose that the number of events that occur is a Poisson random variable with mean λ, and that each event is independently counted with probability p. Show that the number of counted events and the number of uncounted events are independent Poisson random variables with means λp and $\lambda(1-p)$ respectively.

Solution:
Let X denote the total number of events, and let X_c denote the number of them that are counted. To compute the joint moment generating function of X_c, the number of events that are counted, and $X-X_c$, the number that are uncounted, start by conditioning on X to obtain

$$\begin{aligned} E\left[e^{sX_c+t(X-X_c)} \mid X=n\right] &= e^{tn}E\left[e^{(s-t)X_c} \mid X=n\right] \\ &= e^{tn}\left(pe^{s-t}+1-p\right)^n \\ &= \left(pe^s+(1-p)e^t\right)^n, \end{aligned}$$

where the preceding equation follows since conditional on $X=n$, X_c is a binomial random variable with parameters n and p. Hence

$$E\left[e^{sX_c+t(X-X_c)} \mid X\right] = \left(pe^s+(1-p)e^t\right)^X.$$

Taking expectations of both sides of the preceding yields that

$$E\left[e^{sX_c+t(X-X_c)}\right] = E\left[\left(pe^s+(1-p)e^t\right)^X\right].$$

Now, since X is Poisson with mean λ, it follows that $E\left[e^{tX}\right] = e^{\lambda(e^t-1)}$. Therefore, for any positive value a we see (by letting $a=e^t$) that $E\left[a^X\right] = e^{\lambda(a-1)}$. Thus

$$\begin{aligned} E\left[e^{sX_c+t(X-X_c)}\right] &= e^{\lambda(pe^s+(1-p)e^t-1)} \\ &= e^{\lambda p(e^s-1)}e^{\lambda(1-p)(e^t-1)}. \end{aligned}$$

As the preceding is the joint moment generating function of independent Poisson random variables with respective means λp and $\lambda(1-p)$, the result is proven.

Chapter 8

Limit Theorems

8.1 Introduction

The most important theoretical results in probability theory are limit theorems. They are classified either under **laws of large numbers** or under the heading **central limit theorems**. Usually, laws of large numbers are concerned with stating conditions under which the average of a sequence of random variables converges (in some sense) to the expected average. Central limit theorems are concerned with determining conditions under which the sum of a large number of random variables has a probability distribution that is approximately normal.

8.2 Chebyshev's Inequality and the Weak Law of Large Numbers

The two inequalities below are for estimating the tail probabilities using minimal information about the random variable such as its mean or its variance.

Proposition 8.1 (Markov's inequality)**.**
Let X be a nonnegative random variable. For $a > 0$, we have

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

Proof. We only prove this for the continuous random variable X. The dis-

crete case is almost the same, with the integral replaced by summation.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{dx} = \int_0^{\infty} x f_X(x)\,\mathrm{dx} \\ &\geq \int_a^{\infty} x f_X(x)\,\mathrm{dx} \geq \int_a^{\infty} a f_X(x)\,\mathrm{dx} = aP(X \geq a). \end{aligned}$$

□

Proposition 8.2 (Chebyshev's inequality)**.**
Let X be a random variable with finite mean μ and variance σ^2, then for $a > 0$, we have

$$P(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Proof. Apply Markov's inequality to $|X-\mu|^2$ and a^2, we get

$$P(|X-\mu| \geq a) = P(|X-\mu|^2 \geq a^2) \leq \frac{E(|X-\mu|^2)}{a^2} = \frac{\mathrm{var}(X)}{a^2}.$$

□

The importance of Markov's and Chebyshev's inequalities is that they enable us to derive bounds on probabilities when only the mean, or both the mean and the variance, of the probability distribution are known. Of course, if the actual distribution were known, then the desired probabilities could be exactly computed and we would not need to resort to bounds.

Example 8.3. Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50.

(a) What can be said about the probability that this week's production will exceed 75?

(b) If the variance of a week's production is known to equal 25, then what can be said about the probability that this week's production will be between 40 and 60?

Solution:
Let X be the number of items that will be produced in a week:

(a) By Markov's inequality

$$P(X > 75) \leq \frac{E(X)}{75} = \frac{50}{75} = \frac{2}{3}.$$

(b) By Chebyshev's inequality

$$P(|X-50| \geq 10) \leq \frac{\sigma^2}{10^2} = \frac{1}{4}.$$

Hence

$$P(|X-50| < 10) \geq 1 - \frac{1}{4} = \frac{3}{4},$$

so the probability that this week's production will be between 40 and 60 is at least 0.75.

Remark 39. As Chebyshev's inequality is valid for all distributions of the random variable X, we cannot expect the bound on the probability to be very close to the actual probability in most cases.

Consequences of this inequality

Proposition 8.4. *If* $\text{var}(X) = 0$, *then the random variable X is a constant. Or, in other words,*

$$P(X = E(X)) = 1.$$

Remark 40. We say that X is degenerate in this case.

Proof. By Chebyshev's inequality, for any $n \geq 1$,

$$0 \leq P\left(|X-\mu| > \frac{1}{n}\right) \leq \frac{\text{var}(X)}{1/n^2} = 0.$$

Therefore,

$$P\left(|X-\mu| > \frac{1}{n}\right) = 0.$$

Now take limits on both sides, and using the continuity property of probability,

$$0 = \lim_{n\to\infty} P\left(|X-\mu| > \frac{1}{n}\right) = P\left(\lim_{n\to\infty}\left\{|X-\mu| > \frac{1}{n}\right\}\right) = P(X \neq \mu).$$

This means that $P(X = \mu) = 1$. □

Theorem 8.5 (The Weak Law of Large Numbers)**.**
Let $X_1, X_2, \ldots$ *be a sequence of independent and identically distributed random variables, with common mean* μ. *Then, for any* $\varepsilon > 0$,

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \longrightarrow 0 \quad \text{as } n \to \infty.$$

Proof. We shall prove this theorem only under the additional assumption that the random variables have a finite variance σ^2. Now, as

$$E\left[\frac{X_1+\cdots+X_n}{n}\right]=\mu$$

and

$$\operatorname{var}\left(\frac{X_1+\cdots+X_n}{n}\right)=\frac{\operatorname{var}(X_1+\cdots+X_n)}{n^2}=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}.$$

It follows from the Chebyshev's inequality that

$$\begin{aligned} P\left(\left|\frac{X_1+\cdots+X_n}{n}-\mu\right|\geq\varepsilon\right) &\leq \frac{\operatorname{var}([X_1+\cdots+X_n]/n)}{\varepsilon^2} \\ &= \frac{\sigma^2}{n\varepsilon^2}\longrightarrow 0. \end{aligned}$$

□

The weak law of large numbers was originally proven by James Bernoulli for the special case where X_i are Bernoulli random variables. The general form of the weak law of large numbers presented was proved by the Russian mathematician Khintchine.

8.3 Central Limit Theorem

The central limit theorem is one of the most remarkable results in probability theory. Loosely put, it states that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but it also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped (that is, normal) curves.

In its simplest form the central limit theorem is as follows.

Theorem 8.6 (Central Limit Theorem)**.**
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then the distribution of

$$\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n\to\infty$. That is,

$$\lim_{n\to\infty} P\left(\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}\leq x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,\mathrm{d}t\,.$$

The key to the proof of this theorem is the following lemma, which we will not prove.

Lemma 8.7. *Let $Z_1, Z_2, \ldots$ be a sequence of random variables having distribution function F_{Z_n} and moment generating function M_{Z_n}, for $n \geq 1$. Let Z be a random variable having distribution function F_Z and moment generating function M_Z. If $M_{Z_n}(t) \to M_Z(t)$ for all t, then $F_{Z_n}(x) \to F_Z(x)$ for all x at which $F_Z(x)$ is continuous.*

Proof. (Partial proof of Theorem 8.6) We shall assume that moment generating function of X_i's exists.
A reduction: By considering $Y_i := \frac{X_i - \mu}{\sigma}$, we can without loss of generality assume that X_i's have mean 0 and variance 1.
Let $Z_n := \frac{X_1 + \cdots + X_n}{\sqrt{n}}$, then

$$\begin{aligned} M_{Z_n}(t) &:= E\left[e^{t(X_1+\cdots+X_n)/\sqrt{n}}\right] \\ &= E\left[e^{\frac{t}{\sqrt{n}}X_1}\right] \cdots E\left[e^{\frac{t}{\sqrt{n}}X_n}\right] = \left[M(t/\sqrt{n})\right]^n. \end{aligned}$$

According to Lemma 8.7, we need to show that

$$M_{Z_n}(t) \longrightarrow e^{t^2/2}$$

or equivalently,

$$\lim_{n\to\infty} n \log M\left(\frac{t}{\sqrt{n}}\right) = \frac{t^2}{2}. \tag{8.1}$$

Now,

$$\begin{aligned}
\lim_{n\to\infty} n\log M\left(\frac{t}{\sqrt{n}}\right) &= \lim_{n\to\infty} \frac{\log M\left(\frac{t}{\sqrt{n}}\right)}{1/n} \\
&= \lim_{x\to 0+} \frac{\log M(\sqrt{x}t)}{x} \\
&= \lim_{x\to 0+} \frac{M'(\sqrt{x}t)}{M(\sqrt{x}t)} \times \frac{t}{2\sqrt{x}} \quad \text{(L'Hospital's Rule for } 0/0) \\
&= \frac{t}{2}\lim_{x\to 0+} \frac{M'(\sqrt{x}t)}{\sqrt{x}} \\
&= \frac{t}{2}\lim_{x\to 0+} \frac{M''(\sqrt{x}t)\, t/(2\sqrt{x})}{1/(2\sqrt{x})} \\
&= \frac{t^2}{2}M''(0) \\
&= \frac{t^2}{2}\, E(X^2) \\
&= t^2/2,
\end{aligned}$$

proving Equation (8.1). □

The first version of the central limit theorem was proved by DeMoivre around 1733 for the special case where the X_j are Bernoulli random variables with $p = \frac{1}{2}$. This was subsequently extended by Laplace to the case of arbitrary p. Laplace also discovered the more general form of the central limit theorem given. His proof, however, was not completely rigorous and, in fact, cannot easily be made rigorous. A truly rigorous proof of the central limit theorem was first presented by the Russian mathematician Liapounoff in the period 1901 – 1902.

Normal Approximation

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each having mean μ and variance σ^2. Then, for n large, the distribution of

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

is approximately standard normal. In other words, for $-\infty < a < b < \infty$, we have

$$P\left(a < \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le b\right) \approx \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, \mathrm{dt}\,.$$

Example 8.8. Let $X_i,\ i = 1,2,\ldots,10$ be independent random variables, each uniformly distributed over $(0,1)$. Calculate an approximation to $P(X_1 + \cdots + X_{10} > 6)$.

Solution:
$E(X_i) = 1/2$, and $\text{var}(X_i) = 1/12$. Therefore

$$\begin{aligned} P(X_1 + \cdots + X_{10} > 6) &\approx P\left(Z > \frac{6 - 10 \times (1/2)}{\sqrt{10 \times (1/12)}}\right) \\ &= P(Z > 1.095) \\ &= 0.1379. \end{aligned}$$

In the following example, since the random variable takes integer values, we therefore introduce continuity correction as in normal approximation to binomial.

Example 8.9. The number of students who enroll in a psychology class is a Poisson random variable with mean 100. The professor in charge of the course decided that if the number of enrollment is 120 or more, he will teach the course in 2 separate sessions, whereas if the enrollment is under 120 he will teach all the students in a single section. What is the probability that the professor will have to teach 2 sessions?

Solution:
Let X be the enrollment in the class. Note that $X \sim \text{Poisson}(100)$, $E(X) = 100 = \text{var}(X)$. The probability required is $P(X \geq 120)$.
Realize that

$$X = X_1 + \cdots + X_{100}$$

where X_i's are independent Poisson random variables with parameter 1. Then

$$\begin{aligned} P(X \geq 120) &= P(X_1 + \cdots + X_{100} \geq 120) \\ &= P(X_1 + \cdots + X_{100} \geq 119.5) \\ &\approx P\left(Z \geq \frac{119.5 - 100}{\sqrt{100}}\right) \\ &= P(Z \geq 1.95) \\ &= 0.0256. \end{aligned}$$

Note that the exact solution

$$e^{-100} \sum_{i=120}^{\infty} \frac{100^i}{i!}$$

does not readily yield a numerical answer.

Example 8.10. Let $X_1, X_2, \ldots, X_{20}$ be independent uniformly distributed random variables over $(0,1)$. We wish to estimate or upper bound

$$P\left(\sum_{k=1}^{20} X_k > 15\right). \qquad (*)$$

(i) Use Markov's inequality to obtain an upper bound on (*).

(ii) Use Chebyshev's inequality to obtain an upper bound on (*).

(iii) Use the Central Limit Theorem to approximate (*).

Solution:

(i) Applying Markov inequality:

$$P\left(\sum_{k=1}^{20} X_k > 15\right) \leq \frac{E\left[\sum_{k=1}^{20} X_k\right]}{15} = \frac{20 \times 1/2}{15} = \frac{2}{3}.$$

(ii) Applying Chebyshev's inequality:

$$\begin{aligned} P\left(\sum_{k=1}^{20} X_k > 15\right) &\leq P\left(\left|\sum_{k=1}^{20} X_k - 10\right| > 5\right) \\ &\leq \frac{\mathrm{var}(\sum_{k=1}^{20} X_k)}{25} = \frac{20 \times 1/12}{25} = \frac{1}{15}. \end{aligned}$$

(iii) Applying central limit theorem:

$$\begin{aligned} P\left(\sum_{k=1}^{20} X_k > 15\right) &= P\left(\frac{\sum_{k=1}^{20} X_k - 20 \times \frac{1}{2}}{\sqrt{20 \times 1/12}} > \frac{15 - 20 \times \frac{1}{2}}{\sqrt{20 \times 1/12}}\right) \\ &\approx P\left(Z > \sqrt{15}\right) = 0.00005. \end{aligned}$$

8.4 The Strong Law of Large Numbers

The following is the best-known result in probability theory:

Theorem 8.11 (The Strong Law of Large Numbers)**.**
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having a finite mean $\mu = E(X_i)$ Then, with probability 1*,*

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \quad \text{as } n \to \infty.$$

In other words,

$$P\left(\left\{\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = \mu\right\}\right) = 1.$$

Example 8.12. Suppose that a sequence of independent trials of some experiment is performed. Let E be a fixed event of the experiment, and let $P(E)$ be the probability that E occurs on any particular trial.

Let

$$X_i = \begin{cases} 1, & \text{if } E \text{ occurs on the } i\text{th trial} \\ 0, & \text{if } E \text{ does not occur on the } i\text{th trial} \end{cases}.$$

By the Strong Law of Large Numbers, we have, with probability 1, that

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to E(X) = P(E).$$

Note that $\frac{X_1+X_2+\cdots+X_n}{n}$ corresponds to the proportion of times that E occurs in the first n trials. So we may interpret the above result as:

"**With probability 1, the limiting proportion of times that E occurs is $P(E)$.**"