

CALCOLO DELLE PROBABILITA'

Chap All.pdf

1. BASIC CONCEPTS

1.1 Sample spaces

We will deal with an experiment which has a “random” outcome, the possible outcomes being, say, $\omega_1, \omega_2, \dots$. The set of all possible outcomes $\Omega = \{\omega_1, \omega_2, \dots\}$ is the **sample space** of the experiment. A particular point $\omega \in \Omega$ is an **observation**.

Examples of experiments

- (i) Tossing a normal six-faced die, $\Omega = \{1, 2, \dots, 6\}$.
- (ii) Picking a card from a standard pack, Ω is the pack of 52 cards.
- (iii) Administer a drug to 20 patients. For simplicity assume there are only two possible outcomes for each patient, R if the patient recovers, D otherwise. Then

$$\Omega = \{(i_1, \dots, i_{20}); i_j = R \text{ or } D\}$$

where i_j is R or D according as patient j recovers or not, so Ω has 2^{20} points.

- (iv) Pick a point from the unit interval, $[0, 1]$, then $\Omega = [0, 1]$.

All but (iv) are examples of discrete sample spaces.

For the moment Ω will be assumed to be **discrete**, i.e., it is a finite or countable set.

A subset A of Ω will be called an **event**. If the experiment is performed with outcome ω and $\omega \in A$, the event A is said to **occur**. In Example (i) A might be the event that “the outcome is even”, so that $A = \{2, 4, 6\}$, in which case A occurs if one of 2, 4, 6 is shown on the die.

Certain set-theoretic notions have special interpretations in probability; the **complement** in Ω of the event A , A^c (alternatively A' , $\bar{A}$) is the event “not A ”, and occurs if and only if A does not occur. The **union** $A \cup B$ of two events A and B is the event “at least one of A or B occurs”. The **intersection** $A \cap B$ is the event “both A and B occur”. The **inclusion** relation $A \subseteq B$ means “the occurrence of A implies the occurrence of B ”. Events A and B are said to be **mutually exclusive** if they are disjoint, $A \cap B = \emptyset$, and so both cannot occur together.

1.2 Classical probability

Classical probability was concerned with modelling situations where there are just a finite number of possible outcomes of the experiment and each of these outcomes is “equally likely”. For example, tossing a fair coin or an unloaded die, or picking a card from a standard well-shuffled pack. Here Ω is a finite set with N points, say, $\Omega = \{\omega_1, \dots, \omega_N\}$, and attached to each event $A \subseteq \Omega$ is a measure of how “likely” the event A is to occur. Here we take $\mathbb{P}(A)$, the **probability** of the event A to be

$$\mathbb{P}(A) = \frac{N_A}{N} = \frac{\text{number of points in } A}{\text{number of points in } \Omega}$$

where N_A is the number of points in A . Thus the probability $\mathbb{P}(\cdot)$ is just a real-valued function defined on subsets A of Ω .

Properties of $\mathbb{P}(\cdot)$

1. $0 \leq \mathbb{P}(A) \leq 1$.
2. $\mathbb{P}(\Omega) = 1$.
3. If $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

In a while we will take properties 1, 2 and an extension of 3 as the basis of an axiomatic development of probability in a more general context. The following properties follow immediately from the definition of $\mathbb{P}(\cdot)$ above, but they may be deduced easily from 1, 2 and 3 without reference to the definition.

4. $\mathbb{P}(\emptyset) = 0$.
5. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
6. If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
7. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
8. $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

1.3 Combinatorial analysis

Fundamental rule: There are r multiple choices to be made in sequence: there are m_1 possibilities for the first choice; after making the first choice there are m_2 possibilities for the second choice; after making the first two choices there are m_3 possibilities for the third

choice, and so on until after making the first $r - 1$ choices there are m_r possibilities for the r th choice. Then the total number of different possibilities for the set of choices is

$$m_1 m_2 \cdots m_r$$

Example 1.1 A restaurant menu has 6 starters, 7 main courses and 6 puddings. The total number of different three-course meals that may be served is $6 \times 7 \times 6 = 252$. $\square$

Sampling models

Many of the standard calculations that arise in classical probability where outcomes need to be counted may be put in a standard framework of **sampling**. Think of drawing m balls from an urn which initially contains n distinguishable balls (they are numbered 1 to n , say). This may be done in a number of ways:

1. Sampling with replacement and with ordering. The balls are replaced between successive draws and the order in which balls are drawn is noted. Then the Fundamental Rule shows that there are n^m possible ways.

2. Sampling without replacement and with ordering ($m \leq n$). The balls are not replaced after drawing and the order is noted. Then the number of ways is

$$n(n-1) \cdots (n-m+1) = \frac{n!}{(n-m)!} = {}^n P_m;$$

the symbol ${}^n P_m$ represents the number of permutations of n objects m at a time.

An important special case occurs when $n = m$. We get the number of permutations of n distinguishable objects (that is the number of distinguishable ways they may be laid out in a line, say) is $n!$.

3. Sampling without replacement and without ordering. If we take m balls from n and they were ordered there would be $n(n-1) \cdots (n-m+1) = n!/(n-m)!$ ways but each unordered selection may be permuted in $m!$ different ways (or give $m!$ ordered arrangements) so that the total number of unordered ways is

$$\frac{n!}{(n-m)! m!} = \binom{n}{m} = {}^n C_m;$$

the symbol $\binom{n}{m}$ is the binomial coefficient, usually read " n choose m ", and represents the number of ways of picking m objects from n ; note that $\binom{n}{m}$ is the coefficient of x^m in the expansion of $(1+x)^n$ using the binomial theorem.

Suppose that, of the n balls, m_1 are of colour 1, m_2 are of colour 2 and so on up to m_k balls of colour k , where $n = m_1 + \dots + m_k$. Consider the number of permutations of the n balls when the balls are distinguishable only by colour. For example, if $n = 4$ and there are two black balls and two white balls then there are 6 possible arrangements:

BBWW BWBW BWWB WBBW WBWB WWBB

If the balls are distinguishable then there are $n!$ permutations, but within colour i there are $m_i!$ ways of permuting the balls so the Fundamental Rule gives that for each distinguishable arrangement there are $\prod_1^k (m_i!)$ ways of permuting the balls leaving the arrangement the same if the balls are distinguishable only by colour. Thus the number of arrangements distinguishable only by colour is

$$\frac{n!}{m_1! m_2! \dots m_k!} = \binom{n}{m_1 \dots m_k},$$

which is the multinomial coefficient (the coefficient of $x_1^{m_1} x_2^{m_2} \dots x_k^{m_k}$ in the expansion of $(x_1 + \dots + x_k)^n$). An alternative way of seeing this is to think of first choosing the positions for the balls of colour 1, which can be done in $\binom{n}{m_1}$ ways, then choose the positions for the balls of colour 2 in $\binom{n-m_1}{m_2}$ ways and so on to see that the total number of ways, using the Fundamental Rule, is

$$\binom{n}{m_1} \binom{n-m_1}{m_2} \binom{n-m_1-m_2}{m_3} \dots \binom{n-m_1-\dots-m_{k-1}}{m_k} = \binom{n}{m_1 \dots m_k}.$$

4. Sampling with replacement and without ordering. Draw m balls one after another, each time noting the number of the ball and replacing it before the next draw,

Ball No.	1	2	3	4	...	n
Times drawn	✓	✓✓	✓	✓✓	...	✓

so that the number of $\checkmark$ is the number of times the ball is drawn. The number of ways is then the number of ways that $n - 1$ vertical lines may be put between $m \checkmark$ which is

$$\binom{n+m-1}{n-1};$$

think of choosing $n - 1$ slots for the vertical lines out of $n + m - 1$ slots and then the checks go in the remainder.

Allocation models

An alternative way of thinking of these counting schemes is to think about allocating m tokens (labelled $1, \dots, m$) to n boxes (labelled $1, \dots, n$). The four cases considered previously correspond to:

1. Each box may contain any number of tokens and the labels on the tokens are observed.
2. No box may contain more than one token and the labels on the tokens are observed.
3. No box may contain more than one token and the labels on the the tokens are **not** observed.
4. Each box may contain any number of tokens and the labels on the tokens are **not** observed.

In the next examples in which each outcome in the probability space is equally likely we will calculate the probability of the event A by computing the number of points N in the sample space Ω and then computing the number N_A of points in A . In each case the probability will then be

$$\mathbb{P}(A) = \frac{N_A}{N}.$$

Example 1.2 What is the probability that a Poker hand shows five different face values? (A poker hand contains 5 cards.) One way to do this is to think of Ω consisting of all possible poker hands, that is, unordered sets of five cards chosen from a standard pack of 52. We have $N = \binom{52}{5}$ and A consists of those hands showing five different face values so that

$$N_A = \binom{13}{5} \times 4^5$$

since we may think of first choosing the unordered sets of 5 face values and then the different suits for each value, and use the Fundamental Rule. An alternative way in this problem, and in other situations involving sampling without replacement, is to think of Ω comprising all ordered sets of 5 cards (imagine the cards being dealt in sequence). Then $N = 52 \times 51 \times 50 \times 49 \times 48$ and $N_A = 52 \times 48 \times 44 \times 40 \times 36$, thinking of the possible choices for the first card, second card in the hand and so on to give 5 different face values. You should check that these two different approaches give the same probabilities. Either approach is fine, but remember to be consistent – if your sample space has unordered (respectively, ordered) points then the points in A must be unordered (respectively, ordered). $\square$

Example 1.3 What is the probability that a Bridge hand (13 cards) contains 5 spades ($\spadesuit$), 3 hearts ($\heartsuit$), 4 diamonds ($\diamondsuit$) and 1 club ($\clubsuit$)? Take Ω to be the set of unordered Bridge hands, so we have $N = \binom{52}{13}$ while

$$N_A = \binom{13}{5} \times \binom{13}{3} \times \binom{13}{4} \times \binom{13}{1}$$

since we may think of choosing in sequence the spades, hearts, diamonds and club and then use the Fundamental Rule to get the total number of ways. $\square$

♦ **Example 1.4** If there are r people in a room, what is the probability that at least two have the same birthday? In this case the number of points in the sample space Ω is $N = (365)^r$, since the sample space consists of all possible r -tuples of birthdays. If A is the required event, this is a case where it is easier to calculate the number of points in A^c , the complement of A , since A^c is just the event that no two of the r people share a birthday. We then have

$$N_{A^c} = 365 \times 364 \times 363 \times \cdots \times (365 - r + 1),$$

and so

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \frac{365 \times 364 \times 363 \times \cdots \times (365 - r + 1)}{(365)^r} = p_r, \quad \text{say.}$$

It is interesting to note the following values

r	10	15	20	21	22	23	24	25	30	40	55
p_r	0.12	0.25	0.41	0.44	0.48	0.51	0.54	0.57	0.71	0.89	0.99

which shows the well known fact that if there are 23 or more people in a room there is higher than evens chances that at least 2 share a birthday. $\square$

Theorem 1.5 (*Stirling's Formula*) As $n \rightarrow \infty$,

$$\log \left(\frac{n! e^n}{n^{n+\frac{1}{2}}} \right) = \log (\sqrt{2\pi}) + O(1/n).$$

Proof. Let $c_n = \log(n!) + n - (n + \frac{1}{2}) \log n$, then

$$c_n - c_{n+1} = (n + \frac{1}{2}) \log(1 + 1/n) - 1. \quad (1.6)$$

For $0 < x < 1$, if we subtract the two expressions

$$\begin{aligned} \log(1+x) - x &= -\frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \\ \log(1-x) + x &= -\frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \dots \end{aligned}$$

and divide by $2x$, we obtain

$$\frac{1}{2x} \log \left(\frac{1+x}{1-x} \right) - 1 = \frac{x^2}{3} + \frac{x^4}{5} + \frac{x^6}{7} + \dots \quad (1.7)$$

$$\leq \frac{x^2}{3} + \frac{x^4}{3} + \frac{x^6}{3} + \dots = \frac{1}{3} \frac{x^2}{1-x^2}. \quad (1.8)$$

Now put $x = 1/(2n+1)$ and we see from (1.6) that the left-hand side of (1.7) is then $c_n - c_{n+1}$, and furthermore from (1.7) that $c_n - c_{n+1} \geq 0$ and from (1.8) that

$$c_n - c_{n+1} \leq \frac{1}{12} \left(\frac{1}{n} - \frac{1}{n+1} \right).$$

This shows that c_n is monotone non-increasing in n and $c_n - 1/(12n)$ is monotone non-decreasing so that c_n is bounded below and hence converges $c_n \rightarrow c$, for some c .

To determine the value of c , define $I_r = \int_0^{\pi/2} \sin^r \theta d\theta$, for $r \geq 0$, so that $I_0 = \pi/2$ and $I_1 = 1$. Integrating by parts for $r > 1$, we have

$$I_r = [-\sin^{r-1} \theta \cos \theta]_0^{\pi/2} + (r-1) \int_0^{\pi/2} \sin^{r-2} \theta \cos^2 \theta d\theta = (r-1)(I_{r-2} - I_r),$$

so that $rI_r = (r-1)I_{r-2}$. It is immediate that $I_{2n+1} \leq I_{2n} \leq I_{2n-1}$, hence

$$1 \leq \frac{I_{2n}}{I_{2n+1}} \leq \frac{I_{2n-1}}{I_{2n+1}} = \frac{2n+1}{2n} \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

Calculate
$$I_{2n} = \frac{2n-1}{2n} I_{2n-2} = \frac{(2n-1)(2n-3) \cdots 1}{(2n)(2n-2) \cdots 2} I_0 = \frac{(2n)!}{(2^n n!)^2} \frac{\pi}{2}$$

and
$$I_{2n+1} = \frac{2n}{2n+1} I_{2n-1} = \frac{(2n)(2n-2) \cdots 2}{(2n+1)(2n-1) \cdots 1} I_1 = \frac{(2^n n!)^2}{(2n+1)!}.$$

Dividing these we see that, as $n \rightarrow \infty$,

$$\frac{I_{2n}}{I_{2n+1}} = \frac{(2n+1)((2n)!)^2 \pi}{(2^n n!)^4} \frac{1}{2} \rightarrow 1, \quad \text{or} \quad \frac{(2^n n!)^2}{(2n)!} \frac{1}{\sqrt{n}} \rightarrow \sqrt{\pi}.$$

Now note that

$$2c_n - c_{2n} = \log \sqrt{2} + \log \left(\frac{(2^n n!)^2}{(2n)!} \frac{1}{\sqrt{n}} \right) \rightarrow \log \sqrt{2} + \log \sqrt{\pi} = \log \sqrt{2\pi},$$

so that $c = \lim_{n \rightarrow \infty} (2c_n - c_{2n}) = \log \sqrt{2\pi}$, as required. $\square$

The most common statement of Stirling's formula is given as a corollary.

Corollary 1.9 *As $n \rightarrow \infty$, $n! \sim \sqrt{2\pi e}^{-n} n^{n+\frac{1}{2}}$, where, in this context, $\sim$ indicates that the ratio of the two sides tends to 1.*

Example 1.10 Suppose that $4n$ balls, of which $2n$ are red and $2n$ are black, are put at random into two urns, so that each urn contains $2n$ balls. What is the probability that each urn contains n red balls and n black balls? The probability is

$$\begin{aligned} p_n &= \binom{2n}{n} \binom{2n}{n} / \binom{4n}{2n} = \left(\frac{(2n)!}{n!} \right)^4 \frac{1}{(4n)!} \\ &\sim \left(\frac{\sqrt{2\pi e}^{-2n} (2n)^{2n+\frac{1}{2}}}{\sqrt{2\pi e}^{-n} n^{n+\frac{1}{2}}} \right)^4 \frac{1}{\sqrt{2\pi e}^{-4n} (4n)^{4n+\frac{1}{2}}} = \sqrt{\frac{2}{\pi n}} = a_n, \text{ say.} \end{aligned}$$

We may calculate the exact probability p_n , its approximant a_n from Stirling's formula and their ratio for different values of n .

n	1	10	13	20	40	60	100
p_n	0.667	0.248	0.218	0.177	0.126	0.103	0.080
a_n	0.798	0.252	0.221	0.178	0.126	0.103	0.080
p_n/a_n	0.836	0.981	0.986	0.991	0.995	0.997	0.998

The case $n = 13$ corresponds to dividing a standard pack of cards in two and obtaining equal numbers of red and black cards in each half. □

9 January 2007

2. AXIOMATIC PROBABILITY

2.1 The axioms

The formulation for classical probability in which all outcomes or points in the sample space are equally likely is too restrictive to develop a useful theory of probability. In this section we give a general definition. For the moment we will assume that the sample space of outcomes Ω is either a finite or countable set.

A **probability distribution** $\mathbb{P}$ on Ω is a real-valued function defined on subsets (events) of Ω satisfying the following three conditions:

1. $0 \leq \mathbb{P}(A) \leq 1$, for all $A \subseteq \Omega$.
2. $\mathbb{P}(\Omega) = 1$.
3. For a finite or infinite collection of disjoint events $\{A_i\}_i$, that is $A_i \cap A_j = \emptyset$ for $i \neq j$, we have

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i).$$

First observe the following consequence of Axiom 3.

4. $\mathbb{P}(\emptyset) = 0$.

This follows because if we take $A_i = \emptyset$, for each i in Axiom 3 then we get a contradiction unless $\mathbb{P}(\emptyset) = 0$. Note that Axiom 3 here implies the property 3. that we had in Chapter 1 if we take $A_i = \emptyset$ for $i > 2$, $A_1 = A$ and $A_2 = B$. This of course extends to n disjoint events $A_1, \dots, A_n$ for which $\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i)$. An important consequence of this axiom is a continuity property of probabilities in that for events $B_1 \subseteq B_2 \subseteq B_3 \subseteq \dots$ we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \lim_{i \rightarrow \infty} \mathbb{P}(B_i). \quad (2.1)$$

This follows by taking $A_1 = B_1$ and for $i > 1$, $A_i = B_i \cap B_{i-1}^c$, then the $\{A_i\}$ are disjoint with $\bigcup_{j=1}^i A_j = B_i$ and $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ so that $\mathbb{P}(B_i) = \sum_{j=1}^i \mathbb{P}(A_j)$, and

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \lim_{i \rightarrow \infty} \sum_{j=1}^i \mathbb{P}(A_j) = \lim_{i \rightarrow \infty} \mathbb{P}(B_i).$$

As in Chapter 1 we may deduce that

5. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
6. If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
7. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Using (2.1) and Property 5. we may see a corresponding continuity property for probabilities on decreasing events, in that for events $B_1 \supseteq B_2 \supseteq B_3 \supseteq \dots$, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} B_i\right) = \lim_{i \rightarrow \infty} \mathbb{P}(B_i).$$

Boole's Inequality We may deduce from Property 7, that for any events $A_1, \dots, A_n$,

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

The case $n = 2$ follows from Property 7 (and Axiom 1) immediately and the case for general n by induction. We may use (2.1) to extend Boole's Inequality to the case of countably many events $A_1, A_2, \dots$ so that

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

2.2 Inclusion-exclusion formula

The inclusion-exclusion formula gives a method of calculating that at least one of a number of events occurs.

Theorem 2.2 (Inclusion-exclusion) *For any events $A_1, A_2, \dots, A_n$,*

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_i \mathbb{P}(A_i) - \sum_{i_1 < i_2} \mathbb{P}(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum_{i_1 < i_2 < i_3} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \dots + (-1)^{n-1} \mathbb{P}\left(\bigcap_{i=1}^n A_i\right). \end{aligned}$$

Proof. The proof is by induction on n . The case $n = 2$ is just Property 7. Assume the result holds for any n events and consider $A_1, \dots, A_{n+1}$. Now see from Property 7. again that

$$\begin{aligned}\mathbb{P}\left(\bigcup_1^{n+1} A_i\right) &= \mathbb{P}\left(\bigcup_1^n A_i\right) + \mathbb{P}(A_{n+1}) - \mathbb{P}\left(\left(\bigcup_1^n A_i\right) \cap A_{n+1}\right) \\ &= \mathbb{P}\left(\bigcup_1^n A_i\right) + \mathbb{P}(A_{n+1}) - \mathbb{P}\left(\bigcup_1^n (A_i \cap A_{n+1})\right),\end{aligned}\quad (2.3)$$

and then apply the inductive hypothesis to the first and the third terms to complete the induction and obtain the conclusion. $\square$

Example 2.4 What is the probability that in a round of Bridge at least one player holds exactly two aces and two kings? Let Ω be the sample space of all possible (unordered) hands for 4 players, then the number of points in Ω is $N = \binom{52}{13\ 13\ 13\ 13}$. For each player i , $i = 1, 2, 3, 4$, let A_i be the event that player i holds exactly two aces and two kings. Then

$$\mathbb{P}(A_i) = \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9}\binom{39}{13\ 13\ 13}}{\binom{52}{13\ 13\ 13\ 13}} = \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9}}{\binom{52}{13}};$$

think of choosing the aces for player i , then the kings and then 9 other cards and finally distributing the remaining 39 cards among the other three players. If $i \neq j$, then

$$\mathbb{P}(A_i \cap A_j) = \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9\ 9\ 26}\binom{26}{13}}{\binom{52}{13\ 13\ 13\ 13}} = \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9\ 9\ 26}}{\binom{52}{13\ 13\ 26}};$$

as before, think of picking the aces and kings for player i with the remaining aces and kings going to player j , and then picking 9 other cards for each of i and j . If i, j, k are unequal then $A_i \cap A_j \cap A_k = \emptyset$, so that by inclusion exclusion we have the required probability is

$$\mathbb{P}\left(\bigcup_{i=1}^4 A_i\right) = 4 \times \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9}}{\binom{52}{13}} - \binom{4}{2} \times \frac{\binom{4}{2}\binom{4}{2}\binom{44}{9\ 9\ 26}}{\binom{52}{13\ 13\ 26}}.$$

$\square$

• **Example 2.5** Suppose that n students leave their n coats outside a lecture room and when they leave they pick up their coats at random. What is the probability that at least one student has his own coat? Let Ω consist of all permutations of $1, \dots, n$, so that if $(i_1, \dots, i_n) \in \Omega$ then i_j is the index of the coat got by student j . Denote by A_i the event that student i gets his own coat. Then for $i_1 < i_2 < \dots < i_r$, the probability that all of the students $i_1, \dots, i_r$ get their own coats is

$$\mathbb{P}\left(\bigcap_{k=1}^r A_{i_k}\right) = \frac{(n-r)!}{n!} = \frac{1}{\binom{n}{r} r!}$$

since the numerator is the number of ways that the remaining $n-r$ coats may be permuted among the remaining $n-r$ students. Then by inclusion-exclusion we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_{r=1}^n \left[(-1)^{r-1} \sum_{i_1 < \dots < i_r} \mathbb{P}\left(\bigcap_{k=1}^r A_{i_k}\right) \right] \\ &= \sum_{r=1}^n (-1)^{r-1} \binom{n}{r} \frac{(n-r)!}{n!} = \sum_{r=1}^n (-1)^{r-1} \frac{1}{r!},\end{aligned}$$

so that for large n the probability is approximately equal to

$$1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \dots = 1 - e^{-1} \approx 0.632. \quad \square$$

Corollary 2.6 (Bonferroni's Inequalities) *For any events $A_1, A_2, \dots, A_n$ and for any r , $1 \leq r \leq n$,*

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &\leq \sum_i \mathbb{P}(A_i) - \sum_{i_1 < i_2} \mathbb{P}(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum_{i_1 < i_2 < i_3} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \dots + (-1)^{r-1} \sum_{i_1 < \dots < i_r} \mathbb{P}(A_{i_1} \cap \dots \cap A_{i_r}),\end{aligned}$$

according as r is odd or even. That is, if the sum in the inclusion-exclusion formula is truncated after r terms it overestimates or underestimates the probability of the union of the n events according as r is odd or even.

Proof. The proof proceeds by induction on n . Assume it is true for n , then for $n+1$ events it is true when $r = n+1$, by the inclusion-exclusion formula, and for $r \leq n$, apply

the inductive hypothesis to probability of the two unions of n events in relation (2.3) to get the result. $\square$

2.3 Conditional probability

Suppose that $B \subseteq \Omega$ is an event with $\mathbb{P}(B) > 0$. For any event $A \subseteq \Omega$, the **conditional probability of A given B** is

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)};$$

it is the probability that the event A occurs, that is that the outcome of the ‘experiment’ is in A , if it is known that the outcome is in the event B , i.e., that B occurs.

Example 2.7 What is the probability that a Bridge hand contains the ace of hearts given that it contains exactly 5 hearts? Let A be the event that the hand contains the ace of hearts and B the event that the hand contains exactly 5 hearts. Then

$$\mathbb{P}(B) = \frac{\binom{13}{5} \binom{39}{8}}{\binom{52}{13}} \quad \text{and} \quad \mathbb{P}(A \cap B) = \frac{\binom{12}{4} \binom{39}{8}}{\binom{52}{13}},$$

whence $\mathbb{P}(A | B) = \binom{12}{4} / \binom{13}{5} = \frac{5}{13}$. $\square$

The first thing to observe is that $\mathbb{P}(\cdot | B)$ is a probability distribution on the sample space B , because it satisfies the axioms for a probability distribution as follows:

1. For $C \subseteq B$, $\mathbb{P}(C | B) = \mathbb{P}(C)/\mathbb{P}(B)$, so that $0 \leq \mathbb{P}(C | B) \leq 1$.
2. $\mathbb{P}(B | B) = \mathbb{P}(B)/\mathbb{P}(B) = 1$.
3. For disjoint events $C_1, C_2, \dots$ in B ,

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^{\infty} C_i \mid B\right) &= \frac{\mathbb{P}(\bigcup_{i=1}^{\infty} C_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\bigcup_{i=1}^{\infty} C_i)}{\mathbb{P}(B)} \\ &= \frac{\sum_{i=1}^{\infty} \mathbb{P}(C_i)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(C_i | B). \end{aligned}$$

Multiplication rule The next thing to notice is the multiplication rule that

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

so that the probability of two events occurring can be broken up into calculating successive probabilities—firstly the probability that B occurs and then given that B has occurred the probability that A occurs. This is one of two central procedures for calculating probabilities (the second is the Law of Total Probability introduced below).

- **Example 2.8** Suppose that two students are selected, without replacement, from a class of 5 women and 13 men. What is the probability that the first student selected is a man and the second is a woman? Let B be the event that the first is a man and A the event that the second is a woman, then $\mathbb{P}(A \cap B) = \mathbb{P}(A | B)\mathbb{P}(B) = \frac{5}{17} \times \frac{13}{18}$. $\square$

More generally we can write down the multiplication rule for events $A_1, \dots, A_n$,

$$\mathbb{P}(A_1 \cap A_2 \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 | A_1)\mathbb{P}(A_3 | A_1 \cap A_2) \cdots \mathbb{P}(A_n | A_1 \cap \cdots \cap A_{n-1}).$$

Example 2.9 In drawing three cards without replacement from a pack what is the probability of three successive aces? Here A_i would be the event that an ace is obtained on draw i , $i = 1, 2, 3$, then

$$\begin{aligned} \mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_1)\mathbb{P}(A_2 | A_1)\mathbb{P}(A_3 | A_1 \cap A_2) \\ &= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50}. \end{aligned} \quad \square$$

Law of Total Probability A collection $\{B_i\}_{i=1}^{\infty}$ of disjoint events for which $\bigcup_{i=1}^{\infty} B_i = \Omega$ is said to be a **partition** of the sample space Ω . For any partition of the sample space, $\{B_i\}$, and for any event A , we may write

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(A | B_i) \mathbb{P}(B_i), \quad (2.10)$$

where the second summation extends only over those events B_i in the partition for which $\mathbb{P}(B_i) > 0$. The identity in (2.10) is known as the Law of Total Probability. It follows immediately from Axiom 3 and the multiplication rule, because the event A may be represented as $A = \bigcup_{i=1}^{\infty} (A \cap B_i)$.

• **Example 2.11** An urn contains b black balls and r red balls from which two balls are drawn without replacement. What is the probability that the second ball drawn is black? Let A represent the event that the second ball is black and B the event that the first ball is black. Then B and B^c form a partition of the sample space (think of the other events in the partition as being $\emptyset$). Then

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A | B)\mathbb{P}(B) + \mathbb{P}(A | B^c)\mathbb{P}(B^c) \\ &= \left(\frac{b-1}{b+r-1}\right)\left(\frac{b}{b+r}\right) + \left(\frac{b}{b+r-1}\right)\left(\frac{r}{b+r}\right) = \frac{b}{b+r}.\end{aligned}$$

This probability may be seen to be the same as the probability that the second ball is black when sampling with replacement. $\square$

Example 2.12 A survey of US voters shows the following figures for proportions of voters registered with the main parties and the proportions of voters registered for each of the parties who declare an intention to vote for Dubya.

	Registered	Proportion for Bush
Democratic	45%	10%
Republican	35%	60%
Not affiliated	20%	40%

Suppose that a voter is chosen at random, what is the probability (s)he is a Bush voter? Here the partition of the sample space is D (Democrat), R (Republican) and NA (No affiliation), and if B is the event the voter is a Bush supporter then

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B | D)\mathbb{P}(D) + \mathbb{P}(B | R)\mathbb{P}(R) + \mathbb{P}(B | NA)\mathbb{P}(NA) \\ &= 0.1 \times 0.45 + 0.6 \times 0.35 + 0.4 \times 0.2 = 0.335.\end{aligned}$$

$\square$

Bayes' Theorem For any events A and B , for which $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, we have

$$\mathbb{P}(B | A) = \frac{\mathbb{P}(A | B)\mathbb{P}(B)}{\mathbb{P}(A)}. \quad (2.13)$$

Example 2.14 In Example 2.12, given that the voter chosen is a Bush supporter, what is the probability that (s)he is a Republican?

$$\mathbb{P}(R | B) = \frac{\mathbb{P}(B | R)\mathbb{P}(R)}{\mathbb{P}(B)} = 0.6 \times 0.35 / 0.335 \approx 0.63.$$

$\square$

Combining the Law of Total Probability with the statement in (2.13), gives the general statement of Bayes' Theorem:

Theorem 2.15 (Bayes' Theorem) *Suppose that $\{B_i\}_i$ is a partition of the sample space and that A is an event for which $\mathbb{P}(A) > 0$. Then for any event, B_i , in the partition with $\mathbb{P}(B_i) > 0$, we have*

$$\mathbb{P}(B_i | A) = \frac{\mathbb{P}(A | B_i) \mathbb{P}(B_i)}{\sum_j \mathbb{P}(A | B_j) \mathbb{P}(B_j)},$$

where the summation in the denominator extends over all j for which $\mathbb{P}(B_j) > 0$.

• **Example 2.16** Consider a diagnostic test for some disease for which the outcome of the test is either positive, +, or negative, −, and which is 99% accurate, so that if D represents the event that the patient has the disease then

$$\mathbb{P}(+ | D) = 0.99 = \mathbb{P}(- | D^c).$$

Suppose that 0.1% of patients have the disease. A patient is chosen at random and tests positive, what is the probability that (s)he has the disease? Then from Bayes' Theorem

$$\begin{aligned} \mathbb{P}(D | +) &= \frac{\mathbb{P}(+ | D) \mathbb{P}(D)}{\mathbb{P}(+ | D) \mathbb{P}(D) + \mathbb{P}(+ | D^c) \mathbb{P}(D^c)} \\ &= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.01 \times 0.999} \approx 0.09. \end{aligned}$$

At first sight this conclusion, that the probability that someone testing positive has a somewhat low probability of having the disease, may be counter intuitive. It arises because the probability of a 'false positive', $\mathbb{P}(+ | D^c)$, is large compared to the probability of the disease in the population, $\mathbb{P}(D)$. Notice that

$$\mathbb{P}(D | +) = \frac{1}{1 + (\mathbb{P}(+ | D^c) \mathbb{P}(D^c) / \mathbb{P}(+ | D) \mathbb{P}(D))};$$

typically $\mathbb{P}(D^c)$ and $\mathbb{P}(+ | D)$ will be close to 1 so that $\mathbb{P}(D | +)$ will be large or small according as the ratio $\mathbb{P}(+ | D^c) / \mathbb{P}(D)$ is small or large. This becomes less mysterious when you consider a population of 1,000 patients with just one person suffering from the disease. Of the 999 not suffering from the disease there will be about 10 who will test positive so there will be about 11 in the population who will test positive; given that

the person selected has tested positive then there is about a 1 in 11 chance that the person is the one suffering from the disease. A similar calculation to the above shows that $\mathbb{P}(D | -) \approx 0.000001$. $\square$

Example 2.17 *Simpson's paradox* One example of conditional probability that appears counter-intuitive when first seen is the following situation which can arise frequently. Consider one individual chosen at random from 50 men and 50 women applicants to a particular College. Figures on the 100 applicants are given in the following table indicating whether they were educated at a state school or at an independent school and whether they were admitted or rejected.

All applicants	Admitted	Rejected	% Admitted
State	25	25	50%
Independent	28	22	56%

Note that overall the probability that an applicant is admitted is 0.53, but conditional on the candidate being from an independent school the probability is 0.56 while conditional on being from a state school the probability is lower at 0.50. Suppose that when we break down the figures for men and women we have the following figures.

Men only	Admitted	Rejected	% Admitted
State	15	22	41%
Independent	5	8	38%

Women only	Admitted	Rejected	% Admitted
State	10	3	77%
Independent	23	14	62%

It may now be seen that now for both men and women the conditional probability of being admitted is higher for state school applicants, at 0.41 and 0.77, respectively. This may seem to be a surprising result in that while the aggregate data suggests that independent school applicants have a better chance of being admitted, the individual tables for both men and women show that in both cases state school applicants have a higher acceptance rate. A

result of this type in tables of this sort (called **contingency tables** in statistics) is known as Simpson's paradox. Strictly, it is not a paradox in that it has an easy explanation. Note that overall women have a much higher acceptance rate (66%) than men (40%) whereas the proportion of men from state schools is 74% with 26% from independent schools while those proportions are reversed for women. This situation is known in statistics as **confounding**; it occurs when figures for two separate and different populations are aggregated to give misleading conclusions. It may be difficult or impossible to determine whether data such as that presented in the first table arise from two disparate populations and so confounding is present.

The example shows that if A , B , C are three events it is possible to have the three inequalities

$$\mathbb{P}(A | B \cap C) > \mathbb{P}(A | B \cap C^c), \quad \mathbb{P}(A | B^c \cap C) > \mathbb{P}(A | B^c \cap C^c), \quad \text{and} \quad (2.18)$$

$$\mathbb{P}(A | C) > \mathbb{P}(A | C^c), \quad (2.19)$$

holding simultaneously. In this example, A would be the event 'being admitted', B the event 'being a man' (with B^c being a woman) and C being 'state school' (with C^c being independent school). One situation where the three inequalities in (2.18) and (2.19) cannot hold simultaneously is when

$$\mathbb{P}(B | C) = \mathbb{P}(B | C^c). \quad (2.20)$$

To see this, suppose that (2.18) and (2.20) hold, then we see that

$$\mathbb{P}(A \cap B | C) > \mathbb{P}(A \cap B | C^c) \quad \text{and} \quad \mathbb{P}(A \cap B^c | C) > \mathbb{P}(A \cap B^c | C^c)$$

and then adding and observing that, for example,

$$\mathbb{P}(A \cap B | C) + \mathbb{P}(A \cap B^c | C) = \mathbb{P}(A | C),$$

we obtain $\mathbb{P}(A | C) > \mathbb{P}(A | C^c)$, which contradicts (2.19). In this example it would not be possible to engineer that (2.20) holds, but consider the situation where a clinical trial of a new drug for some illness is being conducted to test its effectiveness against a standard drug. Then A would be the event that a patient recovers from the illness and C would

be the event that the patient receives the new drug (C^c the event that (s)he receives the standard drug); again B or B^c would correspond to the patient being a man or woman, respectively. Then to avoid the sort of situation described in this example, (2.20) would require that the trial be designed so that the relative proportions of men and women that receive the new drug are the same as the relative proportions that receive the standard drug. $\square$

2.4 Independence

We say that two events A and B are **independent** if

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

Example 2.21 Roll a die twice and let A be the event that a 1 is obtained on the first roll and B be the event that a 5 is obtained on the second roll. The sample space $\Omega = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$ has 36 points, all of which have the same probability $\frac{1}{36}$. Then $A \cap B$ is just the outcome $(1, 5)$ and $\mathbb{P}(A) = \frac{1}{6} = \mathbb{P}(B)$ so we can conclude that A and B are independent. $\square$

Notice that if $\mathbb{P}(B) > 0$ then A and B are independent if and only if $\mathbb{P}(A | B) = \mathbb{P}(A)$. Furthermore, if A and B are independent then

- (i) A and B^c are independent;
- (ii) A^c and B^c are independent; and
- (iii) A^c and B are independent.

For (i),

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

and (ii) and (iii) follow from (i).

Def We need to generalize the notion of independence to more than two events. Events $A_1, A_2, \dots$ are independent if for all choices of $i_1 < i_2 < \dots < i_r$, we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_r}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_r}). \quad (2.22)$$

The definition in (2.22) implies that if we take any two of the events A_i and A_j (i distinct from j) then they are independent; so the events are said to be **pairwise independent**.

However, it should be noted that **pairwise independence does not imply independence** as the next example shows.

• **Example 2.23** Suppose that a fair coin is tossed twice (by a fair coin we mean the 4 possible outcomes HH, HT, TH and TT are equally likely and equal to $\frac{1}{4}$, so that a Head or a Tail on either toss has probability $\frac{1}{2}$). Let A_1 be the event that a head is obtained on the first toss, A_2 the event that there is a head on the second toss and A_3 the event that exactly 1 head is obtained. Then $\mathbb{P}(A_i) = \frac{1}{2}$ for $i = 1, 2, 3$. It may be seen that the events are pairwise independent since, for example,

$$\mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1 \cap A_2^c) = \frac{1}{4} = \mathbb{P}(A_1)\mathbb{P}(A_3),$$

and similarly for A_2 and A_3 (and A_1 and A_2). But

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = 0 \neq \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = \frac{1}{8},$$

so the three events are not independent. □

We need to see how to model the notion of independent experiments where the outcome of one experiment does not influence the outcome of the other and see how it relates to this definition of independence. Suppose that Ω_1 and Ω_2 are the sample spaces for two experiments with probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively. Let $\Omega = \Omega_1 \times \Omega_2$ be the sample space corresponding to both experiments being performed, so that $\Omega = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$. We can define the probability distribution $\mathbb{P}$ on Ω by specifying $\mathbb{P}$ for any point (singleton) of Ω (see next section), by letting $\mathbb{P}(\{(\omega_1, \omega_2)\}) = \mathbb{P}_1(\{\omega_1\})\mathbb{P}_2(\{\omega_2\})$. If now $A_i \subseteq \Omega_i$, $i = 1, 2$, and both experiments are performed we can identify the event A_1 with the event $A_1 \times \Omega_2$ in Ω and the event A_2 with the event $\Omega_1 \times A_2$ in Ω , and the intersection $A_1 \cap A_2$ with $A_1 \times A_2$, so that

$$\begin{aligned} \mathbb{P}(A_1 \cap A_2) &= \sum_{\omega_1 \in A_1} \sum_{\omega_2 \in A_2} \mathbb{P}(\{(\omega_1, \omega_2)\}) = \sum_{\omega_1 \in A_1} \sum_{\omega_2 \in A_2} \mathbb{P}_1(\{\omega_1\}) \mathbb{P}_2(\{\omega_2\}) \\ &= \sum_{\omega_1 \in A_1} \mathbb{P}_1(\{\omega_1\}) \sum_{\omega_2 \in A_2} \mathbb{P}_2(\{\omega_2\}) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2); \end{aligned}$$

but we then have $\mathbb{P}(A_1 \cap \Omega_2) = \mathbb{P}_1(A_1)$, and $\mathbb{P}(\Omega_1 \cap A_2) = \mathbb{P}_2(A_2)$, whence we can interpret this as $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$ which ties in with the definition of independence given above. This extends to n independent experiments in the obvious way.

2.5 Distributions

We are considering a finite or countable sample space $\Omega = \{\omega_i\}_i$, and for each point $\omega_i \in \Omega$ let $p_i = \mathbb{P}(\{\omega_i\})$ be the probability of the event consisting of the single point ω_i . Then the sequence $\{p_i\}_i$ satisfy

$$p_i \geq 0, \quad \text{for all } i, \quad \text{and} \quad \sum_i p_i = 1, \quad (2.24)$$

because $\Omega = \bigcup_i \{\omega_i\}$. There is a one-to-one correspondence between sequences $\{p_i\}_i$ satisfying (2.24) and probability distributions as defined in this chapter through the relation $\mathbb{P}(A) = \sum_{i: \omega_i \in A} p_i$. Because of this correspondence the term **probability distribution** is also applied to sequences satisfying (2.24). We consider a number of particular distributions:

Example 2.25 *Bernoulli distribution* Consider a sample space with just two points $\Omega = \{H, T\}$, where the two points may be thought of as H = “heads”, and T = “tails”, so we are modelling a coin toss where we will take p as the probability of heads and $1 - p$ as the probability of tails. Then $p = \mathbb{P}(H)$ and $1 - p = \mathbb{P}(T)$, with $0 \leq p \leq 1$. $\square$

● **Example 2.26** *Binomial distribution* This models the number of heads obtained in n successive tosses of the coin in the previous example. Then $\Omega = \{0, 1, 2, \dots, n\}$ and the probability of k heads is

$$p_k = \mathbb{P}(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad 0 \leq k \leq n. \quad (2.27)$$

Notice that this defines a probability distribution since, by the Binomial Theorem

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

To see why this corresponds to the probability of k heads in n tosses of the coin—let $\Omega' = \{(i_1, \dots, i_n) : i_j = H \text{ or } T\}$ be the sample space where an outcome records whether

a head or a tail is obtained on each toss. Then, if the sequence $\underline{i} = (i_1, \dots, i_n)$ represents an outcome in Ω' , let $N(\underline{i})$ represent the number of indices j with $i_j = H$; that is, $N(\underline{i})$ is the number of heads in $\underline{i}$. By independence the probability of the outcome $\underline{i}$ is

$$\mathbb{P}(\{\underline{i}\}) = p^{N(\underline{i})}(1-p)^{n-N(\underline{i})}.$$

Now the number of sequences $\underline{i} \in \Omega'$ for which $N(\underline{i}) = k$, that is the number of sequences for which there are exactly k heads, is $\binom{n}{k}$ (think of choosing the k positions for the heads from the n possible positions) and hence we get (2.27). $\square$

- **Example 2.28 Poisson distribution** This distribution is often used to model the number of occurrences of some event in a specified period of time, such as the number of accidents on a particular stretch of road, for example, or the number of customers who enter a particular shop. Here the probability space is $\Omega = \{0, 1, 2, \dots\}$, the non-negative integers, and the probability of the point k is

$$p_k = \mathbb{P}(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

for some fixed $\lambda > 0$. Check that

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

Suppose that we consider customers entering a shop during an hour-long period, $(0, 1]$. Think of dividing the period into n segments, $((i-1)/n, i/n]$, for $i = 1, \dots, n$, and suppose that 1 customer enters the shop in each segment with probability p , $0 < p < 1$. Then the probability that k customers enter in the hour is the binomial probability

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

Now suppose that $n \rightarrow \infty$ and $p \rightarrow 0$ in such a way that $np \rightarrow \lambda$, then we have, for each fixed k ,

$$\lim_{n \rightarrow \infty} \left[\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \right] = \lim_{n \rightarrow \infty} \left[\left(1 - \frac{\lambda}{n}\right)^{n-k} \frac{(np)^k}{k!} \frac{n!}{n^k(n-k)!} \right] = e^{-\lambda} \frac{\lambda^k}{k!},$$

23

$$np \rightarrow \lambda \Rightarrow p \sim \frac{\lambda}{n}, \text{ infatti } \frac{p}{\frac{\lambda}{n}} = \frac{np}{\lambda} \rightarrow \frac{\lambda}{\lambda} = 1$$

// In realtà, è più semplice porre direttamente $p = \frac{\lambda}{n}$ //

because

$$\lim_{n \rightarrow \infty} \left[\frac{n!}{n^k (n-k)!} \right] = \lim_{n \rightarrow \infty} \left[1 \left(1 - \frac{1}{n} \right) \cdots \left(1 - \frac{k-1}{n} \right) \right] = 1.$$

This has proved the following result.

Theorem 2.29 (Poisson approximation to the binomial) Suppose that $n \rightarrow \infty$ and $p \rightarrow 0$ so that $np \rightarrow \lambda$, then

$$\binom{n}{k} p^k (1-p)^{n-k} \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}, \quad \text{for } k = 0, 1, 2, \dots \quad \square$$

• **Example 2.30** *Geometric distribution* This is a distribution which models the number of tosses of a coin required to obtain the first occurrence of a head, when p , $0 < p < 1$, is the probability of a head on each toss; for the probability space $\Omega = \{1, 2, \dots\}$ we have

$$p_k = \mathbb{P}(k) = p(1-p)^{k-1} \quad \text{for } k = 1, 2, \dots$$

Check that $\sum_1^\infty p_k = p/(1 - (1-p)) = 1$. Note that the term geometric distribution is also applied to the distribution on the probability space $\Omega = \{0, 1, 2, \dots\}$ with $p_k = p(1-p)^k$, for $k \geq 0$; this would be modelling the number of tails before first obtaining a head, but no confusion should arise between the slightly different usage of the term. $\square$

• **Example 2.31** *Hypergeometric distribution* Consider an urn with n_1 red balls and n_2 black balls of which n are drawn without replacement, $n \leq n_1 + n_2$. The probability that there are exactly k red balls drawn is

$$p_k = \frac{\binom{n_1}{k} \binom{n_2}{n-k}}{\binom{n_1+n_2}{n}}, \quad \text{for } \max(0, n-n_2) \leq k \leq \min(n, n_1).$$

For example, the probability that there are exactly 5 hearts in a bridge hand is

$$\frac{\binom{13}{5} \binom{39}{8}}{\binom{52}{13}}. \quad \square$$

9 January 2007

3. DISCRETE RANDOM VARIABLES

3.1 Introduction

When an experiment is conducted there may be a number of quantities associated with the outcome $\omega \in \Omega$ that may be of interest. Suppose that the experiment is choosing a male student at random from the audience of the IA Probability lecture – there are many different measurements, or attributes, of the person chosen that may be of interest: his height, his weight, his IQ, the colour of his eyes, etc. Rather than think of each of these as the outcome of a separate experiment it is more useful to view them as functions of the outcome ω . This leads to the following definition which is a central notion of probability.

Definition A random variable X , taking values in a set S , is a function $X : \Omega \rightarrow S$.

Typically, S may be a subset of the real numbers, $\mathbb{R}$, as would be the case if the height of the student was of interest; or it could be a subset of $\mathbb{R}^k$, if more than one measurement is made on the subject as would be the case, with $k = 2$, if height and weight are measured; or, S could be some arbitrary set such as $S = \{\text{Blue, Green, Brown}\}$, say, if it is the colour of the subject's eyes that are to be recorded. The most frequent situation that we will encounter is the case when $S \subseteq \mathbb{R}$, and X is then said to be a real-valued random variable. Denote by Ω_X the ^{domain} ~~range~~ _{imagined} of X , so that $\Omega_X = \{X(\omega) : \omega \in \Omega\}$. In this chapter we will assume that the sample space Ω is either a finite or a countable set, so that Ω_X is finite or countable.

For $T \subseteq S$, we denote the event $\{\omega : X(\omega) \in T\}$ as $\{X \in T\}$, so that the dependence of X on ω is suppressed in the notation. Suppose that we enumerate the points in Ω_X (equivalently the values taken on by X), so that $\Omega_X = \{x_j : j \in J\}$, then we write the event $\{\omega : X(\omega) = x_j\} = \{X = x_j\}$. If we let $p_j = \mathbb{P}(X = x_j)$, $j \in J$, then $\{p_j : j \in J\}$ is a probability distribution on the space Ω_X , and is referred to as the **probability distribution of the random variable X** . Note that it is a probability distribution on the set Ω_X , not on the underlying sample space Ω .

Example 3.1 Suppose that two standard dice are rolled so that the sample space is

$\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$, and we are interested in the sum of the numbers shown so that the random variable $X : \Omega \rightarrow \mathbb{R}$ is given by $X(i, j) = i + j$. The probability of each point in Ω is $\frac{1}{36}$ with the set of possible values taken on by X being $\Omega_X = \{2, 3, \dots, 12\}$ and, for example,

$$\mathbb{P}(X = 6) = \mathbb{P}(\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}) = \frac{5}{36}.$$

If we set $p_j = \mathbb{P}(X = j)$, for $j = 2, \dots, 12$, then the table

j	2	3	4	5	6	7	8	9	10	11	12
p_j	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

gives the full probability distribution of the random variable X . □

Terminology If the probability distribution of X is a standard distribution such as the binomial distribution (or Poisson, or geometric), we say that X is a binomial (respectively, Poisson, or geometric) random variable. We often write $X \sim \text{Bin}(n, p)$, for example, for the statement that X is binomial distribution where the parameters are n and p , or $X \sim \text{Pois}(\lambda)$ for a Poisson random variable with parameter λ .

Example 3.2 Suppose that a coin is tossed n times and a 1 is recorded whenever a head occurs and a 0 is recorded for each tail. Then $\Omega = \{(i_1, i_2, \dots, i_n) : i_j = 1 \text{ or } 0\}$. If p is the probability of a head and tosses are independent then the probability on Ω is specified by

$$\mathbb{P}(i_1, i_2, \dots, i_n) = p^{i_1 + \dots + i_n} (1 - p)^{n - i_1 - \dots - i_n}.$$

Let X denote the number of heads obtained, so that $X(i_1, \dots, i_n) = i_1 + \dots + i_n$, then X is a binomial random variable since the distribution of X is given by

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \text{for } 0 \leq k \leq n. \quad \square$$

For a function $g : S \rightarrow T$, mapping from the set S to the set T , then if X is a random variable taking values in S , $g(X)$ is the random variable taking values in T , with $g(X) : \Omega \rightarrow T$ specified by $g(X)(\omega) = g(X(\omega))$. For subsets $C \subseteq T$ we have

$\mathbb{P}(g(X) \in C) = \mathbb{P}(X \in g^{-1}(C))$ and the distribution of $g(X)$ may be obtained from that of X by observing that

$$\mathbb{P}(g(X) = y) = \mathbb{P}(X \in g^{-1}(y)) = \sum_{x \in g^{-1}(y)} \mathbb{P}(X = x).$$

A real-valued random variable which takes on just the two values 0 and 1 is known as an **indicator random variable**; suppose that the event on which it takes the value 1 is $A \subseteq \Omega$ then the random variable is denoted by I_A , so that

$$I_A(\omega) = \begin{cases} 1 & \text{for } \omega \in A, \\ 0 & \text{for } \omega \notin A, \end{cases}$$

and I_A is 1 or 0 according as the event A occurs or does not occur. The following properties of indicator random variables should be noted for events A and B :

1. $I_{A^c} = 1 - I_A$.
2. $I_{A \cap B} = I_A I_B$.
3. $I_{A \cup B} = 1 - (1 - I_A)(1 - I_B)$.

and, for events $A_1, A_2, \dots, A_n$, Properties 2 and 3 generalize to $I_{A_1 \cap A_2 \cap \dots \cap A_n} = \prod_{i=1}^n I_{A_i}$, and

$$\begin{aligned} I_{A_1 \cup A_2 \cup \dots \cup A_n} &= 1 - \prod_{i=1}^n (1 - I_{A_i}) \\ &= \sum_i I_{A_i} - \sum_{i_1 < i_2} I_{A_{i_1}} I_{A_{i_2}} + \sum_{i_1 < i_2 < i_3} I_{A_{i_1}} I_{A_{i_2}} I_{A_{i_3}} - \dots + (-1)^{n-1} I_{A_1} \dots I_{A_n} \\ &= \sum_i I_{A_i} - \sum_{i_1 < i_2} I_{A_{i_1} \cap A_{i_2}} + \sum_{i_1 < i_2 < i_3} I_{A_{i_1} \cap A_{i_2} \cap A_{i_3}} - \dots + (-1)^{n-1} I_{A_1 \cap \dots \cap A_n}. \end{aligned}$$

In the next section we see how this last relation provides an alternate proof of the inclusion-exclusion formula.

3.2 Expectation, variance and covariance

From now on, unless we indicate to the contrary, the random variables we will consider will take real values. For a non-negative random variable X , that is one for which $X(\omega) \geq 0$

for all $\omega \in \Omega$, (usually just written as $X \geq 0$), we define the **expectation** (or **expected value** or **mean value**) of X to be

$$\mathbb{E} X = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}(\{\omega\});$$

since all the terms in the sum are non-negative the sum is well defined (although it may take the value $+\infty$). Note that, since $\Omega = \bigcup_{x \in \Omega_X} \{X = x\}$, we have a more useful form for the expectation given by

$$\mathbb{E} X = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}(\{\omega\}) = \sum_{x \in \Omega_X} \sum_{\omega \in \{X=x\}} X(\omega) \mathbb{P}(\{\omega\}) = \sum_{x \in \Omega_X} x \mathbb{P}(X = x).$$

Thus the expectation is the average of the values taken on by the random variable, averaged with weights corresponding to the probabilities of the values.

Example 3.3 Suppose that $X \sim \text{Bin}(n, p)$, so that $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, for $0 \leq k \leq n$, then

NO
$$\begin{aligned} \mathbb{E} X &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = p \sum_{k=1}^n \frac{n!}{(n-k)!(k-1)!} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = np [p + (1-p)]^{n-1} = np. \end{aligned} \quad \square$$

$X = |E_1| + \dots + |E_n| \Rightarrow \mathbb{E} X = \sum \mathbb{E}(|E_i|) = \sum p = np$, $E_i = \text{"test all 'i' involucre"}$

Example 3.4 Suppose that $X \sim \text{Pois}(\lambda)$, so that $\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$, for $k = 0, 1, 2, \dots$, then

$$\mathbb{E} X = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda. \quad \square$$

Def For any random variable X denote by $X_+ = \max(X, 0)$, the positive part of X , and $X_- = \max(-X, 0)$, the negative part of X are non-negative random variables, for which so that $X = X_+ - X_-$ and $|X| = X_+ + X_-$. Provided not both $\mathbb{E} X_+ = \infty$ and $\mathbb{E} X_- = \infty$, we define the expectation of X to be

$$\mathbb{E} X = \mathbb{E} X_+ - \mathbb{E} X_- = \sum_{x \in \Omega_X} x \mathbb{P}(X = x);$$

if both $\mathbb{E}X_+$ and $\mathbb{E}X_-$ are infinite then the expectation of X is not defined. In the following, when we write $\mathbb{E}X$ for a random variable X , it may be assumed that the expectation of X is well defined.

Properties of $\mathbb{E}X$

1. If $X \geq 0$, then $\mathbb{E}X \geq 0$, and $\mathbb{E}X = 0$ implies that $\mathbb{P}(X = 0) = 1$.
2. If c is a constant then $\mathbb{E}(cX) = c\mathbb{E}X$, and $\mathbb{E}c = c$.
3. For random variables X and Y , $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$.

Properties 2 and 3 show the important property that the operator $\mathbb{E}(\cdot)$ is a linear operator and they generalize, by induction, to the case of random variables $X_1, \dots, X_n$ and constants $c_1, \dots, c_n$ so that $\mathbb{E}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i \mathbb{E}X_i$.

$$4. \mathbb{E}g(X) = \sum_{x \in \Omega_X} g(x)\mathbb{P}(X = x).$$

Handwritten note: "By definition"

To see this, let $Y = g(X)$, then

$$\begin{aligned} \mathbb{E}g(X) = \mathbb{E}Y &= \sum_{y \in \Omega_Y} y\mathbb{P}(Y = y) = \sum_{y \in \Omega_Y} y \left(\sum_{x \in g^{-1}(y)} \mathbb{P}(X = x) \right) \\ &= \sum_{y \in \Omega_Y} \sum_{x \in g^{-1}(y)} y\mathbb{P}(X = x) = \sum_{y \in \Omega_Y} \sum_{x \in g^{-1}(y)} g(x)\mathbb{P}(X = x) \\ &= \sum_{x \in \Omega_X} g(x)\mathbb{P}(X = x). \end{aligned}$$

5. For the indicator of any event $A \subseteq \Omega$ we have $\mathbb{E}I_A = \mathbb{P}(A)$.
6. If $X \geq 0$ and X takes integer values, then $\mathbb{E}X = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n)$.

Proof. We have

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

after interchanging the order of the summations. □

Terminology For a random variable X , the expected values of powers of X are known as **moments** of X ; thus $\mathbb{E}(X^r)$ (assuming it is well defined) is the r th moment of X and $\mathbb{E}(|X|^r)$ is the r th absolute moment of X .

Example 3.5 *Another proof of inclusion-exclusion.* For events $A_1, \dots, A_n$, use the previous expression for the product of indicators to calculate

$$\begin{aligned}\mathbb{P}(A_1 \cup \dots \cup A_n) &= \mathbb{E}(I_{A_1 \cup A_2 \cup \dots \cup A_n}) = \mathbb{E}\left(1 - \prod_{i=1}^n (1 - I_{A_i})\right) \\ &= \mathbb{E}\left(\sum_i I_{A_i} - \sum_{i_1 < i_2} I_{A_{i_1} \cap A_{i_2}} + \sum_{i_1 < i_2 < i_3} I_{A_{i_1} \cap A_{i_2} \cap A_{i_3}} - \dots + (-1)^{n-1} I_{A_1 \cap \dots \cap A_n}\right)\end{aligned}$$

then using the linearity of the expectation, this

$$\begin{aligned}&= \sum_i \mathbb{E}(I_{A_i}) - \sum_{i_1 < i_2} \mathbb{E}(I_{A_{i_1} \cap A_{i_2}}) + \dots + (-1)^{n-1} \mathbb{E}(I_{A_1 \cap \dots \cap A_n}) \\ &= \sum_i \mathbb{P}(A_i) - \sum_{i_1 < i_2} \mathbb{P}(A_{i_1} \cap A_{i_2}) + \dots + (-1)^{n-1} \mathbb{P}(A_1 \cap \dots \cap A_n),\end{aligned}$$

which is the required expression for the inclusion-exclusion formula. $\square$

For any random variable, X with finite mean, the **variance** is defined to be

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2,$$

and it is a measure of how much the distribution of X is spread out around the mean; the smaller the distribution the more the distribution of X is concentrated close to $\mathbb{E}X$. The quantity $\sqrt{\text{Var}(X)}$ is known as the **standard deviation** of X . When we use the notation $\text{Var}(X)$ we will assume implicitly that it is a finite quantity.

Properties of $\text{Var}(X)$

$$1. \text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

Proof. We have, using Properties 2 and 3 of the expectation,

$$\begin{aligned}\mathbb{E}(X - \mathbb{E}X)^2 &= \mathbb{E}(X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2) \\ &= \mathbb{E}X^2 - 2\mathbb{E}X\mathbb{E}X + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.\end{aligned}$$

$\square$

2. If c is a constant, $\text{Var}(cX) = c^2 \text{Var}(X)$.
3. If c is a constant, $\text{Var}(X + c) = \text{Var}(X)$.
4. $\text{Var}(X) \geq 0$, and $\text{Var}(X) = 0$ if and only if $\mathbb{P}(X = c) = 1$, for some constant c .
5. The expression $\mathbb{E}(X - c)^2$ is minimized over constants c when $c = \mathbb{E}X$, so that $\mathbb{E}(X - c)^2 \geq \text{Var}(X)$, for all c , with equality when $c = \mathbb{E}X$.

Proof. Expand out the expression

$$\mathbb{E} (X - c)^2 = \mathbb{E} (X^2 - 2cX + c^2) = \mathbb{E} X^2 - 2c\mathbb{E} X + c^2,$$

and minimize the right-hand side in c to see that the minimum occurs at $c = \mathbb{E} X$. $\square$

• **Example 3.6** For $X \sim \text{Bin}(n, p)$, we have

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

then it follows that

$$\mathbb{E} X^2 = \mathbb{E} (X(X-1)) + \mathbb{E} X = n(n-1)p^2 + np,$$

since we had seen that $\mathbb{E} X = np$; hence $\text{Var}(X) = \mathbb{E} X^2 - (\mathbb{E} X)^2 = np(1-p)$. $\square$



• **Example 3.7** Suppose that $X \sim \text{Pois}(\lambda)$, then a similar calculation to that in the previous Example gives

$$\mathbb{E} (X(X-1)) = \sum_{k=0}^{\infty} k(k-1) e^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=2}^{\infty} e^{-\lambda} \frac{\lambda^k}{(k-2)!} = \lambda^2 e^{-\lambda} \sum_{r=0}^{\infty} \frac{\lambda^r}{r!} = \lambda^2;$$

recalling that in this case $\mathbb{E} X = \lambda$, we have $\mathbb{E} X^2 = \lambda^2 + \lambda$, so that $\text{Var}(X) = \lambda$, showing that for a Poisson random variable the mean is the same as the variance. $\square$

Example 3.8 *Use of indicators* Return to the situation, considered in Chapter 2, where n students leave their n coats outside the lecture room and when they leave they pick up their coats at random. Let N be the number of students who get their own coat, then $N = \sum_{i=1}^n I_{A_i}$, where A_i is the event that student i obtains his own coat. It follows that

$$\begin{aligned} \mathbb{E} N &= \mathbb{E} \left(\sum_{i=1}^n I_{A_i} \right) = \sum_{i=1}^n \mathbb{E} (I_{A_i}) = \sum_{i=1}^n \mathbb{P}(A_i) = \sum_{i=1}^n \frac{1}{n} = 1, \quad \text{and} \\ \mathbb{E} N^2 &= \mathbb{E} \left(\sum_{i=1}^n I_{A_i} \right)^2 = \mathbb{E} \left(\sum_{i=1}^n (I_{A_i})^2 + \sum_i \sum_{j \neq i} I_{A_i} I_{A_j} \right); \end{aligned}$$

since $I_{A_i} I_{A_j} = I_{A_i \cap A_j}$ we have $(I_{A_i})^2 = I_{A_i}$, and we see that



$X = |E_1| + \dots + |E_m|$; E_i indep $\Rightarrow \text{Var } X = \sum \text{Var } |E_i| \stackrel{?}{=} np$
 $\text{Var } |E_i| = \mathbb{E} |E_i|^2 - (\mathbb{E} |E_i|)^2 = p - p^2 = pq$

$$\begin{aligned}\mathbb{E} N^2 &= \mathbb{E} \left(\sum_{i=1}^n I_{A_i} + \sum_i \sum_{j \neq i} I_{A_i \cap A_j} \right) = \sum_{i=1}^n \mathbb{P}(A_i) + \sum_i \sum_{j \neq i} \mathbb{P}(A_i \cap A_j) \\ &= \sum_{i=1}^n \frac{1}{n} + \sum_i \sum_{j \neq i} \frac{1}{n(n-1)} = n \times \frac{1}{n} + n(n-1) \times \frac{1}{n(n-1)} = 1 + 1 = 2.\end{aligned}$$

That gives $\text{Var}(N) = \mathbb{E} N^2 - (\mathbb{E} N)^2 = 1$. The fact that the mean and the variance are both the same might suggest that the distribution of the random variable N is close to being Poisson (with mean $\lambda = 1$) as is indeed the case when n is large. If we let $p_n = \mathbb{P}(N = 0)$, the probability that when there are n students none of them gets his own coat, then we have seen previously (using inclusion-exclusion) that

$$p_n = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} \rightarrow e^{-1}, \quad \text{as } n \rightarrow \infty;$$

take $p_0 = 1$. The probability that exactly k students get their own coats is

$$\mathbb{P}(N = k) = \binom{n}{k} \frac{1}{n!} ((n-k)!) p_{n-k} = \frac{1}{k!} p_{n-k} \rightarrow \frac{1}{k!} e^{-1}, \quad \text{as } n \rightarrow \infty,$$

showing that the distribution of N is approximately Poisson. □

• **Theorem 3.9** (Cauchy–Schwarz inequality) *For any random variables X and Y ,*

$$(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2) \mathbb{E}(Y^2);$$

if $\mathbb{E}(Y^2) > 0$, equality occurs if and only if $X = aY$ for some constant $a \in \mathbb{R}$.

Proof. For any $a \in \mathbb{R}$, observe that $\mathbb{E}(X - aY)^2 \geq 0$, so that

$$0 \leq \mathbb{E}(X^2 - 2aXY + a^2Y^2) = \mathbb{E}(X^2) - 2a\mathbb{E}(XY) + a^2\mathbb{E}(Y^2),$$

showing that the quadratic in a on the right-hand side has at most one real root, whence the discriminant $4((\mathbb{E}(XY))^2 - \mathbb{E}(X^2)\mathbb{E}(Y^2)) \leq 0$, giving the inequality. There is clearly equality if $X = aY$ for some $a \in \mathbb{R}$, whereas if $\mathbb{E}(Y^2) > 0$ and the discriminant is 0 then the quadratic is 0 for $a = \mathbb{E}(XY)/\mathbb{E}(Y^2)$, and for that value of a , $\mathbb{E}(X - aY)^2 = 0$ and so $X = aY$. Of course, if $\mathbb{E}(Y^2) = 0$ then $Y = 0$ and equality occurs. □

For two random variable X and Y , we define the **covariance** between X and Y as

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

We shall see that this is a measure of the dependence between the random variables X and Y .

Properties of $\text{Cov}(X, Y)$

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
2. $\text{Cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)$.

Proof. We have

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}(XY - X(\mathbb{E}Y) - Y(\mathbb{E}X) + (\mathbb{E}X)(\mathbb{E}Y)) \\ &= \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) - (\mathbb{E}X)(\mathbb{E}Y) + (\mathbb{E}X)(\mathbb{E}Y) \\ &= \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y). \end{aligned} \quad \square$$

3. $\text{Cov}(X, X) = \text{Var}(X)$.
4. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.

Proof. We have

$$\begin{aligned} \text{Var}(X + Y) &= \mathbb{E}(X + Y - \mathbb{E}X - \mathbb{E}Y)^2 = \mathbb{E}((X - \mathbb{E}X) + (Y - \mathbb{E}Y))^2 \\ &= \mathbb{E}((X - \mathbb{E}X)^2 + (Y - \mathbb{E}Y)^2 + 2(X - \mathbb{E}X)(Y - \mathbb{E}Y)) \\ &= \mathbb{E}(X - \mathbb{E}X)^2 + \mathbb{E}(Y - \mathbb{E}Y)^2 + 2\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y). \end{aligned} \quad \square$$

5. If c is a constant, $\text{Cov}(X, c) = 0$.
6. If c is a constant, $\text{Cov}(X + c, Y) = \text{Cov}(X, Y)$.
7. If c is a constant, $\text{Cov}(cX, Y) = c\text{Cov}(X, Y)$.
8. $\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y)$.

These last two generalize to the case of random variables $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ and constants $c_1, \dots, c_n$ and $d_1, \dots, d_n$ to give, by induction,

$$\text{Cov}\left(\sum_{i=1}^n c_i X_i, \sum_{j=1}^n d_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^n c_i d_j \text{Cov}(X_i, Y_j). \quad (3.10)$$

Using the fact that $\text{Var}(X) = \text{Cov}(X, X)$, we see that a special case of this is

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i=1}^n \sum_{j \neq i} \text{Cov}(X_i, X_j), \quad (3.11)$$

for any random variables $X_1, \dots, X_n$.

The **correlation coefficient** (or just the **correlation**) between random variables X and Y with $\text{Var}(X) > 0$ and $\text{Var}(Y) > 0$ is

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

Notice that by the Cauchy-Schwarz inequality

$$|\text{Corr}(X, Y)| \leq 1, \quad \text{for all } X \text{ and } Y;$$

this follows by applying the inequality to the random variables $\bar{X} = X - \mathbb{E}X$ and $\bar{Y} = Y - \mathbb{E}Y$. It may be further seen that $|\text{Corr}(X, Y)| = 1$ if and only if $X = aY + b$ for some constants a and b . One property of correlation that we should note is that for constants a, b, c and d with $ac \neq 0$, we have

$$\text{Corr}(aX + b, cY + d) = \begin{cases} \text{Corr}(X, Y) & \text{when } ac > 0, \\ -\text{Corr}(X, Y) & \text{when } ac < 0. \end{cases}$$

This follows easily from the definition of correlation and the properties of the covariance and variance; notice that when $ac = 0$, $\text{Cov}(aX + b, cY + d) = 0$, and the correlation is not defined because at least one of $\text{Var}(aX + b) = 0$ or $\text{Var}(cY + d) = 0$.

Notice that one consequence of this fact is that the correlation between two random variables is scale invariant – if we multiply the observation of X and Y by positive constants we do not alter the correlation.

3.3 Independence

Discrete random variables $X_1, X_2, \dots, X_n$ are **independent** if, for all choices of $x_i \in \Omega_{X_i}$, $1 \leq i \leq n$, we have

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n \mathbb{P}(X_i = x_i). \quad (3.12)$$

Prop 1

Notice that $X_1, X_2, \dots, X_n$ are independent if and only if, for all choices of subsets $S_i \subseteq \Omega_{X_i}$, $1 \leq i \leq n$, we have

$$\mathbb{P}(X_1 \in S_1, X_2 \in S_2, \dots, X_n \in S_n) = \prod_{i=1}^n \mathbb{P}(X_i \in S_i). \quad (3.13)$$

To see this, if (3.13) holds, take $S_i = \{x_i\}$ for each i and we see that (3.12) is true; conversely, the left hand side of (3.13) is

$$\sum_{x_1 \in S_1} \sum_{x_2 \in S_2} \cdots \sum_{x_n \in S_n} \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

and we see that, if (3.12) holds, then this expression is

$$\sum_{x_1 \in S_1} \sum_{x_2 \in S_2} \cdots \sum_{x_n \in S_n} \prod_{i=1}^n \mathbb{P}(X_i = x_i) = \prod_{i=1}^n \left(\sum_{x_i \in S_i} \mathbb{P}(X_i = x_i) \right) = \prod_{i=1}^n \mathbb{P}(X_i \in S_i),$$

which gives (3.13).

Notice that events $A_1, \dots, A_n$ are independent, as defined in the previous chapter, if and only if their indicator random variables $I_{A_1}, \dots, I_{A_n}$ are independent random variables. Observe also that if random variables are independent then they are independent in pairs (this follows by taking $S_i = \Omega_{X_i}$ for all but two of the subsets S_i in (3.13)) – they are said to be pairwise independent; a similar argument shows that if any collection of random variables is independent then any sub-collection of them is independent. By considering indicators, the example from the last chapter shows that pairwise independence of random variables does not imply independence in general.

Properties of independent random variables

1. If $X_1, \dots, X_n$ are independent random variables and $g_i : \mathbb{R} \rightarrow \mathbb{R}$, $1 \leq i \leq n$, are functions then $g_1(X_1), \dots, g_n(X_n)$ are independent random variables.

Proof. For $y_i \in \Omega_{g_i(X_i)}$, $1 \leq i \leq n$, we have

$$\begin{aligned} \mathbb{P}(g_1(X_1) = y_1, \dots, g_n(X_n) = y_n) &= \mathbb{P}(X_1 \in g_1^{-1}(y_1), \dots, X_n \in g_n^{-1}(y_n)) \\ &= \prod_{i=1}^n \mathbb{P}(X_i \in g_i^{-1}(y_i)) = \prod_{i=1}^n \mathbb{P}(g_i(X_i) = y_i) \end{aligned}$$

↑
prop 1 35

after using (3.13), showing that the random variables $g_1(X_1), \dots, g_n(X_n)$ are independent. $\square$

2. If $X_1, \dots, X_n$ are independent random variables, then

$$\mathbb{E} \left(\prod_{i=1}^n X_i \right) = \prod_{i=1}^n \mathbb{E} (X_i);$$

that is, the expectation of the product of independent random variables is the product of their expectations.

Proof. In a similar way to the previous proof, we may represent the event

$$\left(\prod_{i=1}^n X_i = y \right) = \bigcup (X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

as a disjoint union of events over values of $x_1, \dots, x_n$ with $\prod_i x_i = y$. Then

$$\begin{aligned} \mathbb{E} \left(\prod_{i=1}^n X_i \right) &= \sum_y y \mathbb{P} \left(\prod_{i=1}^n X_i = y \right) \stackrel{\substack{\text{la } x_i \text{ stanno nelle immagini, quindi } (X_1=x_1, \dots, X_n=x_n) \text{ e } (X_1=x'_1, \dots, X_n=x'_n) = \emptyset}}}{=} \sum_y y \sum_{x_i: \prod_i x_i = y} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \\ &= \sum_y \sum_{x_i: \prod_i x_i = y} y \prod_{i=1}^n \mathbb{P}(X_i = x_i), \quad \text{by independence} \\ &= \sum_{x_1, \dots, x_n} \prod_{i=1}^n (x_i \mathbb{P}(X_i = x_i)) \stackrel{*}{=} \prod_{i=1}^n \left(\sum_{x_i} x_i \mathbb{P}(X_i = x_i) \right) = \prod_{i=1}^n \mathbb{E}(X_i), \end{aligned}$$

as required. $\square$

3. If X and Y are independent random variables then $\text{Cov}(X, Y) = 0$ (and hence $\text{Corr}(X, Y) = 0$). The converse is not true in general (see Example 3.14 below): that is, $\text{Cov}(X, Y) = 0$ does not imply that X and Y are independent.

Proof. Property 1 shows that $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$ are independent random variables and then by Property 2,

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) = \mathbb{E}(X - \mathbb{E}X) \mathbb{E}(Y - \mathbb{E}Y) = 0$$

since $\mathbb{E}(X - \mathbb{E}X) = \mathbb{E}(X) - \mathbb{E}(X) = 0$ (and similarly $\mathbb{E}(Y - \mathbb{E}Y) = 0$). $\square$

$$\begin{aligned} * \quad x_1 x_2'' + x_1 x_2' + x_1' x_2 + x_1' x_2' &= (x_1 + x_1')(x_2 + x_2') \\ \text{in generale} \quad \sum_{\substack{x_1 \in \{x_1, x_1', x_1'', \dots\} \\ x_2 \in \{x_2, x_2', x_2'', \dots\}}} \prod_{i=1}^2 x_i &= \prod_{i=1}^2 \sum_{x_i \in \{x_i, x_i', x_i'', \dots\}} x_i \end{aligned}$$

4. If $X_1, \dots, X_n$ are independent random variables then

$$\text{Var} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{Var} (X_i);$$

that is, the variance of the sum of independent random variables is the sum of their variances.

Proof. Use Property 3 to see that for $j \neq i$, $\text{Cov} (X_i, X_j) = 0$ and the result follows from the relation (3.11). $\square$

5. If $X_1, \dots, X_n$ are independent random variables then the conditional probability

$$\mathbb{P}(X_1 = x_1, \dots, X_{n-1} = x_{n-1} \mid X_n = x_n) = \mathbb{P}(X_1 = x_1, \dots, X_{n-1} = x_{n-1}),$$

for all choices of $x_i \in \Omega_{X_i}$, $1 \leq i \leq n$.

Proof. We have the conditional probability on the left-hand side is

$$\frac{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)}{\mathbb{P}(X_n = x_n)} = \frac{\prod_{i=1}^n \mathbb{P}(X_i = x_i)}{\mathbb{P}(X_n = x_n)} = \prod_{i=1}^{n-1} \mathbb{P}(X_i = x_i),$$

which equals the right-hand side, again by independence. $\square$

Terminology Random variables with the same distribution are usually said to be **identically distributed**, and if they are also independent they are **i.i.d.** (independent and identically distributed). If $X_1, \dots, X_n$ are i.i.d. then, from Property 4,

$$\text{Var} \left(\frac{X_1 + \dots + X_n}{n} \right) = \frac{\text{Var}(X_1)}{n}.$$

• **Example 3.14** *Covariance equal to 0 does not imply independence* Suppose that X is a random variable with distribution determined by

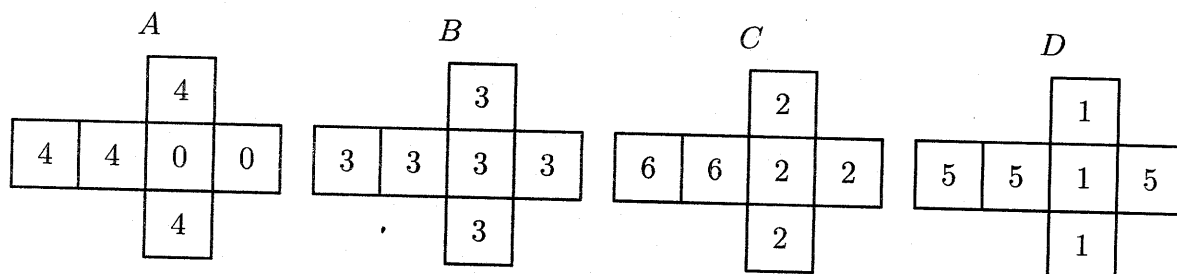
x	2	1	-1	-2
$\mathbb{P}(X = x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

and let $Y = X^2$. Then $\mathbb{E} X = 0$ and $\mathbb{E}(X^3) = 0$ so that $\text{Cov}(X, Y) = \mathbb{E}(X^3) = 0$, but

$$\mathbb{P}(X = 2, Y = 4) = \frac{1}{4} \neq \mathbb{P}(X = 2)\mathbb{P}(Y = 4) = \frac{1}{4} \times \frac{1}{2},$$

so that X and Y are not independent. □

Example 3.15 *Efron's dice* An interesting example showing that odds are not transitive is given by a set of 4 dice with the following faces:



If each of the dice is rolled with respective outcomes A, B, C and D then

$$\mathbb{P}(A > B) = \mathbb{P}(B > C) = \mathbb{P}(C > D) = \mathbb{P}(D > A) = \frac{2}{3}.$$
□

3.4 Probability generating functions

Consider a random variable, X , taking values in the non-negative integers $0, 1, 2, \dots$ with distribution determined by $p_r = \mathbb{P}(X = r)$, $r = 0, 1, 2, \dots$. The **probability generating function** (p.g.f.) of X is defined to be

$$p(z) = \mathbb{E}(z^X) = \sum_{r=0}^{\infty} p_r z^r, \quad \text{for } 0 \leq z \leq 1. \quad // 0^0 = 1 //$$

Since the terms in the sum are all non-negative and $0 \leq \sum_r p_r z^r \leq \sum_r p_r = 1$, the probability generating function is well defined and takes values in $[0, 1]$. Its importance stems from the following result.

Theorem 3.16 *The probability generating function of X , $p(z)$, $0 \leq z \leq 1$, determines the probability distribution of X uniquely.*

Proof. Suppose that $p(z) = \sum_{r=0}^{\infty} p_r z^r = \sum_{r=0}^{\infty} q_r z^r$, for all $0 \leq z \leq 1$, where $p_r \geq 0$, and $q_r \geq 0$ for each r , and $\sum_{r=0}^{\infty} p_r = 1 = \sum_{r=0}^{\infty} q_r$. We will show by induction on n that $p_n = q_n$ for all n . First see, by setting $z = 0$, that $p_0 = q_0$. Now assume that $p_i = q_i$ for $0 \leq i \leq n$, then for $0 < z \leq 1$

$$\sum_{r=n+1}^{\infty} p_r z^r = \sum_{r=n+1}^{\infty} q_r z^r.$$

Coroll. X, Y hanno la stessa p.g.f. $\Rightarrow$ hanno la stessa distrib. di probabilità.

Divide through both sides by z^{n+1} and let $z \downarrow 0$ to see that $p_{n+1} = q_{n+1}$ to complete the induction. $\square$

In addition to determining the distribution uniquely, the probability generating function may be used to compute moments of the random variable by evaluating derivatives of the function.

Theorem 3.17 *Let X be a random variable with probability generating function $p(z)$, then the mean of X is*

$$\mathbb{E} X = \lim_{z \uparrow 1} p'(z) = p'(1-).$$

Proof. First assume that $\mathbb{E} X < \infty$. For $0 < z < 1$,

$$p'(z) = \sum_{r=1}^{\infty} r p_r z^{r-1} \leq \sum_{r=1}^{\infty} r p_r = \mathbb{E} X.$$

We see that $p'(z)$ is non-decreasing in z so that $\lim_{z \uparrow 1} p'(z) \leq \mathbb{E} X$. Take $\epsilon > 0$, and choose

N so that $\sum_{r=1}^N r p_r \geq \mathbb{E} X - \epsilon$. Then

$$\lim_{z \uparrow 1} p'(z) \geq \lim_{z \uparrow 1} \sum_{r=1}^N r p_r z^{r-1} = \sum_{r=1}^N r p_r \geq \mathbb{E} X - \epsilon;$$

this is true for each $\epsilon > 0$, whence $\lim_{z \uparrow 1} p'(z) \geq \mathbb{E} X$ and it follows that $\lim_{z \uparrow 1} p'(z) = \mathbb{E} X$. If

$\mathbb{E} X = \infty$, then for any $M > 0$ choose N so that $\sum_{r=1}^N r p_r \geq M$, and, as above, see that

$$\lim_{z \uparrow 1} p'(z) \geq \lim_{z \uparrow 1} \sum_{r=1}^N r p_r z^{r-1} = \sum_{r=1}^N r p_r \geq M;$$

this is true for any M , whence $\lim_{z \uparrow 1} p'(z) = \infty$. $\square$

Note By considering the second derivative of $p(z)$, a similar argument to that of Theorem 3.17 may be used to show that

$$p''(1-) = \lim_{z \uparrow 1} p''(z) = \lim_{z \uparrow 1} \sum_{r=1}^{\infty} r(r-1) p_r z^{r-2} = \mathbb{E} (X(X-1)),$$

and by considering the k th derivative, $k \geq 1$, we have

$$\begin{aligned} p^{(k)}(1-) &= \lim_{z \uparrow 1} p^{(k)}(z) = \lim_{z \uparrow 1} \sum_{r=1}^{\infty} r(r-1) \cdots (r-k+1) p_r z^{r-2} \\ &= \mathbb{E}(X(X-1) \cdots (X-k+1)). \end{aligned}$$

In particular, $\text{Var}(X) = p''(1-) + p'(1-) - (p'(1-))^2$.

• **Example 3.18** *Geometric distribution* Let X be a random variable with probability distribution given by $\mathbb{P}(X=r) = p(1-p)^r = pq^r$, $r = 0, 1, 2, \dots$, where $0 < p = 1 - q < 1$. Then X may be thought of as the number of tails obtained before getting the first head when successively tossing a coin with probability p of heads on each toss. The probability generating function of X is

$$p(z) = \mathbb{E}(z^X) = \sum_{r=0}^{\infty} p_r z^r = \sum_{r=0}^{\infty} pq^r z^r = \frac{p}{1-qz}.$$

We have $p'(z) = pq/(1-qz)^2$, so that $\mathbb{E}X = p'(1) = q/p$. Also, $p''(z) = 2pq^2/(1-qz)^3$, so that $\mathbb{E}(X(X-1)) = 2q^2/p^2$, from which we deduce that

$$\mathbb{E}(X^2) = \frac{2q^2}{p^2} + \frac{q}{p} \quad \text{and} \quad \text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \frac{q}{p^2}. \quad \square$$

Note The term geometric distribution is often also given to the situation where $\mathbb{P}(X=r) = pq^{r-1}$, $r = 1, 2, \dots$ for $0 < p = 1 - q < 1$. Here, X would be the number of tosses required to achieve the first head where the probability of heads is p . This just corresponds to replacing X in Example 3.18 by $X+1$, so the probability generating function becomes $pz/(1-qz)$, the mean is $1/p$ and the variance is unchanged at q/p^2 . $\square$

Another use for probability generating functions is that they provide an easy way of dealing with sums of independent random variables. Suppose that $X_1, \dots, X_n$ are independent random variables with probability generating functions $p_1(z), \dots, p_n(z)$ respectively. Then, since $z^{X_1}, \dots, z^{X_n}$ are independent, we have that the probability generating function of $X_1 + \dots + X_n$ is

$$\mathbb{E}(z^{X_1 + \dots + X_n}) = \prod_{i=1}^n \mathbb{E}(z^{X_i}) = \prod_{i=1}^n p_i(z).$$

In the special case when $X_1, \dots, X_n$ are i.i.d. with common probability generating function $p(z)$ we have

$$\mathbb{E}(z^{X_1 + \dots + X_n}) = (p(z))^n.$$

- **Example 3.19** *Sums of Binomial random variables* Consider independent random variables $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, where $0 < p = 1 - q < 1$. The probability generating function of X is

$$\mathbb{E}(z^X) = \sum_{r=0}^n \binom{n}{r} p^r q^{n-r} z^r = (pz + q)^n,$$

so that the probability generating function of Y is $(pz + q)^m$. It follows that the probability generating function of $X + Y$ is the product of the two generating functions and is therefore $(pz + q)^{m+n}$. From Theorem 3.16 we conclude that $X + Y \sim \text{Bin}(n + m, p)$. The probabilistic interpretation is immediate, of course; X is the number of heads in n tosses of a coin with probability p of heads and Y is the number of ~~tosses~~ ^{heads} in m (independent) tosses of the coin, so that $X + Y$ is the number of heads in $n + m$ tosses. This of course generalizes, by induction, to the case of independent random variables $X_1, \dots, X_k$ with $X_i \sim \text{Bin}(n_i, p)$, to give $X_1 + \dots + X_k \sim \text{Bin}\left(\sum_1^k n_i, p\right)$. $\square$

- **Example 3.20** *Sums of Poisson random variables* Consider independent random variables $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$, where $\lambda > 0$ and $\mu > 0$. The probability generating function of X is

$$\mathbb{E}(z^X) = \sum_{r=0}^{\infty} z^r e^{-\lambda} \frac{\lambda^r}{r!} = e^{-\lambda(1-z)}.$$

The probability generating function is the same expression with μ replacing λ and the probability generating function of $X + Y$ is

$$e^{-\lambda(1-z)} e^{-\mu(1-z)} = e^{-(\lambda+\mu)(1-z)};$$

from Theorem 3.16 we conclude that $X + Y \sim \text{Pois}(\lambda + \mu)$; for an alternative argument see Example 3.22 below. $\square$

Example 3.21 *Negative binomial distribution* Consider a random variable X which has distribution given by

$$\mathbb{P}(X = r) = \binom{r-1}{n-1} p^n (1-p)^{r-n}, \quad \text{for } r = n, n+1, \dots,$$

where $0 < p = 1 - q < 1$, and $n \geq 1$. Here, X represents the number of tosses of a coin to get n heads for the first time, where the probability of heads is p . The probability generating function of X is

$$\mathbb{E}(z^X) = \sum_{r=n}^{\infty} z^r \binom{r-1}{n-1} p^n q^{r-n} = (pz)^n \sum_{r=n}^{\infty} \binom{r-1}{n-1} (qz)^{r-n} = (pz/(1-qz))^n.$$

From the note following Example 3.18 we see that X may be represented as the sum $X_1 + \dots + X_n$ of n i.i.d. random variables each with the same geometric distribution

$$\mathbb{P}(X_1 = r) = pq^{r-1}, \quad \text{for } r = 1, 2, \dots$$

The distribution of X is usually referred to as the negative binomial distribution. $\square$

3.5 Conditional distributions

The **joint distribution** of random variables $X_1, \dots, X_n$, is given by

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \quad \text{for } x_1 \in \Omega_{X_1}, \dots, x_n \in \Omega_{X_n},$$

and it is a probability distribution on $\Omega_{X_1} \times \dots \times \Omega_{X_n}$, and the **marginal distribution** of X_i is

$$\mathbb{P}(X_i = x_i) = \sum \mathbb{P}(X_1 = x_1, \dots, X_n = x_n),$$

where the summation is over $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$; this identity is a consequence of the law of total probability. Now consider the case $n = 2$ and (to avoid unnecessary subscripts) consider the random variables X and Y . The **conditional distribution** of X , given $Y = y$, is a probability distribution on Ω_X given by

$$\mathbb{P}(X = x | Y = y) \quad \text{for } x \in \Omega_X,$$

42

(*) $\{X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = x_i, X_{i+1} = x_{i+1}, \dots, X_n = x_n\}$ $\xrightarrow{\text{una part. di } \Omega}$
 Quindi $\mathbb{P}(X_i = x_i) = \mathbb{P}(X_i = x_i \cap \bigcup_{\Omega} \dots) = \sum \mathbb{P}(X_i = x_i \cap \dots) = \sum \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$

where, of course, $\mathbb{P}(X = x | Y = y) = \mathbb{P}(X = x, Y = y) / \mathbb{P}(Y = y)$. Again, by the law of total probability

$$\mathbb{P}(X = x) = \sum_{y \in \Omega_Y} \mathbb{P}(X = x, Y = y) = \sum_{y \in \Omega_Y} \mathbb{P}(X = x | Y = y) \mathbb{P}(Y = y).$$

• **Example 3.22** *Sum of two independent random variables* Suppose that X and Y are independent random variables, then we may express the distribution of their sum as follows

$$\begin{aligned} \mathbb{P}(X + Y = z) &= \sum_{y \in \Omega_Y} \mathbb{P}(X + Y = z | Y = y) \mathbb{P}(Y = y) = \sum_{y \in \Omega_Y} \mathbb{P}(X = z - y) \mathbb{P}(Y = y) \\ &= \sum_{x \in \Omega_X} \mathbb{P}(X = x) \mathbb{P}(Y = z - x), \end{aligned} \quad (3.23)$$

where the last expression is obtained if we condition on X initially instead of Y . This procedure gives the **convolution** of the distributions of X and Y . For example, if $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$,

$$\begin{aligned} \mathbb{P}(X + Y = n) &= \sum_{r=0}^{\infty} \mathbb{P}(X = n - r) \mathbb{P}(Y = r) \\ &= \sum_{r=0}^n e^{-\lambda} \frac{\lambda^{n-r}}{r!} e^{-\mu} \frac{\mu^r}{(n-r)!}, \quad \text{since } \mathbb{P}(X = k) = 0, \text{ for } k < 0, \\ &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{r=0}^n \binom{n}{r} \lambda^{n-r} \mu^r = e^{-(\lambda+\mu)} \frac{(\lambda + \mu)^n}{n!}, \end{aligned}$$

so that $X + Y \sim \text{Pois}(\lambda + \mu)$, as seen in Example 3.20 previously using generating functions. $\square$

The **conditional expectation of X given $Y = y$** is

$$\mathbb{E}(X | Y = y) = \sum_{x \in \Omega_X} x \mathbb{P}(X = x | Y = y) = \sum_{\omega: Y(\omega) = y} X(\omega) \mathbb{P}(\{\omega\}) / \mathbb{P}(Y = y).$$

Note that $\mathbb{E}(X | Y = y)$ is a function of y , $g(y)$ say, then the random variable $g(Y)$ is known as the **conditional expectation of X given Y** and is written $\mathbb{E}(X | Y)$. It is important to emphasize that $\mathbb{E}(X | Y)$ is a random variable and it is a function of Y , in contrast to $\mathbb{E}(X | Y = y)$, which is a real number.

$$\mathbb{E}(X|Y): \Omega_Y \rightarrow \mathbb{R} \\ y \mapsto \mathbb{E}(X|Y=y)$$

(Bisio)

• **Example 3.24** Consider tossing a coin n times where the probability of a head is p , $0 < p = 1 - q < 1$, and let $X_i = 1$ if the i th toss produces a head and $X_i = 0$, otherwise. Let $Y = X_1 + \cdots + X_n$ denote the total number of heads so that $Y \sim \text{Bin}(n, p)$. Then, for $r \geq 1$,

$$\begin{aligned} \mathbb{P}(X_1 = 1 \mid Y = r) &= \frac{\mathbb{P}(X_1 = 1, Y = r)}{\mathbb{P}(Y = r)} = \frac{\mathbb{P}(X_1 = 1, X_1 + \cdots + X_n = r)}{\mathbb{P}(Y = r)} \\ &= \frac{\mathbb{P}(X_1 = 1, X_2 + \cdots + X_n = r - 1)}{\mathbb{P}(Y = r)}, \end{aligned}$$

then by independence and the fact that $X_2 + \cdots + X_n \sim \text{Bin}(n - 1, p)$ this

$$\begin{aligned} &= \frac{\mathbb{P}(X_1 = 1) \mathbb{P}(X_2 + \cdots + X_n = r - 1)}{\mathbb{P}(Y = r)} \\ &= \frac{p \binom{n-1}{r-1} p^{r-1} q^{n-r}}{\binom{n}{r} p^r q^{n-r}} = \frac{r}{n}; \end{aligned}$$

we may see also that $\mathbb{P}(X_1 = 1 \mid Y = 0) = 0$. Then

$$\mathbb{E}(X_1 \mid Y = r) = 1 \times \mathbb{P}(X_1 = 1 \mid Y = r) + 0 \times \mathbb{P}(X_1 = 0 \mid Y = r) = \frac{r}{n}, \quad 0 \leq r \leq n.$$

In this case we have $\mathbb{E}(X_1 \mid Y) = Y/n$. □

Properties of conditional expectation

1. For c , a constant, then $\mathbb{E}(cX \mid Y) = c\mathbb{E}(X \mid Y)$ and $\mathbb{E}(c \mid Y) = c$.
2. For random variables $X_1, \dots, X_n$, $\mathbb{E}\left(\sum_i X_i \mid Y\right) = \sum_i \mathbb{E}(X_i \mid Y)$.
3. $\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}(X)$.

Proof. We have

$$\begin{aligned} \mathbb{E}(\mathbb{E}(X \mid Y)) &= \sum_{y \in \Omega_Y} \left(\sum_{x \in \Omega_X} x \mathbb{P}(X = x \mid Y = y) \right) \mathbb{P}(Y = y) \\ &= \sum_{x \in \Omega_X} x \sum_{y \in \Omega_Y} \mathbb{P}(X = x, Y = y) = \sum_{x \in \Omega_X} x \mathbb{P}(X = x) = \mathbb{E}(X). \quad \square \end{aligned}$$

4. When X and Y are independent, $\mathbb{E}(X \mid Y) = \mathbb{E}(X)$.

Proof. For $y \in \Omega_Y$,

$$\mathbb{E}(X \mid Y = y) = \sum_{x \in \Omega_X} x \mathbb{P}(X = x \mid Y = y) = \sum_{x \in \Omega_X} x \mathbb{P}(X = x) = \mathbb{E}(X). \quad \square$$

5. When Y and Z are independent, $\mathbb{E}(\mathbb{E}(X|Y)|Z) = \mathbb{E}(X)$.

Proof. Since $\mathbb{E}(X|Y)$ is a function of Y it is independent of Z , so using Property 4 and then Property 3, we have

$$\mathbb{E}(\mathbb{E}(X|Y)|Z) = \mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X). \quad \square$$

6. For any function $h: \mathbb{R} \rightarrow \mathbb{R}$, we have $\mathbb{E}(h(Y)X|Y) = h(Y)\mathbb{E}(X|Y)$.

Proof. We have, for $y \in \Omega_Y$,

$$\mathbb{E}(h(Y)X|Y=y) = \sum_{\omega: Y(\omega)=y} h(Y(\omega))X(\omega)\mathbb{P}(\{\omega\})/\mathbb{P}(Y=y) = h(y)\mathbb{E}(X|Y=y).$$

A particular consequence of this and Property 1 is that $\mathbb{E}(\mathbb{E}(X|Y)|Y) = \mathbb{E}(X|Y)$. $\square$

7. The conditional expectation $\mathbb{E}(X|Y)$ is that function $h(Y)$ of Y which minimizes $\mathbb{E}(X - h(Y))^2$ over all functions h .

Proof. Write

$$\mathbb{E}(X - h(Y))^2 = \mathbb{E}[X - \mathbb{E}(X|Y) + \mathbb{E}(X|Y) - h(Y)]^2,$$

which may be expanded to

$$\mathbb{E}[X - \mathbb{E}(X|Y)]^2 + \mathbb{E}[\mathbb{E}(X|Y) - h(Y)]^2 + 2\mathbb{E}[(X - \mathbb{E}(X|Y))(\mathbb{E}(X|Y) - h(Y))].$$

Now consider half the cross-product term,

$$\mathbb{E}[(X - \mathbb{E}(X|Y))(\mathbb{E}(X|Y) - h(Y))] \stackrel{3.}{=} \mathbb{E}(\underbrace{(X - \mathbb{E}(X|Y))}_{=0} \underbrace{(\mathbb{E}(X|Y) - h(Y))}_{=g(Y)}|Y)$$

by using Property 3, and then, using Property 6, this

$$\mathbb{E}(\underbrace{(X - \mathbb{E}(X|Y))}_{=0}|Y) = \mathbb{E}(\underbrace{(\mathbb{E}(X|Y) - h(Y))}_{=g(Y)}|Y) = \mathbb{E}((\mathbb{E}(X|Y) - h(Y))\mathbb{E}[(X - \mathbb{E}(X|Y))|Y]);$$

but $\mathbb{E}[(X - \mathbb{E}(X|Y))|Y] \stackrel{4.}{=} \mathbb{E}(X|Y) - \mathbb{E}(X|Y) = 0$, so that

$$\mathbb{E}(X - h(Y))^2 = \mathbb{E}[X - \mathbb{E}(X|Y)]^2 + \mathbb{E}[\mathbb{E}(X|Y) - h(Y)]^2$$

from which the result follows, since the first term in this expression does not involve h and the second term is minimized by $h(Y) = \mathbb{E}(X|Y)$. (per la prop 2.5 pg 17 Alt) $\square$

$$\mathbb{E}(\underbrace{\mathbb{E}(X|Y)}_{f(Y)}|Y) = \mathbb{E}(f(Y)|Y) \stackrel{2.}{=} f(Y) \mathbb{E}(1|Y) \stackrel{1.}{=} f(Y) \quad (*)$$

Example 3.25 *Sum of a random number of random variables* Let $X_1, X_2, \dots$ be independent and identically distributed random variables with common probability generating function $p(z)$. Let N be a non-negative integer valued random variable independent of the $\{X_i\}$ and having probability generating function $q(z)$. We consider the p.g.f. of the random variable $X_1 + \dots + X_N$; (here the sum is 0 if $N = 0$).

$$\begin{aligned} r(z) &= \mathbb{E} (z^{X_1 + \dots + X_N}) = \mathbb{E} (\mathbb{E} (z^{X_1 + \dots + X_N} \mid N)) = \mathbb{E} \left((\mathbb{E} z^{X_1})^N \right) \\ &= \mathbb{E} \left((p(z))^N \right) = q(p(z)). \end{aligned}$$

If at a first reading you find the second equality too cryptic, you might wish to spell out the argument as

$$\begin{aligned} \mathbb{E} (z^{X_1 + \dots + X_N}) &= \sum_{n=0}^{\infty} \mathbb{E} (z^{X_1 + \dots + X_N} \mid N = n) \mathbb{P}(N = n) \\ &= \sum_{n=0}^{\infty} \mathbb{E} (z^{X_1 + \dots + X_n} \mid N = n) \mathbb{P}(N = n) \\ &= \sum_{n=0}^{\infty} \mathbb{E} (z^{X_1 + \dots + X_n}) \mathbb{P}(N = n) = \sum_{n=0}^{\infty} (p(z))^n \mathbb{P}(N = n) = q(p(z)). \end{aligned}$$

After some practice you should find the conditional expectation shorthand notation given first more helpful. It follows from the expression for $r(z)$ that $r'(z) = q'(p(z))p'(z)$, so that

$$\mathbb{E} (X_1 + \dots + X_N) = q'(p(1-))p'(1-) = (\mathbb{E} N) (\mathbb{E} X_1),$$

since $p(1-) = 1$. Furthermore, since $r''(z) = q''(p(z))(p'(z))^2 + q'(p(z))p''(z)$, and the fact that $q''(1-) = \mathbb{E} (N)^2 - \mathbb{E} N$ and $p''(1-) = \mathbb{E} (X_1)^2 - \mathbb{E} X_1$, we may calculate that

$$\begin{aligned} \text{Var} (X_1 + \dots + X_N) &= r''(1-) + r'(1-) - (r'(1-))^2 \\ &= (\mathbb{E} N) \text{Var} (X_1) + (\mathbb{E} X_1)^2 \text{Var} (N). \end{aligned}$$

Notice that the variance of $X_1 + \dots + X_N$ is increased over what it would be if N is constant, $N \equiv \mathbb{E} N = n$, say, by the amount $(\mathbb{E} X_1)^2 \text{Var} (N)$; if $\text{Var} (N) = 0$ and N is constant we get the usual expression for the variance of a sum of n i.i.d. random variables. $\square$

3.6 Branching processes

As an example of conditional expectations and of generating functions we will consider a model of population growth and extinction known as the Bienaymé-Galton-Watson process. Consider a sequence of random variables $X_0, X_1, \dots$, where X_n represents the number of individuals in the n th generation. We will assume that the population is initiated by one individual, take $X_0 \equiv 1$, and when he dies he is replaced by k individuals with probability g_k , $k = 0, 1, 2, \dots$. These individuals behave independently and identically to the parent individual, as do those in subsequent generations. The number in the $(n+1)$ st generation, X_{n+1} , depends on the number in the n th generation and is given by

$$X_{n+1} = \begin{cases} Y_1^n + Y_2^n + \dots + Y_{X_n}^n & \text{when } X_n \geq 1, \\ 0 & \text{when } X_n = 0. \end{cases}$$

Here $\{Y_j^n : n \geq 1, j \geq 1\}$ are independent, identically distributed random variables with $\mathbb{P}(Y_j^n = k) = g_k$, for $k \geq 0$ and Y_j^n represents the number of offspring of the j th individual in the n th generation, $j \leq X_n$.

Assumptions (i) $g_0 > 0$; and (ii) $g_0 + g_1 < 1$.

Assumption (i) means that the population can die out (extinction) since in each generation there is positive probability that all individuals have no offspring; assumption (ii) means that the population may grow, there is positive probability that the next generation has more individuals than the present one. Now let $G(z) = \sum_{k=0}^{\infty} g_k z^k = \mathbb{E}(z^{X_1})$ and set $G_n(z) = \mathbb{E}(z^{X_n})$, for $n \geq 1$, so that $G_1 = G$.

Theorem 3.26 For all $n \geq 1$, $G_{n+1}(z) = G_n(G(z)) = G(\dots(G(z))\dots) = G(G_n(z))$.

Proof. Note that $Y_1^n, Y_2^n, \dots$ are independent of X_n , so that

$$\begin{aligned} G_{n+1}(z) &= \mathbb{E}(z^{X_{n+1}}) = \sum_{k=0}^{\infty} \mathbb{E}(z^{X_{n+1}} \mid X_n = k) \mathbb{P}(X_n = k) \\ &= \sum_{k=0}^{\infty} \mathbb{E}(z^{Y_1^n + \dots + Y_k^n}) \mathbb{P}(X_n = k) = \sum_{k=0}^{\infty} (G(z))^k \mathbb{P}(X_n = k) \\ &= \mathbb{E}((G(z))^{X_n}) = G_n(G(z)). \end{aligned}$$

□

Corollary 3.27 For $m = \mathbb{E}(X_1) = \sum_{k=1}^{\infty} kg_k$ and $\sigma^2 = \text{Var}(X_1) = \sum_{k=0}^{\infty} (k-m)^2 g_k$, then for $n \geq 1$, we have

$$\mathbb{E}(X_n) = m^n, \quad \text{Var}(X_n) = \begin{cases} \frac{\sigma^2 m^{n-1} (m^n - 1)}{m - 1} & \text{when } m \neq 1, \\ n\sigma^2 & \text{when } m=1. \end{cases}$$

Proof. Differentiating $G_n(z) = G_{n-1}(G(z))$ to obtain $G'_n(z) = G'_{n-1}(G(z))G'(z)$ and letting $z \uparrow 1$, it follows that

$$\mathbb{E}(X_n) = m\mathbb{E}(X_{n-1}) = \dots = m^n \mathbb{E}(X_0) = m^n,$$

since $X_0 = 1$. Differentiating $G_n(z)$ a second time gives

$$G''_n(z) = G''_{n-1}(G(z))(G'(z))^2 + G'_{n-1}(G(z))G''(z),$$

and letting $z \uparrow 1$ again we have

$$\mathbb{E}(X_n(X_n - 1)) = m^2 \mathbb{E}(X_{n-1}(X_{n-1} - 1)) + (\sigma^2 + m^2 - m) \mathbb{E}(X_{n-1}).$$

We then have, using the fact that $\mathbb{E}X_n = m^n$,

$$\begin{aligned} \text{Var}(X_n) &= \mathbb{E}(X_n(X_n - 1)) + \mathbb{E}(X_n) - (\mathbb{E}X_n)^2 \\ &= m^2 \mathbb{E}(X_{n-1}(X_{n-1} - 1)) + (\sigma^2 + m^2 - m) \mathbb{E}(X_{n-1}) + m^n - m^{2n} \\ &= m^2 [\text{Var}(X_{n-1}) - \mathbb{E}(X_{n-1}) + (\mathbb{E}X_{n-1})^2] + (\sigma^2 + m^2) m^{n-1} - m^{2n} \\ &= m^2 \text{Var}(X_{n-1}) + \sigma^2 m^{n-1}. \end{aligned}$$

Iterating this, we see that

$$\begin{aligned} \text{Var}(X_n) &= m^2 \text{Var}(X_{n-1}) + \sigma^2 m^{n-1} = m^4 \text{Var}(X_{n-2}) + \sigma^2 (m^{n-1} + m^n) = \dots \\ &= m^{2n} \text{Var}(X_0) + \sigma^2 (m^{n-1} + \dots + m^{2n-2}) \\ &= \sigma^2 (m^{n-1} + \dots + m^{2n-2}), \text{ since } \text{Var}(X_0) = 0 \text{ because } X_0 = 1, \end{aligned}$$

and then the result may be obtained immediately. □

Probability of extinction Notice that $X_n = 0$ implies that $X_{n+1} = 0$ so that if we let $A_n = (X_n = 0)$, the event that the population is extinct at or before generation n , we

have $A_n \subseteq A_{n+1}$ and $A = \bigcup_{n=1}^{\infty} A_n$ represents the event that extinction ever occurs. Notice that $\mathbb{P}(A_n) = G_n(0)$ and by the continuity property of probabilities on increasing events we see that the extinction probability, q , say, is

$$q = \mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \lim_{n \rightarrow \infty} G_n(0) = \lim_{n \rightarrow \infty} \mathbb{P}(X_n = 0).$$

Theorem 3.28 *The extinction probability q is the smallest positive root of the equation $G(z) = z$. When m , the mean number of offspring per individual, satisfies $m \leq 1$ then $q = 1$; when $m > 1$ then $q < 1$.*

Proof. The fact that the extinction probability q is well defined follows from the above and since G is continuous and $q = \lim_{n \rightarrow \infty} G_n(0)$ we have $G\left(\lim_{n \rightarrow \infty} G_n(0)\right) = \lim_{n \rightarrow \infty} G_{n+1}(0)$, so that $G(q) = q$, that is q is a root of $G(z) = z$; note that 1 is always a root since $G(1) = \sum_{r=0}^{\infty} g_r = 1$. Now let $H(z) = G(z) - z$, then $H'' = \sum_{r=0}^{\infty} r(r-1)g_r z^{r-2} > 0$ for $0 < z < 1$ provided $g_0 + g_1 < 1$, so the derivative of H is strictly increasing in the range $0 < z < 1$, hence H can have at most one root different from 1 in $[0, 1]$ (Rolle's Theorem).

Firstly, suppose that H has no root in $[0, 1)$ then, since $H(0) = g_0 > 0$ we must have $H(z) > 0$ for all $0 < z < 1$, so $H(1) - H(z) < H(1) = 0$ and so

$$H'(1-) = \lim_{z \uparrow 1} \frac{H(1) - H(z)}{1 - z} \leq 0, \quad \text{whence } m = G'(1-) \leq 1.$$

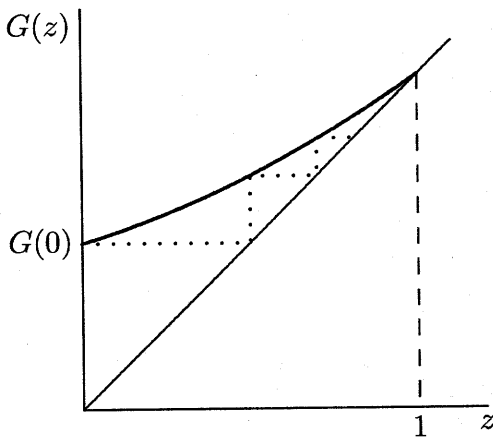


Fig. 1: $m \leq 1, q = 1$

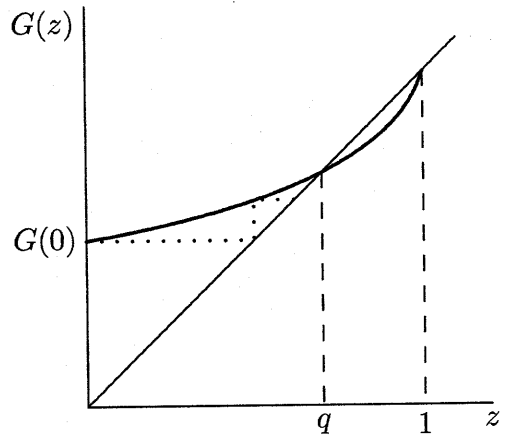


Fig. 2: $m > 1, q < 1$

Next, suppose that H has a unique root r in $[0, 1)$, then H' must have a root in $[r, 1)$; that is $H'(z) = G'(z) - 1 = 0$ for some z , $r \leq z < 1$. The function G' is strictly increasing (since $g_0 + g_1 < 1$) so that $m = G'(1-) > G'(z) = 1$. Thus we see that $m \leq 1$, if and only if, $q = 1$. Furthermore, if $m > 1$, let $\alpha > 0$ be the smallest positive root of $G(z) = z$, so that because G is increasing, $\alpha = G(\alpha) \geq G(0)$, and repeating n times we have $\alpha \geq G_n(0)$, whence $\alpha \geq \lim_{n \rightarrow \infty} G_n(0) = q$, so that we must have $\alpha = q$, completing the proof. $\square$

Note Figures 1 and 2 illustrate the two situations $m \leq 1$ and $m > 1$; the dotted lines illustrate the iteration $G_{n+1}(0) = G(G_n(0))$ tending to the smallest positive root, q .

3.7 Random walks

Let $X_1, X_2, \dots$ be i.i.d. random variables and set $S_n = S_0 + X_1 + \dots + X_k$ for $k \geq 1$ where S_0 is a constant then $\{S_k, k \geq 0\}$ is known as a (one-dimensional) **random walk**. When each X_i just takes the two values $+1$ and -1 with probabilities p and $q = 1 - p$, respectively, it is a **simple random walk** and further when $p = q = \frac{1}{2}$ it is a **simple, symmetric random walk**. We will consider simple random walks.

Recurrence relations The problems we will look at for the simple random walk often reduce to the solution of recurrence relations (or difference equations). We consider the general solution of such equations in the simplest situations which have constant coefficients.

1. First-order equations: The general first-order equation is $x_{n+1} = ax_n + b$, for $n \geq 0$, where a and b are constants; the case $b = 0$ gives the general first-order homogeneous equation $x_{n+1} = ax_n$, which trivially may be solved as $x_n = a^n x_0$; if y_n is any solution of the inhomogeneous equation, then the general solution of the inhomogeneous equation is of the form $x_n = Ca^n + y_n$ for some constant C (because $x_n - y_n$ must be a solution of the homogeneous equation). The constant is determined by a boundary condition.
2. Second-order equations: $x_{n+1} = ax_n + bx_{n-1} + c$, for $n \geq 1$, where a , b and c are constants. First consider the homogeneous case where $c = 0$. Then write the

relation in matrix form as follows:

$$\begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_n \\ x_{n-1} \end{pmatrix} = A \begin{pmatrix} x_n \\ x_{n-1} \end{pmatrix}, \quad \text{where } A = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix}.$$

It follows that

$$\begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix} = A^n \begin{pmatrix} x_1 \\ x_0 \end{pmatrix};$$

find the eigenvalues of A , by solving

$$\begin{vmatrix} a - \lambda & b \\ 1 & -\lambda \end{vmatrix} = 0, \quad \text{to give the equation } \lambda^2 - a\lambda - b = 0,$$

with roots λ_1 and λ_2 , say. This equation is known as the **auxiliary equation** of the recurrence relation; it corresponds to seeking a solution of the form $x_n = \lambda^n$.

If λ_1 and λ_2 are distinct then for some matrix Λ we may write

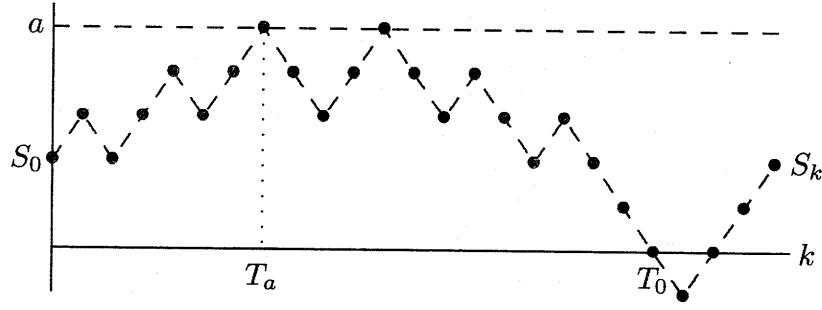
$$A = \Lambda^{-1} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \Lambda \quad \text{and then} \quad A^n = \Lambda^{-1} \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix} \Lambda,$$

so that the general solution of the homogeneous equation may be seen to be of the form $x_n = C\lambda_1^n + D\lambda_2^n$ for some constants C and D . If the eigenvalues are not distinct, $\lambda_1 = \lambda_2 = \lambda$, then

$$A = \Lambda^{-1} \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \Lambda \quad \text{and then} \quad A^n = \Lambda^{-1} \begin{pmatrix} \lambda^n & n\lambda^{n-1} \\ 0 & \lambda^n \end{pmatrix} \Lambda,$$

and then the general solution of the homogeneous equation may be seen to be of the form $x_n = \lambda^n(C + Dn)$ for some constants C and D . As before, if y_n is any particular solution of the inhomogeneous equation, the general solution is of the form $x_n + y_n$ where x_n is the general solution of the homogeneous equation.

Example 3.29 *Gambler's ruin* For the simple random walk, $\{S_k\}$ may represent the fortune of a gambler after k plays of a game where on each play he either wins £1, with probability p , or loses £1 with probability $q = 1 - p$; his initial fortune is £ S_0 and a classical problem is to calculate the probability that his fortune achieves the level a , $a > S_0$, before the time of ruin, that is the time that he goes bankrupt (his fortune hits the level 0). If T_a denotes the first time that the random walk hits the level a and T_0 the time the random



walk first hits the level 0, we would wish to calculate $\mathbb{P}(T_a < T_0)$, given that his fortune starts at $S_0 = r$, $0 < r < a$.

The figure illustrates a path of the random walk—although, in the case of the game, it finishes at the instant T_0 , the time of bankruptcy! Let $x_r = \mathbb{P}(T_a < T_0)$ when $S_0 = r$, for $0 \leq r \leq a$, so that we have the boundary conditions $x_a = 1$ and $x_0 = 0$. A general rule in problems of this type in probability may be summed up as ‘condition on the first thing that happens’, which here would be a shorthand for using the Law of Total Probability to express the probability conditional on the outcome of the first play of the game, that is, whether $X_1 = 1$ or $X_1 = -1$, or equivalently, $S_1 = r + 1$ or $S_1 = r - 1$. Thus, for $0 < r < a$,

$$\begin{aligned} x_r &= \mathbb{P}(T_a < T_0 \mid S_1 = r + 1) \mathbb{P}(X_1 = 1) + \mathbb{P}(T_a < T_0 \mid S_1 = r - 1) \mathbb{P}(X_1 = -1) \\ &= px_{r+1} + qx_{r-1}. \end{aligned}$$

The auxiliary equation for this recurrence relation is $p\lambda^2 - \lambda + q = 0$, and since $p + q = 1$, this may be factored as $(\lambda - 1)(p\lambda - q) = 0$ to give roots $\lambda = 1$ and $\lambda = q/p$.

Case $p \neq q$: the roots are distinct and the general solution is of the form $x_r = A + B(q/p)^r$ for some constants A and B ; the boundary conditions at $r = a$ and $r = 0$, fix A and B and we conclude that

$$x_r = \mathbb{P}(T_a < T_0) = \frac{1 - (q/p)^r}{1 - (q/p)^a}, \quad \text{for } 0 \leq r \leq a.$$

Case $p = q = \frac{1}{2}$: here $\lambda = 1$ is a repeated root of the auxiliary equation so that the general solution of the recurrence relation is $x_r = A + Br$, which, after using the boundary conditions, leads to the solution $x_r = r/a$, $0 \leq r \leq a$.

We do not know necessarily that at least one of T_0 and T_a must be finite, but if we interchange p and q and replace r by $a - r$, (or just calculate directly as above) we may obtain, for $S_0 = r$, $0 \leq r \leq a$, that

$$\mathbb{P}(T_0 < T_a) = \begin{cases} \frac{(q/p)^r - (q/p)^a}{1 - (q/p)^a} & \text{when } p \neq q, \\ 1 - r/a & \text{when } p = q = \frac{1}{2}. \end{cases}$$

It follows, in both cases, that $\mathbb{P}(T_a < T_0) + \mathbb{P}(T_0 < T_a) = 1$, so that at least one of the two barriers, 0 or a , must be reached with certainty. $\square$

Example 3.30 *Probability of ruin* From the previous calculation we may derive an expression for $\mathbb{P}(T_0 < \infty)$ given $S_0 = r > 0$, which is the probability that ruin ever happens. We see that the event that ruin occurs may be written as

$$(T_0 < \infty) = \bigcup_{a=r+1}^{\infty} (T_0 < T_a);$$

the events in the union are expanding as a increases, so by the continuity of the probability on expanding events, we have

$$\mathbb{P}(T_0 < \infty) = \lim_{a \rightarrow \infty} \mathbb{P}(T_0 < T_a) = \begin{cases} (q/p)^r & \text{when } p > q, \\ 1 & \text{when } p \leq q, \end{cases}$$

so that ruin is certain except in the case when the probability of winning a play is strictly larger than $\frac{1}{2}$. $\square$

Example 3.31 *Expected duration of the game* Suppose that the gambler plays either until his fortune reaches a or until he goes bankrupt, whichever is sooner. That is the number of plays is $\min(T_0, T_a) = T_0 \wedge T_a$. We will derive the expected length of the game, $\mathbb{E}(T_0 \wedge T_a)$, given that $S_0 = r$, $0 \leq r \leq a$, which we will denote by m_r . We do not know whether m_r is finite. Consider blocks of jumps of the random walk of length a , that is

$$\begin{array}{cccc} X_1 & X_2 & \cdots & X_a \\ X_{a+1} & X_{a+2} & \cdots & X_{2a} \\ X_{2a+1} & X_{2a+2} & \cdots & X_{3a} \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

and for $i \geq 1$ set $Y_i = 1$ if either $X_{(i-1)a+1} = X_{(i-1)a+2} = \cdots = X_{ia} = 1$ or $X_{(i-1)a+1} = X_{(i-1)a+2} = \cdots = X_{ia} = -1$, otherwise $Y_i = 0$. Thus $Y_i = 1$ if and only if the i th block of plays is a run of all wins or all losses, and $\mathbb{P}(Y_i = 1) = 1 - \mathbb{P}(Y_i = 0) = p^a + q^a = \theta$, say. If we let Z be the first i such that $Y_i = 1$, then Z has a geometric distribution $\mathbb{P}(Z = j) = (1 - \theta)^{j-1}\theta$, $j \geq 1$, and so $\mathbb{E}(Z) = 1/\theta < \infty$. But it is clear that $T_0 \wedge T_a \leq aZ$, hence we see that $\mathbb{E}(T_0 \wedge T_a) \leq a\mathbb{E}(Z) < \infty$. To compute m_r , we again condition on the first thing to happen, that is whether the first play is a win or loss, to see that for $0 < r < a$,

$$m_r = p(1 + m_{r+1}) + q(1 + m_{r-1}) = 1 + pm_{r+1} + qm_{r-1}, \quad \text{with } m_0 = m_a = 0;$$

here the 1 in the recurrence relation counts the initial play of the game. The solution of the homogeneous equation is again $m_r = A + B(q/p)^r$ when $p \neq q$ and $m_r = A + Br$ for the case $p = q = \frac{1}{2}$.

Case $p \neq q$: look for a particular solution of the inhomogeneous equation with $m_r = cr$, then

$$cr = 1 + pc(r+1) + qc(r-1), \quad \text{so that } c = 1/(q-p),$$

so that the general solution is $m_r = r/(q-p) + A + B(q/p)^r$, and after using the boundary conditions we have

$$m_r = \frac{r}{q-p} - \left(\frac{a}{q-p} \right) \frac{1 - (q/p)^r}{1 - (q/p)^a}.$$

Case $p = q = \frac{1}{2}$: a particular solution of the inhomogeneous equation is $-r^2$, so the general solution is $m_r = A + Br - r^2$ and after using the boundary conditions we have $m_r = r(a-r)$. □

10 February 2007

4. CONTINUOUS RANDOM VARIABLES

4.1 Introduction

Up to now we have restricted consideration to sample spaces Ω which are finite, or countable; we will now relax that assumption. We assume that we have a probability $\mathbb{P}(\cdot)$ defined on subsets (events) of Ω satisfying the axioms given previously. We will be interested in random variables which may take on uncountably many values. Here if $X : \Omega \rightarrow \mathbb{R}$, define the **distribution function** (sometimes called the **cumulative distribution function**) of X as *(fonction de répartition)*

$$F(x) = \mathbb{P}(X \leq x), \quad -\infty < x < \infty,$$

so that $F : \mathbb{R} \rightarrow [0, 1]$. Note that $\mathbb{P}(X > x) = 1 - F(x)$.

Properties of the distribution function $F(x)$

- 1. $F(x)$ is non-decreasing in x , $-\infty < x < \infty$.

Proof. If $x \leq y$, then the event $(X \leq x) \subseteq (X \leq y)$, so that

$$F(x) = \mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y) = F(y). \quad \square$$

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

Proof. We have

$$\begin{aligned} \mathbb{P}(a < X \leq b) &= \mathbb{P}((X \leq a)^c \cap (X \leq b)) \\ &= \mathbb{P}((X \leq a)^c) + \mathbb{P}(X \leq b) - \mathbb{P}((X \leq b) \cup (X \leq a)^c) \\ &= 1 - \mathbb{P}(X \leq a) + \mathbb{P}(X \leq b) - \mathbb{P}(\Omega) = F(b) - F(a). \quad \square \end{aligned}$$

- 3. $F(x)$ is right continuous in x ; that is, when $y \downarrow x$ we have $F(y) \downarrow F(x)$; since F is non-decreasing the limit from the left $\lim_{y \uparrow x} F(y) = F(x-) \leq F(x)$ always exists.

Proof. Fix x , then for $n \geq 1$, consider the event

$$A_n = (x < X \leq x + 1/n) = (X \leq x + 1/n) \cap (X \leq x)^c;$$

then the $\{A_n\}$ are decreasing events $A_n \supseteq A_{n+1}$, and $\bigcap A_n = \emptyset$, so by the continuity property of probabilities $\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 0$. But $\mathbb{P}(A_n) = F(x + 1/n) - F(x)$, from which the conclusion follows. $\square$

4. $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$. *Proof: $F(x)$ non-decreasing and limited a.s. 1*

We say that a random variable X is **continuous** if its distribution function, F , is a continuous function. We have seen that a distribution function is necessarily right continuous, then if X is a continuous random variable, F must also be left continuous. This is equivalent to the statement that $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$, since as in the proof of Property 2, we will have $\mathbb{P}(X = x) \stackrel{?}{=} \lim_{y \uparrow x} \mathbb{P}(y < X \leq x) = \lim_{y \uparrow x} [F(x) - F(y)]$.

In discussing continuous random variables we will restrict consideration to the situation where F is not only continuous but also differentiable, and we will set $f(x) = F'(x)$; $f(\cdot)$ is known as the **probability density function** (p.d.f) of the random variable X . A probability density function satisfies the following two conditions:

$$(i) \quad f(x) \geq 0, \text{ for all } x \in \mathbb{R}, \quad (ii) \quad \int_{-\infty}^{\infty} f(x) dx = 1,$$

and then $F(x) = \int_{-\infty}^x f(y) dy$.

Note that for a discrete random variable the distribution function is a right-continuous step function as illustrated in Figure 1, with the heights of the steps being $\mathbb{P}(X = x_i)$ for the possible values x_i , while for a continuous random variable the distribution function is a continuous non-decreasing function as in Figure 2.

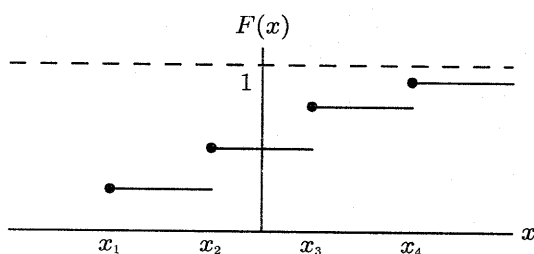


Fig. 1: X discrete

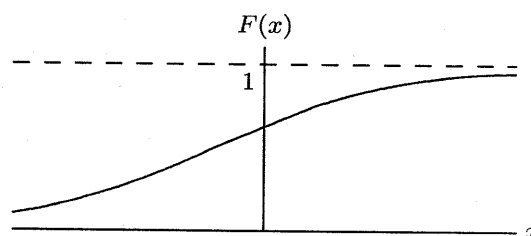


Fig. 2: X continuous

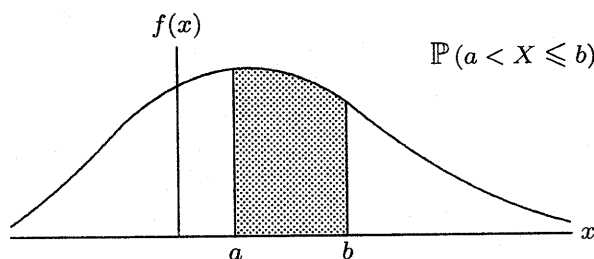
Q (i) $F(x)$ is non-decreasing $\Rightarrow f(x) = F'(x) \geq 0$
(ii) $\int_{-\infty}^{\infty} f(x) dx = [F(x)]_{-\infty}^{\infty} = F(\infty) - F(-\infty) = 1 - 0 = 1$

Note that there is not a straight split between discrete and continuous random variables, it is possible to have a random variable which is continuous over some ranges of values while at the same time taking certain values with positive probabilities; however, in this course we will deal with the two cases separately.

The intuitive interpretation of the p.d.f. is that for small Δx ,

$$\mathbb{P}(x < X \leq x + \Delta x) = F(x + \Delta x) - F(x) = \int_x^{x+\Delta x} f(y) dy \approx f(x)\Delta x,$$

so that while $f(x)$ does not represent a probability, the probability that X lies in a small interval around x is proportional to $f(x)$, and for this reason many intuitive arguments involving probabilities carry over to probability density functions. Note that areas under the probability density function represent probabilities as illustrated in the figure.



More generally, for a set $S \subseteq \Omega_X$, we have $\mathbb{P}(X \in S) = \int_{x \in S} f(x) dx$.

4.2 Expectation, variance and standard distributions

Consider a continuous random variable X with distribution function F and p.d.f. f .

Then set

$$\mathbb{E}(X_+) = \int_0^\infty x f(x) dx, \quad \text{and} \quad \mathbb{E}(X_-) = \int_{-\infty}^0 (-x) f(x) dx,$$

and if not both $\mathbb{E}(X_+)$ and $\mathbb{E}(X_-)$ are infinite then define the expectation of X to be

$$\mathbb{E}(X) = \mathbb{E}(X_+) - \mathbb{E}(X_-) = \int_{-\infty}^\infty x f(x) dx;$$

otherwise, the expectation is not defined.

For a continuous non-negative random variable $X \geq 0$, we may write

$$\mathbb{E} X = \int_0^\infty (1 - F(x)) dx,$$

$$\Delta \sum_{y=0}^{\infty} \sum_{x=0}^y f(x,y) = \sum_{x=0}^{\infty} \sum_{y=x}^{\infty} f(x,y), \text{ infact: } \{(x,y) \mid y \geq 0, 0 \leq x \leq y\} = \{(x,y) \mid x \geq 0, y \geq x\}$$

$(X \geq 0, X < 0 \Rightarrow P(X \leq x) = 0 \forall x < 0)$

since

$$\begin{aligned} \mathbb{E} X &= \int_0^{\infty} y f(y) dy = \int_{y=0}^{\infty} \left(\int_{x=0}^y dx \right) f(y) dy \\ &= \int_{x=0}^{\infty} \left(\int_{y=x}^{\infty} f(y) dy \right) dx = \int_0^{\infty} (1 - F(x)) dx, \end{aligned}$$

by interchanging the order of integration. By considering X_+ and X_- , we may see that for any continuous random variables we may write

$$\mathbb{E} X = \int_0^{\infty} (1 - F(x)) dx - \int_{-\infty}^0 F(x) dx.$$

Observe that the properties of expectation as set out for discrete random variables carry over to the situation here with one change, which is that for a function $g(\cdot)$,

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

We may define the variance of a continuous random variable in exactly the same way, $\text{Var} (X) = \mathbb{E} (X - \mathbb{E} X)^2$, and its properties are exactly as before; in particular $\text{Var} (X) = \mathbb{E} (X^2) - (\mathbb{E} X)^2$. The standard deviation of X is again just $\sqrt{\text{Var} (X)}$.

● **Example 4.1** *The exponential distribution* One of the two most important continuous distribution is the exponential distribution for which the random variable X has the probability density function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, with $f(x) = 0$ for $x < 0$, where $\lambda > 0$ is a constant. We write $X \sim \text{Exp}(\lambda)$. First note that $\int_0^{\infty} \lambda e^{-\lambda x} dx = 1$, so that f is a genuine p.d.f.. Then, for $x \geq 0$,

$$F(x) = \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x}.$$

We may calculate

$$\mathbb{E} (X) = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \int_0^{\infty} x d(-e^{-\lambda x}) = [-x e^{-\lambda x}]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Furthermore, using integration by parts again, we may also obtain that

$$\mathbb{E} (X^2) = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = \int_0^{\infty} x^2 d(-e^{-\lambda x}) = [-x^2 e^{-\lambda x}]_0^{\infty} + 2 \int_0^{\infty} x e^{-\lambda x} dx = \frac{2}{\lambda^2},$$

using the previous calculation, so that $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = 1/\lambda^2$.

The exponential distribution is sometimes used to model the lifetime of a component. If X is the lifetime and $X \sim \text{Exp}(\lambda)$, then the probability that the component survives a length of time $x > 0$ is $\mathbb{P}(X > x) = e^{-\lambda x}$. Then for $x > 0$ and $y > 0$,

$$\begin{aligned}\mathbb{P}(X > x + y \mid X > y) &= \frac{\mathbb{P}(X > x + y, X > y)}{\mathbb{P}(X > y)} = \frac{\mathbb{P}(X > x + y)}{\mathbb{P}(X > y)} \\ &= \frac{e^{-\lambda(x+y)}}{e^{-\lambda y}} = e^{-\lambda x},\end{aligned}$$

so that, given the component has survived a length of time y the probability that it will survive a further time x is the same as if it has just been installed. This property, which is crucial to the study of stochastic processes, is known as the lack of memory property of the exponential distribution. $\square$

Theorem 4.2 Suppose that X is a continuous random variable with p.d.f. $f(x)$ and $g: \mathbb{R} \rightarrow \mathbb{R}$ is a continuous function which is either strictly increasing or strictly decreasing and with g^{-1} is differentiable, then $g(X)$ is a continuous random variable with p.d.f.

$$f(g^{-1}(x)) \left| \frac{d}{dx} g^{-1}(x) \right|.$$

Proof. Suppose that g is strictly increasing (then g^{-1} is also, so its derivative is positive), we see that the distribution of $g(X)$ is

$$\mathbb{P}(g(X) \leq x) = \mathbb{P}(X \leq g^{-1}(x)) = F(g^{-1}(x));$$

differentiating with respect to x to obtain the p.d.f. gives the result. When g is decreasing so also is its inverse (so $\frac{d}{dx} g^{-1}(x)$ is negative) and we have

$$\mathbb{P}(g(X) \leq x) = \mathbb{P}(X \geq g^{-1}(x)) = \mathbb{P}(X > g^{-1}(x)) = 1 - F(g^{-1}(x)),$$

because $\mathbb{P}(X = g^{-1}(x)) = 0$, since X is continuous, and the result follows by differentiating. $\square$

$$\mathbb{P}(X=x) = 0 \quad \forall x \in \mathbb{R}$$

59

Notes: g^{-1} derivable $\Rightarrow g^{-1}$ continua $\Rightarrow g$ continua

● **Example 4.3** *The normal distribution* The normal distribution (also known as the Gaussian distribution) is the most important continuous distribution; its significance stems from the Central Limit Theorem which we will consider later. The probability density is specified by two parameters μ , $-\infty < \mu < \infty$, and $\sigma > 0$, and is given by

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

First, we must check that this is indeed a p.d.f. in that it integrates to 1. By making the substitution $u = (x - \mu)/\sigma$ we see that

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = 2 \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du,$$

by the symmetry of the integrand around $u = 0$. Then we may calculate as follows,

$$I^2 = \frac{2}{\pi} \int_{u=0}^{\infty} \int_{v=0}^{\infty} e^{-\frac{1}{2}(u^2+v^2)} dudv,$$

then going to polar coordinates $u = r \cos \theta$ and $v = r \sin \theta$, this

$$= \frac{2}{\pi} \int_{r=0}^{\infty} \int_{\theta=0}^{\pi/2} e^{-\frac{1}{2}r^2} r dr d\theta = \frac{2}{\pi} \int_{\theta=0}^{\pi/2} \left(\int_{r=0}^{\infty} e^{-\frac{1}{2}r^2} d(r^2/2) \right) d\theta = 1,$$

showing that $I = 1$. To calculate the mean, by making the substitution $u = (x - \mu)/\sigma$, we see that

$$\mathbb{E} X = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx = \sigma \int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \mu,$$

because the first integral is 0, since the integrand is an odd function, and the second integral is 1, as we have just established. The same substitution shows that

$$\text{Var}(X) = \mathbb{E}(X - \mu)^2 = \int_{-\infty}^{\infty} \frac{(x - \mu)^2}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx = \sigma^2 \int_{-\infty}^{\infty} \frac{u^2}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \sigma^2 \int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

then, integrating by parts, this

$$= \sigma^2 \int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi}} d(-e^{-\frac{1}{2}u^2}) = \sigma^2 \left(\left[-\frac{u}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \right) = \sigma^2.$$

We see that the two parameters μ and σ^2 of the normal distribution represent the mean and variance of X , (σ is the standard deviation of X); we usually write $X \sim N(\mu, \sigma^2)$. The

special case $\mu = 0$ and $\sigma^2 = 1$ gives what is known as the **standard normal distribution**, $N(0, 1)$; the distribution function in this case is usually denoted by $\Phi(x)$ and is given by

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

Denote the p.d.f. of the standard normal distribution by $\phi(x) = \Phi'(x) = e^{-x^2/2}/\sqrt{2\pi}$, then, since $\phi(x) = \phi(-x)$, we have that

$$\Phi(x) = 1 - \Phi(-x), \quad -\infty < x < \infty. \quad (*)$$

Note that if $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ where a and b are constants with $a \neq 0$, then $Y \sim N(a\mu + b, a^2\sigma^2)$. To see this, apply Theorem 4.2 with $y = g(x) = ax + b$, so that the inverse is $g^{-1}(y) = (y - b)/a$, to show that the p.d.f. of $Y = g(X)$ evaluated at y is

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-(g^{-1}(y)-\mu)^2/(2\sigma^2)} \left| \frac{d}{dy} g^{-1}(y) \right| = \frac{1}{\sqrt{2\pi}|a|\sigma} e^{-(y-a\mu-b)^2/(2a^2\sigma^2)},$$

as required. Note that, when $X \sim N(\mu, \sigma^2)$, it follows that $((X - \mu)/\sigma) \sim N(0, 1)$. This fact is important since it enables the calculation of a probability for any $X \sim N(\mu, \sigma^2)$ to be expressed in terms of the standard normal distribution, by subtracting off the mean μ and dividing by the standard deviation σ , as for example

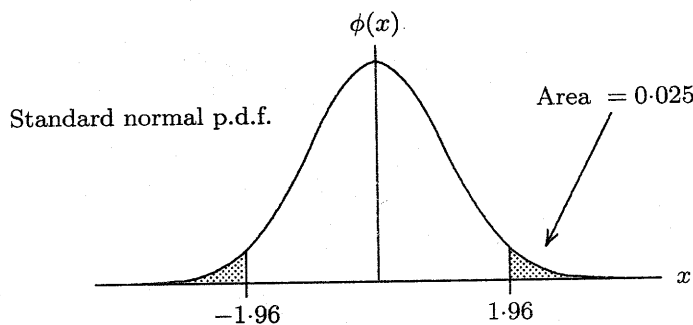
$$\mathbb{P}(X \leq a) = \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right) = \Phi\left(\frac{a - \mu}{\sigma}\right).$$

Important points of the standard normal distribution function are

x	1.28	1.64	1.96	2.33
$\Phi(x)$	0.90	0.95	0.975	0.99

The third of these points leads to an important observation: for $X \sim N(\mu, \sigma^2)$,

$$\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = \mathbb{P}\left(\left|\frac{X - \mu}{\sigma}\right| \leq 2\right) \geq \mathbb{P}\left(\left|\frac{X - \mu}{\sigma}\right| \leq 1.96\right) = 0.95,$$



$$\Phi[-\Phi(-v)] = \Phi'(-v) = \phi(-v)$$

$$\Phi(x) = \int_{-\infty}^x \phi(v) dv = \int_{-\infty}^x \phi(v) dv = [-\Phi(-v)]_{-\infty}^x = -\Phi(-x) + \underbrace{\Phi(\infty)}_1 = 1 - \Phi(-x) \quad (*)$$

which is usually summed up in the statement “more than 95% of the normal distribution is within two standard deviations of the mean”. $\square$

Example 4.4 *The uniform distribution* For constants $a < b$, let $f(x) = 1/(b-a)$ for $a \leq x \leq b$, and $f(x) = 0$, otherwise. Then the random variable has the uniform distribution on the interval $[a, b]$, and we write $X \sim U[a, b]$. Note that

$$\mathbb{E} X = \int_a^b x/(b-a) dx = (a+b)/2,$$

and similarly $\mathbb{E} (X^2) = (a^2 + ab + b^2)/3$, which implies that $\text{Var} (X) = \frac{1}{12}(b-a)^2$.

In the case where $X \sim U(0, 1]$, let $Y = -\log(X)$, then for $y \geq 0$,

$$\mathbb{P}(Y \leq y) = \mathbb{P}(-\log(X) \leq y) = \mathbb{P}(X \geq e^{-y}) = \int_{e^{-y}}^1 dx = 1 - e^{-y},$$

so that $Y \sim \text{Exp}(1)$; that is, Y has the exponential distribution with parameter 1. $\square$

A result that is important for computer simulation of random variables is the following.

Theorem 4.5 Suppose that $U \sim U[0, 1]$, then for any continuous distribution function F , the random variable $X = F^{-1}(U)$ has distribution function F .

Proof. Note that for $u \in [0, 1]$, $\mathbb{P}(U \leq u) \stackrel{(*)}{=} u$, so we have

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

$\uparrow$
 $F \text{ non-decreasing.}$

which gives the result. $\square$

Note There is a corresponding result for discrete random variables. Suppose that F is the distribution function of a discrete random variable and that $p_j = F(x_j) - F(x_{j-1}) > 0$, $j = 1, 2, \dots$, for values $x_1, x_2, \dots$, where $\sum_j p_j = 1$. Now suppose that $U \sim U[0, 1]$ and define a random variable X , by setting $X = x_1$ when $0 < U \leq p_1$, and for $j > 1$, set

$$X = x_j, \quad \text{when} \quad \sum_{i=1}^{j-1} p_i < U \leq \sum_{i=1}^j p_i;$$

then $\mathbb{P}(X = x_j) = p_j$, for each j , and X has the distribution function F . As a consequence, in order to simulate any random variable it is only necessary to use a random number

(*) $\mathbb{P}(U \leq u) = \int_{-\infty}^u f(x) dx = \int_0^u \frac{1}{b-a} dx = \frac{1}{b-a} u = u$
 $\uparrow$
 $f(x) = \frac{1}{b-a} \forall x \in [0, 1]$

generator to provide a random number uniform in $[0, 1]$ and then use the above procedures in the continuous and discrete cases. $\square$

The **median** m of a continuous random variable X with density function f is the point which satisfies

$$\mathbb{P}(X \geq m) = \int_m^\infty f(x)dx = \int_{-\infty}^m f(x)dx = \mathbb{P}(X \leq m) = \frac{1}{2}.$$

Thus half the distribution lies on one side of m and half on the other. For a discrete random variable, X , a median m is a point satisfying

$$\mathbb{P}(X \geq m) \geq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X \leq m) \geq \frac{1}{2}.$$

[Note that for the normal distribution $N(\mu, \sigma^2)$ the mean is equal to the median (and this is true for any symmetric distribution).

A **mode** of a continuous random variable, with density function f , is a point m for which $f(m) \geq f(x)$ for all x ; that is the density function is maximized at a mode. For a discrete random variable a mode is a point, m , for which $\mathbb{P}(X = m) \geq \mathbb{P}(X = x)$ for all possible values x .

In the case of the normal distribution function, the mean and median are also the mode. For example, for the $\text{Exp}(\lambda)$ distribution with density function $\lambda e^{-\lambda x}$ for $x > 0$, we have seen that the mean is $1/\lambda$; it is easy to check that the median is $\log 2/\lambda$ and the mode is 0.

4.3 Joint distribution functions

To start with, to keep the notation simpler, consider just the case of two random variables. The **joint distribution function** of X and Y is

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) \quad \text{for} \quad -\infty < x < \infty, -\infty < y < \infty,$$

so that $F: \mathbb{R}^2 \rightarrow [0, 1]$. If there exists a function $f(\cdot, \cdot)$ with

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv, \quad \text{so that} \quad f(x, y) = \frac{\partial^2 F}{\partial x \partial y},$$

then f is the **joint probability density function** of X and Y . Note that, for any region $C \subseteq \mathbb{R}^2$,

$$\mathbb{P}((X, Y) \in C) = \int \int_{(x, y) \in C} f(x, y) dx dy.$$

Furthermore,

$$f_X(x) = \int_{y=-\infty}^{\infty} f(x, y) dy \quad \text{and} \quad f_Y(y) = \int_{x=-\infty}^{\infty} f(x, y) dx$$

are the **marginal** probability density functions of X and Y , respectively.

Properties of the joint distribution function $F(x, y)$

1. $F(x, y)$ is non-decreasing in y for each fixed x , and in x for each fixed y .
2. $F(x, y)$ is right continuous in y for each fixed x , and in x for each fixed y .
3. $F(\infty, \infty) = \lim_{x \uparrow \infty} \lim_{y \uparrow \infty} F(x, y) = 1$; for each fixed x , $F(x, -\infty) = \lim_{y \downarrow -\infty} F(x, y) = 0$ and for each fixed y , $F(-\infty, y) = \lim_{x \downarrow -\infty} F(x, y) = 0$. Furthermore, $F(x, \infty) = \mathbb{P}(X \leq x)$ and $F(\infty, y) = \mathbb{P}(Y \leq y)$ are the **marginal** probability distributions of X and Y , respectively.
4. For all x_1, x_2, y_1 and y_2 with $x_1 < x_2, y_1 < y_2$,

$$F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \geq 0.$$

Proof. The result follows from the observation that the expression that the left-hand side

$$\mathbb{P}(X \leq x_2, Y \leq y_2) - \mathbb{P}(X \leq x_1, Y \leq y_2) - \mathbb{P}(X \leq x_2, Y \leq y_1) + \mathbb{P}(X \leq x_1, Y \leq y_1)$$

equals $\mathbb{P}(x_1 < X \leq x_2, y_1 < Y \leq y_2) \geq 0$. This is most easily seen by plotting in $\mathbb{R}^2$ the different regions in which (X, Y) lies corresponding to the different probabilities. $\square$

Properties of the joint probability density function $f(x, y)$

1. $f(x, y) \geq 0$, for all x, y .
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$.
3. For any random variable of the form $g(X, Y)$, for some function g , we compute the expectation as

$$\mathbb{E} g(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy,$$

in particular, we may obtain the covariance in the continuous case with the same definition as in the discrete case

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

and it has the same properties as set out previously. Likewise for the correlation coefficient in the context of continuous random variables; it is defined in the same way as for discrete random variables, $\text{Corr}(X, Y) = \text{Cov}(X, Y) / \sqrt{\text{Var}(X)\text{Var}(Y)}$, and it has the same properties as mentioned in the discrete case.

- We define the **conditional density of X given $Y = y$** to be

$$f_{X|Y}(x | y) = \frac{f(x, y)}{f_Y(y)};$$

- note that the Law of Total Probability here is that the marginal density of X may be expressed as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x | y) f_Y(y) dy.$$

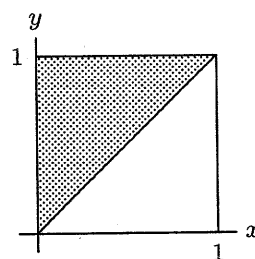
- Then the conditional expectation of X given $Y = y$ is

$$\mathbb{E}(X | Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx.$$

If we set $g(y) = \mathbb{E}(X | Y = y)$, then the random variable $g(Y) = \mathbb{E}(X | Y)$ is the conditional expectation of X given Y and has the same properties as given for the conditional expectation in the discrete case.

Example 4.6 Consider the joint density for X and Y given by

$$f(x, y) = \begin{cases} 8xy & \text{for } 0 \leq x \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$



Here, (X, Y) are distributed over the upper half of the unit square as illustrated in the diagram. You should check that this is indeed a joint p.d.f. in that it integrates to 1 over the region. Compute the marginal densities of X and Y ,

$$f_X(x) = \int_x^1 8xy \, dy = 4x(1 - x^2) \quad \text{and} \quad f_Y(y) = \int_0^y 8xy \, dx = 4y^3,$$

for $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Calculate that

$$\mathbb{E} X = \int_0^1 x f_X(x) dx = \int_0^1 4x^2(1-x^2) dx = 4 \left(\frac{1}{3} - \frac{1}{5} \right) = \frac{8}{15}$$

and similarly $\mathbb{E} Y = \frac{4}{5}$. The conditional densities are

$$f_{X|Y}(x|y) = \frac{8xy}{4y^3} = \frac{2x}{y^2} \quad \text{and} \quad f_{Y|X}(y|x) = \frac{8xy}{4x(1-x^2)} = \frac{2y}{1-x^2},$$

for $0 \leq x \leq y \leq 1$. We then have

$$\mathbb{E}(X|Y=y) = \int_0^y x \frac{2x}{y^2} dx = \frac{2y}{3} \quad \text{and} \quad \mathbb{E}(Y|X=x) = \int_x^1 y \frac{2y}{1-x^2} dx = \frac{2(1-x^3)}{3(1-x^2)}.$$

We see that $\mathbb{E}(X|Y) = 2Y/3$ and $\mathbb{E}(Y|X) = 2(1-X^3)/(3(1-X^2))$. Check that we have $\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}X$, and $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}Y$. $\square$

The joint distribution function and density function extends to any number of random variables, in the obvious way. For random variables $X_1, \dots, X_n$, the joint distribution function is

$$\begin{aligned} F(x_1, \dots, x_n) &= \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) \quad \text{for } -\infty < x_i < \infty, 1 \leq i \leq n, \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(u_1, \dots, u_n) du_1 \dots du_n, \end{aligned}$$

where $f(u_1, \dots, u_n)$ is the joint probability density function. Note that

$$f(x_1, \dots, x_n) = \frac{\partial^n F}{\partial x_1 \cdots \partial x_n}.$$

The expectation of a function of $X_1, \dots, X_n$ is computed as

$$\mathbb{E} g(X_1, \dots, X_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \dots dx_n.$$

Independence for continuous random variables may be defined similarly to the discrete case. Random variables $X_1, \dots, X_n$ are **independent** if

$$\mathbb{P}(X_1 \in S_1, X_2 \in S_2, \dots, X_n \in S_n) = \mathbb{P}(X_1 \in S_1) \mathbb{P}(X_2 \in S_2) \cdots \mathbb{P}(X_n \in S_n),$$

for all $S_i \subseteq \Omega_{X_i}$, $1 \leq i \leq n$; this is equivalent to each of the statements that the joint distribution function

$$F(x_1, x_2, \dots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n), \quad \text{for all } x_i, 1 \leq i \leq n,$$

Infact, $F(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod \mathbb{P}(X_i \leq x_i)$
 $\mathbb{P}(X_i \leq x_i) = \mathbb{P}(X_i \leq x_i, \infty, \infty, \dots) = F(x_i, \infty, \infty, \dots) = F_{X_i}(x_i)$

factors into the product of the marginal distribution functions, F_{X_i} , and the joint probability density function

$$f(x_1, x_2, \dots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n), \quad \text{for all } x_i, 1 \leq i \leq n,$$

factors into the product of the marginal densities, f_{X_i} . It follows that if $X_1, \dots, X_n$ are independent then, for functions $g_1, \dots, g_n$,

$$\begin{aligned} \mathbb{E} \left(\prod_{i=1}^n g_i(X_i) \right) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{i=1}^n g_i(x_i) \right) f(x_1, x_2, \dots, x_n) dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{i=1}^n g_i(x_i) \right) f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n) dx_1 \cdots dx_n \\ &= \prod_{i=1}^n \left(\int_{-\infty}^{\infty} g_i(x_i) f_{X_i}(x_i) dx_i \right) = \prod_{i=1}^n (\mathbb{E}(g_i(X_i))), \end{aligned}$$

that is, as in the discrete case, the expectation of the product is the product of the expectations. This shows, as in the discrete case, that if X and Y are independent then $\text{Cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) = 0$.

Note that for independent random variables X, Y the conditional density of X given $Y = y$ is

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

which is of course just the unconditioned density function of X .

Example 4.7 Suppose that X and Y are independent random variables each with the $U[0, 1]$ distribution and that we wish to calculate $\mathbb{P}(X < Y)$. There are several ways that we might proceed. Firstly, the joint p.d.f. of X and Y is $f(x, y) = f_X(x)f_Y(y) = 1$ for $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Then,

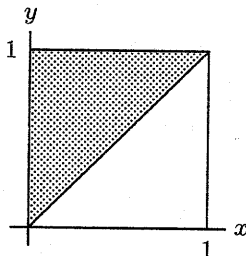
$$\begin{aligned} \mathbb{P}(X < Y) &= \int_{0 \leq x < y \leq 1} f(x, y) dx dy = \int_{x=0}^1 \int_{y=x}^1 dx dy \\ &= \int_{x=0}^1 (1 - x) dx = [x - x^2/2]_0^1 = \frac{1}{2}. \end{aligned}$$

Alternatively we could write, using the Law of Total Probability,

$$\begin{aligned} \mathbb{P}(X < Y) &= \int_0^1 \mathbb{P}(X < Y | Y = y) f_Y(y) dy = \int_0^1 \mathbb{P}(X < y) dy \\ &= \int_0^1 y dy = [y^2/2]_0^1 = \frac{1}{2}. \end{aligned}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} ; \text{ ok } P(B) = \int y(y)$$

Or, finally in this case we can argue graphically, since the joint distribution of X and Y is uniform over the unit square,



then $\mathbb{P}(X < Y)$ is just the area of the shaded region, which is $\frac{1}{2}$. $\square$

For independent random variables X and Y , the density function of $X + Y$ may be expressed in terms of the densities of X and Y as

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx; \quad (4.8)$$

this is known as the **convolution** of the two densities. It is derived from the corresponding statements involving distribution functions, when $F_{X+Y}(z) = \mathbb{P}(X + Y \leq z)$, which are

$$\begin{aligned} F_{X+Y}(z) &= \int_{-\infty}^{\infty} \mathbb{P}(X + Y \leq z \mid Y = y) f_Y(y)dy = \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X + Y \leq z \mid X = x) f_X(x)dx = \int_{-\infty}^{\infty} f_X(x)F_Y(z-x)dx. \end{aligned} \quad (4.9)$$

Then (4.8) is obtained by differentiating with respect to z either of the two expressions in (4.9).

Example 4.10 *Minimum of exponentials is exponential* Suppose that $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ are independent then consider the distribution of $\min(X, Y)$. Using the independence, we see that for $x \geq 0$,

$$\begin{aligned} \mathbb{P}(\min(X, Y) \leq x) &= 1 - \mathbb{P}(\min(X, Y) > x) = 1 - \mathbb{P}(X > x, Y > x) \\ &= 1 - \mathbb{P}(X > x)\mathbb{P}(Y > x) = 1 - e^{-\lambda x}e^{-\mu x} = 1 - e^{-(\lambda+\mu)x}, \end{aligned}$$

so that $\min(X, Y) \sim \text{Exp}(\lambda + \mu)$.

We may extend this, using induction on n , to see that if $X_1, \dots, X_n$ are independent, with $X_i \sim \text{Exp}(\lambda_i)$, then $\min_{1 \leq i \leq n} X_i \sim \text{Exp}(\lambda_1 + \dots + \lambda_n)$. In particular, when $X_1, \dots, X_n$ are i.i.d. with each $X_i \sim \text{Exp}(\lambda)$, then $\min_{1 \leq i \leq n} X_i \sim \text{Exp}(n\lambda)$. $\square$

Example 4.11 *Order statistics of a random sample* Independent, identically random variables $X_1, \dots, X_n$ each having the continuous distribution $F(x)$ are said to be a **random sample** from the distribution F . The values of these random variables arranged in increasing order are usually written as

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n-1)} \leq X_{(n)}.$$

The values $Y_i = X_{(i)}$ are said to be the **order statistics** of the sample. Thus, $Y_1 = \min_{1 \leq i \leq n} X_i$ is the smallest of the random variables, Y_2 is the second smallest and so on with $Y_n = \max_{1 \leq i \leq n} X_i$. As in the previous example, we may calculate the distribution of Y_1 ,

$$\begin{aligned} \mathbb{P}(Y_1 \leq x) &= \mathbb{P}\left(\min_{1 \leq i \leq n} X_i \leq x\right) = 1 - \mathbb{P}\left(\min_{1 \leq i \leq n} X_i > x\right) \\ &= 1 - \mathbb{P}(X_1 > x, \dots, X_n > x) = 1 - \prod_{i=1}^n \mathbb{P}(X_i > x) = 1 - (1 - F(x))^n. \end{aligned}$$

Then the p.d.f. of Y_1 is $n(1 - F(x))^{n-1} f(x)$, where $f(x) = F'(x)$ is the p.d.f. of the $\{X_i\}$. A similar calculation shows that for Y_n ,

$$\mathbb{P}(Y_n \leq x) = \mathbb{P}\left(\max_{1 \leq i \leq n} X_i \leq x\right) = (F(x))^n \quad \text{and its p.d.f. is } n(F(x))^{n-1} f(x).$$

We may also see that the joint p.d.f. of $Y_1, \dots, Y_n$ is given by

$$g(y_1, \dots, y_n) = \begin{cases} n! f(y_1) \cdots f(y_n) & \text{for } y_1 < \dots < y_n, \\ 0 & \text{otherwise.} \end{cases}$$

To see this consider the joint probabilities that $Y_i \in (y_i, y_i + dy_i)$, $1 \leq i \leq n$, and see that there are n choices from the $\{X_i\}$ for the smallest order statistic, $n - 1$ choices for the second smallest and so on to understand how the factor $n!$ in the expression for the joint density is obtained. $\square$

4.4 Moment generating functions

The **moment generating function** (m.g.f.) of a random variable X , with p.d.f. $f(x)$, is

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

defined for those values of θ for which the expectation is finite. Note that it is always defined for $\theta = 0$, and that $m(0) = 1$. When discussing moment generating functions we will assume that we are considering random variables for which the m.g.f. is defined for some non-trivial interval of values θ , (including 0). The m.g.f. plays the same role for more general random variables as the p.g.f. does for non-negative integer-valued random variables. Its importance stems from the following result, which we will not prove.

Theorem 4.12 *The moment generating function $m(\theta) = \mathbb{E}(e^{\theta X})$ determines the distribution of X uniquely provided it is defined for some open interval of values of θ .*

The name moment generating function stems from the following result.

Theorem 4.13 *If the moment generating function $m(\theta) = \mathbb{E}(e^{\theta X})$ is defined for some open interval of values of θ , then for each $r \geq 1$, $m^{(r)}(0) = \mathbb{E}(X^r)$, where $m^{(r)}$ is the r th derivative of m .*

Here, it is possible that $m(\theta)$ is not differentiable at $\theta = 0$ since it is possible that $m(\theta)$ is not defined for, say, $\theta > 0$, (or alternately for $\theta < 0$), but we may interpret $m^{(r)}(0)$ as $\lim_{\theta \uparrow 0} m^{(r)}(\theta)$ or $\lim_{\theta \downarrow 0} m^{(r)}(\theta)$, as appropriate, and the result is still true. We will not give a formal proof of Theorem 4.13, but to see intuitively why it holds, observe that

$$e^{\theta X} = 1 + \theta X + \frac{(\theta X)^2}{2!} + \frac{(\theta X)^3}{3!} + \dots,$$

so that, after taking expectations we see that

$$m(\theta) = 1 + \theta \mathbb{E}(X) + \frac{\theta^2 \mathbb{E}(X^2)}{2!} + \frac{\theta^3 \mathbb{E}(X^3)}{3!} + \dots;$$

now differentiate r times with respect to θ and set $\theta = 0$.

The other important application for moment generating functions is for studying sums of independent random variables since, if $X_1, \dots, X_n$ are independent random variables with m.g.f.s $m_{X_1}(\theta), \dots, m_{X_n}(\theta)$, respectively, then the m.g.f. of $X_1 + \dots + X_n$ is

$$m_{X_1 + \dots + X_n}(\theta) = \mathbb{E}(e^{\theta(X_1 + \dots + X_n)}) = \prod_{i=1}^n \mathbb{E}(e^{\theta X_i}) = \prod_{i=1}^n m_{X_i}(\theta),$$

just the product of the individual generating functions.

Example 4.14 *The Gamma distribution* A random variable X with p.d.f. $f(x) = e^{-\lambda x} \lambda^n x^{n-1} / ((n-1)!)$, for $x \geq 0$, ($f(x) = 0$ for $x < 0$), is said to have a Gamma distribution with parameters $\lambda > 0$ and integer $n \geq 1$, usually written $X \sim \Gamma(n, \lambda)$. Notice that the case $n = 1$ is the exponential distribution introduced previously. We need to check that the function f is indeed a p.d.f., that is, it integrates to 1, but this follows by integration by parts since, for $n > 1$,

$$I_n = \int_0^\infty e^{-\lambda x} \frac{\lambda^n x^{n-1}}{(n-1)!} dx = \int_0^\infty \frac{(\lambda x)^{n-1}}{(n-1)!} d(-e^{-\lambda x}) = \left[-e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} \right]_0^\infty + I_{n-1} = I_{n-1},$$

and $I_1 = 1$. The moment generating function of X , for $\theta < \lambda$, is

$$\begin{aligned} m(\theta) &= \mathbb{E}(e^{\theta X}) = \int_0^\infty e^{\theta x} e^{-\lambda x} \frac{\lambda^n x^{n-1}}{(n-1)!} dx \\ &= \left(\frac{\lambda}{\lambda - \theta} \right)^n \int_0^\infty e^{-(\lambda - \theta)x} \frac{(\lambda - \theta)^n x^{n-1}}{(n-1)!} dx = \left(\frac{\lambda}{\lambda - \theta} \right)^n, \end{aligned}$$

since the last integral is 1 by the above argument (replacing λ by $\lambda - \theta$). In particular, if $X \sim \text{Exp}(\lambda)$ then X has m.g.f. $\lambda/(\lambda - \theta)$. Then

$$m'(\theta) = \frac{n\lambda^n}{(\lambda - \theta)^{n+1}}, \quad \text{so that } \mathbb{E}(X) = m'(0) = \frac{n}{\lambda},$$

and similarly $\mathbb{E}(X^2) = m''(0) = n(n+1)/\lambda^2$, so that $\text{Var}(X) = n/\lambda^2$. Now if Y is independent of X and $Y \sim \Gamma(m, \lambda)$ the the m.g.f. of $X + Y$ is

$$\mathbb{E}(e^{\theta(X+Y)}) = \mathbb{E}(e^{\theta X}) \mathbb{E}(e^{\theta Y}) = \left(\frac{\lambda}{\lambda - \theta} \right)^n \left(\frac{\lambda}{\lambda - \theta} \right)^m = \left(\frac{\lambda}{\lambda - \theta} \right)^{n+m},$$

so that $X + Y \sim \Gamma(n + m, \lambda)$. Using induction, we may deduce that if $X_1, \dots, X_n$ are i.i.d. with $X_1 \sim \text{Exp}(\lambda)$, then $X_1 + \dots + X_n \sim \Gamma(n, \lambda)$. Note that this gives an alternate explanation of why for the Gamma distribution the mean and variance are n/λ and n/λ^2 , respectively. Note further that the Gamma distribution generalizes to non-integer parameter $\alpha > 0$ (replacing n) if $(n-1)!$ is replaced in the definition of the probability density by the Gamma function $\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx$. $\square$

Example 4.15 *The Normal distribution* Suppose that $X \sim N(\mu, \sigma^2)$, then the m.g.f. is

$$m(\theta) = \mathbb{E}(e^{\theta X}) = \int_{-\infty}^\infty e^{\theta x} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx,$$

but the argument of the exponential in the integral is

$$\theta x - \frac{(x - \mu)^2}{2\sigma^2} = \mu\theta + \frac{\theta^2\sigma^2}{2} - \frac{(x - \mu - \theta\sigma^2)^2}{2\sigma^2},$$

so that

$$m(\theta) = e^{\mu\theta + \theta^2\sigma^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x - \mu - \theta\sigma^2)/(2\sigma^2)} dx = e^{\mu\theta + \theta^2\sigma^2/2},$$

since the integrand is just the p.d.f. of the $N(\mu + \theta\sigma^2, \sigma^2)$ -distribution. We may check the fact that we established previously that a linear transformation of X has a normal distribution, that is $aX + b \sim N(a\mu + b, a^2\sigma^2)$, for constants a and b since the m.g.f. of $aX + b$ is

$$\mathbb{E} \left(e^{\theta(aX+b)} \right) = e^{b\theta} \mathbb{E} \left(e^{(a\theta)X} \right) = e^{b\theta} e^{a\theta\mu + a^2\theta^2\sigma^2/2} = e^{\theta(a\mu+b) + a^2\theta^2\sigma^2/2},$$

which has the required form. If $Y \sim N(\nu, \tau^2)$ is independent of X we see that the m.g.f. of $X + Y$ is

$$e^{\mu\theta + \theta^2\sigma^2/2} e^{\nu\theta + \theta^2\tau^2/2} = e^{(\mu+\nu)\theta + \theta^2(\sigma^2 + \tau^2)/2},$$

which is the m.g.f. of the $N(\mu + \nu, \sigma^2 + \tau^2)$ -distribution; we conclude that if we sum independent normally-distributed random variables we get a normally-distributed random variable—sum the means and sum the variances. $\square$

4.5 Transformations of random variables

We first consider the case of two random variables X, Y , with joint p.d.f. $f(x, y)$, and suppose that U and V are random variables which are functions of X and Y derived from a one-to-one transformation $(x, y) \mapsto (u, v)$, so that $U = a(X, Y)$, $V = b(X, Y)$, say, and moreover X and Y may be written as functions of U and V as $X = A(U, V)$ and $Y = B(U, V)$. In order to obtain the joint p.d.f. $g(u, v)$ of the pair U and V , recall the definition of the Jacobian

$$\frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u}.$$

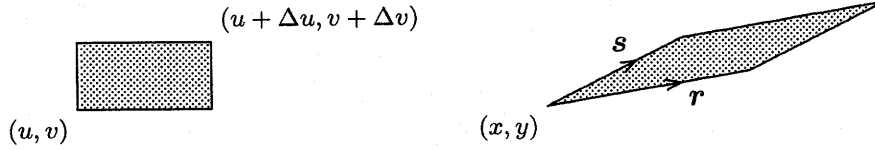
of the transformation $(u, v) \mapsto (x, y)$. Then the joint p.d.f. $g(u, v)$ is given by

$$g(u, v) = f(x, y) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|. \quad (4.16)$$

This follows from the fact that if a region S in the (x, y) -plane maps into the region T in the (u, v) -plane then we must have

$$\mathbb{P}((X, Y) \in S) = \iint_S f(x, y) dx dy = \iint_T g(u, v) du dv = \mathbb{P}((U, V) \in T).$$

The change-of-variable formula in multiple integration comes from the following idea: the element of area, which may be thought of as a rectangle in the (u, v) -plane with sides of length Δu and Δv , maps into a parallelogram in the (x, y) -plane bounded by vectors $\mathbf{r}$ and $\mathbf{s}$ (which we think of as being in $\mathbb{R}^3$) as illustrated,



where

$$\begin{aligned} \mathbf{r} &= (x(u + \Delta u, v) - x(u, v), y(u + \Delta u, v) - y(u, v), 0) \\ &\approx \Delta u \left(\frac{\partial x}{\partial u}, \frac{\partial y}{\partial u}, 0 \right) = \Delta u \left(\frac{\partial x}{\partial u} \mathbf{i} + \frac{\partial y}{\partial u} \mathbf{j} \right), \end{aligned}$$

and similarly, $\mathbf{s} \approx \Delta v \left(\frac{\partial x}{\partial v} \mathbf{i} + \frac{\partial y}{\partial v} \mathbf{j} \right)$; here $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are the standard basic unit vectors in $\mathbb{R}^3$. Then by the determinant rule, the cross product between $\mathbf{r}$ and $\mathbf{s}$ is

$$\mathbf{r} \times \mathbf{s} = \Delta u \Delta v \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & 0 \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & 0 \end{vmatrix} = \Delta u \Delta v \frac{\partial(x, y)}{\partial(u, v)} \mathbf{k}.$$

It follows that the area of the parallelogram is $|\mathbf{r} \times \mathbf{s}| = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| \Delta u \Delta v$, from which we see the relation (4.16).

Example 4.17 Suppose that X and Y are independent, identically distributed random variables each with the $\text{Exp}(\lambda)$ distribution. Let $U = X + Y$ and $V = X/(X + Y)$. The joint probability density function of X and Y is

$$f_{X,Y}(x, y) = \lambda^2 e^{-\lambda(x+y)}, \quad 0 < x < \infty, \quad 0 < y < \infty.$$

Then we have $u = x + y$ and $v = x/(x + y)$, so solving for x and y in terms of u and v gives

$$x = uv, \quad y = u(1 - v), \quad \text{for } 0 < u < \infty, \quad 0 < v < 1.$$

We calculate the Jacobian,

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 1 - v & -u \end{vmatrix} = -vu - u(1 - v) = -u.$$

The joint density of U and V is then

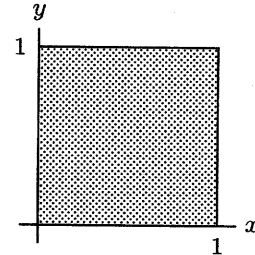
$$g_{U,V}(u, v) = f_{X,Y}(uv, u(1 - v)) |J| = \lambda^2 u e^{-\lambda u}, \quad \text{for } 0 < u < \infty, \quad 0 < v < 1.$$

We see that this can be viewed as the product of the two probability densities, $g_U(u) = \lambda^2 u e^{-\lambda u}$, which is the density of the $\Gamma(2, \lambda)$ distribution, and $g_V(v) = 1$, which is the density of the $U(0, 1)$ distribution; we can conclude that U and V are independent with g_U and g_V as their marginal density functions.

Whenever we calculate a joint probability density function in this way and we see that it splits into a product of functions of the variables separately in such a way that we may normalize the functions so that they become the marginal probability densities of the two random variables, then we may conclude that the random variables are independent. $\square$

Example 4.18 Suppose that X and Y have joint p.d.f. given by

$$f(x, y) = \begin{cases} 4xy & \text{for } 0 < x < 1, 0 < y < 1, \\ 0 & \text{otherwise,} \end{cases}$$

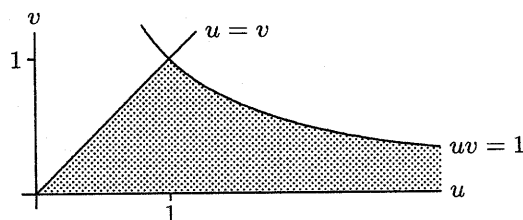


and that $U = X/Y$ and $V = XY$. Then $x = \sqrt{uv}$ and $y = \sqrt{v/u}$, and the Jacobian is

$$\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2}\sqrt{\frac{v}{u}} & \frac{1}{2}\sqrt{\frac{u}{v}} \\ -\frac{1}{2}\frac{\sqrt{v}}{u^{3/2}} & \frac{1}{2}\frac{1}{\sqrt{uv}} \end{vmatrix} = \frac{1}{4u} + \frac{1}{4u} = \frac{1}{2u}.$$

We see from (4.17) that the joint density of U and V (when it is non-zero) is then of the form $2v/u$; however, U and V are not independent since the region over which the density is positive does not allow the joint density to split into the product of the marginal densities. We have

$$g(u, v) = \begin{cases} \frac{2v}{u} & \text{for } 0 < uv < 1, 0 < v/u < 1, \\ 0 & \text{otherwise,} \end{cases}$$



which is concentrated on the region shown. We may calculate the marginal density of U ,

$$\text{for } u \leq 1, \quad g_U(u) = \int_0^1 g(u, v) dv = \int_0^u \frac{2v}{u} dv = \left[\frac{v^2}{u} \right]_0^u = u, \text{ while}$$

$$\text{for } u > 1, \quad g_U(u) = \int_0^1 g(u, v) dv = \int_0^{1/u} \frac{2v}{u} dv = \left[\frac{v^2}{u} \right]_0^{1/u} = \frac{1}{u^3}.$$

Calculating the marginal density of V , for $0 < v < 1$, we obtain

$$g_V(v) = \int_0^\infty g(u, v) du = \int_v^{1/v} \frac{2v}{u} du = [2v \log u]_v^{1/v} = -4v \log v,$$

and we see that $g(u, v) \neq g_U(u)g_V(v)$. □

Example 4.19 Sums and Convolution Suppose that X and Y have joint probability density function $f(x, y)$ and let $U = X + Y$ and $V = Y$, so that $X = U - V$ and $Y = V$.

The Jacobian

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1,$$

so that the joint density of U and V is $g(u, v) = f(u - v, v)$. We may then derive the marginal density of $X + Y$ as

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f(u - v, v) dv.$$

In the particular case that X and Y are independent we have $f(x, y) = f_X(x)f_Y(y)$ and we derive the formula for the convolution of two independent random variables

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(u - v)f_Y(v)dv,$$

that we had derived previously in (4.10). □

Example 4.20 Suppose that X and Y are i.i.d. each with the $N(0, 1)$ distribution and let $D = X^2 + Y^2$ and $\Theta = \tan^{-1}(Y/X)$. The joint density function of X and Y is

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

Then for $d = x^2 + y^2$ and $\theta = \tan^{-1}(y/x)$, consider the Jacobian

$$J = \begin{vmatrix} \frac{\partial d}{\partial x} & \frac{\partial d}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{vmatrix} = \begin{vmatrix} 2x & 2y \\ -\frac{y}{x^2 + y^2} & \frac{x}{x^2 + y^2} \end{vmatrix} = 2,$$

so the Jacobian of the inverse transformation is $\frac{1}{2}$. It follows that the joint density of D and Θ is

$$g(d, \theta) = \frac{1}{4\pi} e^{-d/2}, \quad 0 \leq d < \infty, \quad 0 \leq \theta \leq 2\pi,$$

which we may see can be expressed as the product of the marginal densities of D and Θ as $g(d, \theta) = g_D(d)g_\Theta(\theta)$, where

$$g_D(d) = \frac{1}{2}e^{-d/2}, \quad 0 \leq d < \infty, \quad \text{and} \quad g_\Theta(\theta) = \frac{1}{2\pi}, \quad 0 \leq \theta \leq 2\pi.$$

This means that $D \sim \text{Exp}(\frac{1}{2})$ and $\Theta \sim U[0, 2\pi]$ and they are independent random variables. This suggests a way of simulating $N(0, 1)$ random variables. Take U_1 and U_2 as

independent $U[0, 1]$ random variables. Then $D = -2 \log(U_1)$ has the $\text{Exp}(\frac{1}{2})$ distribution, while $\Theta = 2\pi U_2$ has the $U[0, 2\pi]$ distribution and we see that

$$X = \sqrt{D} \cos \Theta = \sqrt{-2 \log U_1} \cos(2\pi U_2) \quad \text{and} \quad Y = \sqrt{D} \sin \Theta = \sqrt{-2 \log U_1} \sin(2\pi U_2),$$

are independent standard normals. $\square$

We may generalize these ideas to one-to-one transformations of n random variables. Suppose that $X_1, \dots, X_n$ are random variables with joint probability density function $f(x_1, \dots, x_n)$ and that the random variables $U_1, \dots, U_n$ are given as functions $U_i = a_i(X_1, \dots, X_n)$ which we can invert so that $X_i = A_i(U_1, \dots, U_n)$. The Jacobian of the transformation is

$$\frac{\partial(x_1, x_2, \dots, x_n)}{\partial(u_1, u_2, \dots, u_n)} = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \dots & \frac{\partial x_1}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \dots & \frac{\partial x_n}{\partial u_n} \end{vmatrix},$$

and the joint probability density function of $U_1, \dots, U_n$ is obtained by setting

$$g(u_1, \dots, u_n) = f(x_1, \dots, x_n) \left| \frac{\partial(x_1, x_2, \dots, x_n)}{\partial(u_1, u_2, \dots, u_n)} \right|.$$

In particular, if the $\{X_i\}$ are just a linear transformation of the $\{U_j\}$, so that in vector notation

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \mathbf{A}\mathbf{U} = \mathbf{A} \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix},$$

where $\mathbf{A}$ is an $n \times n$ matrix, then the Jacobian of the transformation is $\det \mathbf{A}$. We then have $g(\mathbf{u}) = f(\mathbf{A}\mathbf{u}) |\det \mathbf{A}|$.

Example 4.21 Suppose that $X_1, \dots, X_n$ are independent identically distributed random variables with $X_i \sim \text{Exp}(\lambda)$, for each i , $1 \leq i \leq n$. Let $Y_1, \dots, Y_n$ be the order statistics of the $\{X_i\}$ so that $Y_1 = \min_i X_i$ is the smallest of the $\{X_i\}$, Y_2 is the second smallest, and so on, with $Y_n = \max_i X_i$. Think of $X_1, \dots, X_n$ representing the lifetimes of n components which are plugged in simultaneously at time 0, then Y_1 is the time of the first failure, Y_2 is the time of the second failure and so on. Set

$$Z_1 = Y_1, \quad Z_2 = Y_2 - Y_1, \quad \dots \quad Z_n = Y_n - Y_{n-1},$$

so that

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = A \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \text{with} \quad A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix};$$

note that $\det A = 1$ and that $y_j = \sum_1^j z_i$ for each j . Recall that the joint p.d.f. of the order statistics $Y_1, \dots, Y_n$ is

$$g(y_1, \dots, y_n) = n! f(y_1) \cdots f(y_n) \quad \text{where} \quad f(x) = \lambda e^{-\lambda x},$$

we then obtain the joint p.d.f. of $Z_1, \dots, Z_n$ as

$$\begin{aligned} h(z_1, \dots, z_n) &= n! \lambda^n e^{-\lambda(y_1 + \cdots + y_n)} \\ &= n! \lambda^n e^{-\lambda(nz_1 + (n-1)z_2 + \cdots + 2z_{n-1} + z_n)} = \prod_{i=1}^n \left(\lambda(n-i+1) e^{-\lambda(n-i+1)z_i} \right). \end{aligned}$$

As the joint p.d.f. factors into n individual probability densities we conclude that the random variables $Z_1, \dots, Z_n$ are independent with $Z_i \sim \text{Exp}(\lambda(n-i+1))$.

Note that this puts together formally two ideas that we have seen from our previous consideration of the exponential distribution: the time until the first failure is the minimum of n i.i.d. exponential random variables, with parameter λ , and so has the exponential distribution with parameter $n\lambda$; by the lack of memory property of the exponential distribution, when the first failure of a component occurs, the time from then until the failure of the other components is exponential with the same parameter λ , so the time until the second failure is the minimum of $n-1$ i.i.d. exponentials and thus is exponential with parameter $(n-1)\lambda$, and so on. $\square$

4.6 Bivariate normal distribution

Recall that the random variable X has the $N(\mu, \sigma^2)$ -distribution if its probability density function is

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

and that $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{Var}(X)$. We say that random variables X and Y have a **bivariate normal distribution** (or **bivariate Gaussian distribution** or **joint normal distribution**) if their joint probability density function has the form

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma\tau\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu)^2}{\sigma^2} - 2\rho \frac{(x-\mu)(y-\nu)}{\sigma\tau} + \frac{(y-\nu)^2}{\tau^2} \right) \right]$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$ where the parameters satisfy $-\infty < \mu < \infty$, $-\infty < \nu < \infty$, $\sigma > 0$, $\tau > 0$ and $-1 < \rho < 1$. The first task is to check that this expression is indeed a joint density function in that it integrates to 1. By making the substitutions $u = (x - \mu)/(\sigma\sqrt{1-\rho^2})$ and $v = (y - \nu)/(\tau\sqrt{1-\rho^2})$, we have

$$\begin{aligned} I &= \int_{-\infty < x, y < \infty} f_{X,Y}(x,y) dx dy = \int_{-\infty < u, v < \infty} \frac{\sqrt{1-\rho^2}}{2\pi} e^{-\frac{1}{2}(u^2 - 2\rho uv + v^2)} du dv \\ &= \int_{-\infty < u, v < \infty} \frac{\sqrt{1-\rho^2}}{2\pi} e^{-\frac{1}{2}((u-\rho v)^2 + (1-\rho^2)v^2)} du dv. \end{aligned}$$

Now put $w = u - \rho v$ and $z = v\sqrt{1-\rho^2}$, or $u = w + \rho z/\sqrt{1-\rho^2}$ and $v = z/\sqrt{1-\rho^2}$, and calculate the Jacobian of this transformation

$$\frac{\partial(u,v)}{\partial(w,z)} = \begin{vmatrix} 1 & \frac{\rho}{\sqrt{1-\rho^2}} \\ 0 & \frac{1}{\sqrt{1-\rho^2}} \end{vmatrix} = \frac{1}{\sqrt{1-\rho^2}};$$

then we see that

$$I = \int_{-\infty < w, z < \infty} \frac{1}{2\pi} e^{-(w^2+z^2)/2} dw dz = \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw \right)^2 = 1.$$

Marginal distributions To see the relationship with the ordinary (univariate) normal distribution and to determine the marginal distributions, consider the random variables

$$U = X, \quad V = Y - \nu - \rho\tau(X - \mu)/\sigma.$$

Putting X and Y in terms of U and V gives

$$X = U, \quad Y = V + \nu + \rho\tau(U - \mu)/\sigma.$$

The Jacobian of this transformation is

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ \rho\tau/\sigma & 1 \end{vmatrix} = 1.$$

We may now calculate the joint density function of U and V , evaluated at (u, v) , as

$$\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(u-\mu)^2/(2\sigma^2)} \right) \left(\frac{1}{\sqrt{2\pi}\tau\sqrt{1-\rho^2}} e^{-v^2/(2\tau^2(1-\rho^2))} \right),$$

and we recognize these two expressions, the first in u is the density of the $N(\mu, \sigma^2)$ distribution, and the second in v is the density of the $N(0, \tau^2(1-\rho^2))$ distribution, and moreover, because the joint density factors into the product of these two densities, U and V are independent random variables. We conclude that the marginal distribution of X is $N(\mu, \sigma^2)$ and, by the symmetry of the joint density of X and Y , we can see that the marginal density of Y is $N(\nu, \tau^2)$. To interpret the remaining parameter ρ , calculate

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(U, V + \nu + \rho\tau(U - \mu)/\sigma) \\ &= \text{Cov}(U, V) + \text{Cov}(U, \rho\tau(U - \mu)/\sigma), \quad \text{since } \nu \text{ is constant,} \\ &= \text{Cov}(U, \rho\tau(U - \mu)/\sigma), \quad \text{since } U \text{ and } V \text{ are independent,} \\ &= \rho\tau\text{Var}(U)/\sigma = \rho\sigma\tau = \rho\sqrt{\text{Var}(X)\text{Var}(Y)}. \end{aligned}$$

Thus the parameter $\rho = \text{Corr}(X, Y)$ is the correlation coefficient of the random variables X and Y . We may see immediately that

$$f_{X,Y}(x, y) = \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \right) \left(\frac{1}{\sqrt{2\pi}\tau} e^{-(y-\nu)^2/(2\tau^2)} \right) = f_X(x)f_Y(y),$$

for all x and y , if and only if $\rho = 0$, or equivalently if and only if $\text{Cov}(X, Y) = 0$. Thus random variables which have a joint normal distribution are independent if and only if their covariance is zero. Recall that in general the covariance between random variables being zero does not imply independence of the random variables, we see here the important and useful property that the covariance being zero is sufficient to show independence for normally distributed variables.

Conditional distributions We may calculate the conditional density of one of the random variables Y , say, given the value of the other variable $X = x$, that is, the density $f_{Y|X}(y | x) = f_{X,Y}(x, y)/f_X(x)$, which equals

$$\begin{aligned} & \frac{\exp \left[-\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu)^2}{\sigma^2} - 2\rho \frac{(x-\mu)(y-\nu)}{\sigma\tau} + \frac{(y-\nu)^2}{\tau^2} \right) \right]}{2\pi\sigma\tau\sqrt{1-\rho^2}} \bigg/ \frac{\exp \left[-\frac{(x-\mu)^2}{2\sigma^2} \right]}{\sigma\sqrt{2\pi}} \\ &= \exp \left[-\frac{1}{2(1-\rho^2)} \left(\frac{\rho^2(x-\mu)^2}{\sigma^2} - 2\rho \frac{(x-\mu)(y-\nu)}{\sigma\tau} + \frac{(y-\nu)^2}{\tau^2} \right) \right] \bigg/ \tau\sqrt{2\pi(1-\rho^2)} \\ &= \exp \left[-\frac{1}{2\tau^2(1-\rho^2)} (y-\nu-\rho\tau(x-\mu)/\sigma)^2 \right] \bigg/ \tau\sqrt{2\pi(1-\rho^2)}. \end{aligned}$$

We recognize this last expression as being the density (in y) of the normal distribution with mean $\nu + \rho\tau(x - \mu)/\sigma$ and variance $\tau^2(1 - \rho^2)$, so that, in shorthand notation,

$$Y | X \sim N(\nu + \rho\tau(X - \mu)/\sigma, \tau^2(1 - \rho^2)).$$

Notice that the conditional expectation of Y given X , which is

$$\mathbb{E}(Y | X) = \nu + \rho\tau(X - \mu)/\sigma,$$

depends on X , but the variance of Y conditional on X is the constant $\tau^2(1 - \rho^2)$, which is less than the unconditioned variance of Y , that is τ^2 .

Linear transformations A further property that you might wish to check is that if X and Y have a joint normal distribution and we define random variables R and S by

$$\begin{pmatrix} R \\ S \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} \theta \\ \phi \end{pmatrix},$$

where a, b, c, d, θ and ϕ are constants with $ad \neq bc$, then R and S have a joint normal distribution, so that normal distributions are preserved under linear transformations. You should check that the condition $ad \neq bc$ is needed to ensure that $|\text{Corr}(R, S)| \neq 1$; even if this condition does not hold, the random variables R and S will individually have normal distributions but their correlation coefficient will be 1 or -1.

Multivariate normal distribution We may generalize the above to define the joint normal distribution for n random variables. Suppose that $Z_1, \dots, Z_n$ are i.i.d. random

variables each with the standard $N(0, 1)$ distribution. Suppose that A is a $n \times n$ invertible matrix and (using vector notation) suppose that

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} + A \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \boldsymbol{\mu} + A\mathbf{Z},$$

where $\mu_1, \dots, \mu_n$ are constants. Since each of the random variables $\{Z_j\}$ has mean zero, we see first that $\mathbb{E} X_i = \mu_i$, for each i . The joint probability density function of the components of $\mathbf{Z}$ at $\mathbf{z} = (z_1, \dots, z_n)^\top$ is

$$f(\mathbf{z}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\sum_{i=1}^n z_i^2/2} = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\mathbf{z}^\top \mathbf{z}/2}.$$

Writing $\mathbf{z} = A^{-1}(\mathbf{x} - \boldsymbol{\mu})$, the Jacobian of the transformation is $|\det A|^{-1}$, so that the joint density for $\mathbf{X}$ is

$$\begin{aligned} g(\mathbf{x}) &= \frac{1}{|\det A|} f(A^{-1}(\mathbf{x} - \boldsymbol{\mu})) = \frac{1}{|\det A|} \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}(A^{-1}(\mathbf{x} - \boldsymbol{\mu}))^\top (A^{-1}(\mathbf{x} - \boldsymbol{\mu}))} \\ &= \frac{1}{|\det A|} \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (A^{-1})^\top A^{-1}(\mathbf{x} - \boldsymbol{\mu})} \\ &= \frac{1}{\sqrt{|\det V|}} \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top V^{-1}(\mathbf{x} - \boldsymbol{\mu})}, \quad \text{where } V = AA^\top. \end{aligned} \quad (4.22)$$

To interpret the matrix V we see that for any pair (i, j) , $1 \leq i, j \leq n$,

$$\begin{aligned} \text{Cov}(X_i, X_j) &= \mathbb{E}((X_i - \mu_i)(X_j - \mu_j)) = \mathbb{E}\left(\left(\sum_r A_{ir} Z_r\right)\left(\sum_s A_{js} Z_s\right)\right) \\ &= \sum_r A_{ir} A_{jr} = (AA^\top)_{ij} = V_{ij}, \end{aligned}$$

so that the entries of the matrix V are the covariances between the components of the random vector $\mathbf{X}$. Any joint density of the form (4.22) is a **multivariate normal distribution** with mean $\boldsymbol{\mu}$ and **covariance matrix** V , usually written $N(\boldsymbol{\mu}, V)$.

Notice that V is a symmetric matrix and it is positive definite in that $\mathbf{x}^\top V \mathbf{x} > 0$ for all vectors $\mathbf{x} \neq 0$; this follows because $\mathbf{x}^\top V \mathbf{x} = \|A^\top \mathbf{x}\|^2 > 0$, since A is invertible.

Furthermore, in the case when $n = 2$ and X and Y have the bivariate normal distribution described above we see that if, for any angle θ , we take A to be the matrix

$$A = \begin{pmatrix} \sigma \cos(\theta + \cos^{-1} \rho) & \sigma \sin(\theta + \cos^{-1} \rho) \\ \tau \cos \theta & \tau \sin \theta \end{pmatrix}$$

we see that

$$AA^T = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} = V,$$

and

$$A^{-1} \begin{pmatrix} X - \mu \\ Y - \nu \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},$$

where Z_1 and Z_2 are independent random variables each with the standard normal distribution, $N(0, 1)$.

4.7 Multivariate moment generating functions

For random variables $X_1, \dots, X_n$ and real numbers $\theta_1, \dots, \theta_n$ set $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$ and $\mathbf{X} = (X_1, \dots, X_n)^T$, then we define

$$m(\boldsymbol{\theta}) = m(\theta_1, \dots, \theta_n) = \mathbb{E} (e^{\theta_1 X_1 + \dots + \theta_n X_n}) = \mathbb{E} (e^{\boldsymbol{\theta}^T \mathbf{X}}),$$

to be the **joint moment generating function** of the random variables. The moment generating function is only defined for those $\boldsymbol{\theta}$ for which $m(\boldsymbol{\theta}) < \infty$. The properties of the multivariate generating function are similar to those we have seen previously for the moment generating function of a single random variable.

Properties of $m(\boldsymbol{\theta})$

1. Provided $m(\boldsymbol{\theta})$ is finite for a non-trivial range of θ_i for each i , then $m(\boldsymbol{\theta})$ determines the joint distribution of $X_1, \dots, X_n$.
2. We may determine moments of the X_i from partial derivatives of m ,

$$\left. \frac{\partial^r m}{\partial \theta_i^r} \right|_{\boldsymbol{\theta}=0} = \mathbb{E} (X_i^r) \quad \text{and} \quad \left. \frac{\partial^{r+s} m}{\partial \theta_i^r \partial \theta_j^s} \right|_{\boldsymbol{\theta}=0} = \mathbb{E} (X_i^r X_j^s), \quad \text{for } r \geq 1, s \geq 1.$$

In particular, we may calculate covariances as

$$\text{Cov} (X_i, X_j) = \mathbb{E} (X_i X_j) - (\mathbb{E} X_i) (\mathbb{E} X_j) = \left[\frac{\partial^2 m}{\partial \theta_i \partial \theta_j} - \left(\frac{\partial m}{\partial \theta_i} \right) \left(\frac{\partial m}{\partial \theta_j} \right) \right]_{\boldsymbol{\theta}=0}.$$

3. The moment generating function factors

$$m(\boldsymbol{\theta}) = \prod_{i=1}^n (\mathbb{E} (e^{\theta_i X_i})),$$

into the product of the moment generating functions of the individual random variables if and only if $X_1, \dots, X_n$ are independent. $\square$

For the particular case of random variables X and Y having the bivariate normal distribution considered in the previous section, then we may use the form for the moment generating function of the normal distribution

$$\mathbb{E} (e^{\theta X}) = e^{\theta\mu + \frac{1}{2}\theta^2\sigma^2}, \quad \text{when } X \sim N(\mu, \sigma^2),$$

and the form of the conditional distribution of Y given X to calculate (here, to avoid subscripts take $\theta_1 = \theta$ and $\theta_2 = \phi$),

$$\begin{aligned} \mathbb{E} (e^{\theta X + \phi Y}) &= \mathbb{E} (\mathbb{E} (e^{\theta X + \phi Y} \mid X)) = \mathbb{E} (e^{\theta X} \mathbb{E} (e^{\phi Y} \mid X)) \\ &= \mathbb{E} \left(e^{\theta X} e^{\phi \mathbb{E}(Y|X) + \frac{1}{2}\phi^2 \text{Var}(Y|X)} \right) = \mathbb{E} \left(e^{\theta X + \phi(\nu + \rho\tau(X - \mu)/\sigma) + \frac{1}{2}\phi^2\tau^2(1 - \rho^2)} \right) \\ &= e^{\phi(\nu - \mu\rho\tau/\sigma) + \frac{1}{2}\phi^2\tau^2(1 - \rho^2)} \mathbb{E} \left(e^{(\theta + \phi\rho\tau/\sigma)X} \right) \\ &= e^{\phi(\nu - \mu\rho\tau/\sigma) + \frac{1}{2}\phi^2\tau^2(1 - \rho^2)} e^{(\theta + \phi\rho\tau/\sigma)\mu + \frac{1}{2}\sigma^2(\theta + \phi\rho\tau/\sigma)^2} \\ &= e^{\theta\mu + \phi\nu + \frac{1}{2}(\theta^2\sigma^2 + \phi^2\tau^2 + 2\theta\phi\rho\sigma\tau)}. \end{aligned}$$

We see that this factors into the product

$$\left(e^{\theta\mu + \frac{1}{2}\theta^2\sigma^2} \right) \left(e^{\phi\nu + \frac{1}{2}\phi^2\tau^2} \right)$$

of the individual generating functions of X and Y for all θ and ϕ if and only if $\rho = 0$; as we have seen previously, the random variables are independent in this case if and only if their covariance is zero.

15 May 2007

5. INEQUALITIES, LIMIT THEOREMS AND GEOMETRIC PROBABILITY

5.1 Inequalities

Suppose that $X \geq 0$ is a random variable taking non-negative values and that $c > 0$ is a constant. Then

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c},$$

is **Markov's inequality**. It follows because $X \geq c \Rightarrow \frac{X}{c} \geq 1 = I_{(X \geq c)}$; $X < c \Rightarrow \frac{X}{c} < 1 = I_{(X \geq c)}$; $X \geq 0, c > 0$

$$\mathbb{P}(X \geq c) = \mathbb{E}(I_{(X \geq c)}) \leq \mathbb{E}\left(\frac{X}{c} I_{(X \geq c)}\right) \leq \mathbb{E}\left(\frac{X}{c}\right) = \frac{\mathbb{E}(X)}{c}.$$

From this result we may deduce **Chebyshev's inequality**: for any random variable X and constant $c > 0$,

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}(X^2)}{c^2}.$$

This follows by observing that

$$\mathbb{P}(|X| \geq c) = \mathbb{P}(X^2 \geq c^2),$$

and applying Markov's inequality for the random variable $Y = X^2$ and constant c^2 . We should note the following points about Chebyshev's inequality:

1. The inequality is 'distribution free'; it holds for all random variables irrespective of the distribution of the random variable.
2. If $\mathbb{E}(X^2) \geq c^2$, then the inequality provides no bound on the probability.
3. If $\mathbb{E}(X^2) < c^2$, the inequality is the best possible in the sense that given c there is a random variable X for which the inequality holds with equality. To see this suppose that $d < c^2$ and let X be the random variable

$$X = \begin{cases} c & \text{with probability } \frac{d}{2c^2}, \\ -c & \text{with probability } \frac{d}{2c^2}, \\ 0 & \text{with probability } 1 - \frac{d}{c^2}. \end{cases}$$

Then $\mathbb{E}(X^2) = d$ and $\mathbb{P}(|X| \geq c) = d/c^2 = \mathbb{E}(X^2)/c^2$.

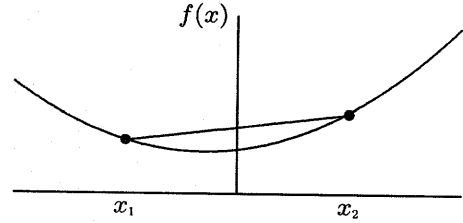
Suppose that $\phi : [0, \infty) \rightarrow [0, \infty)$ is a non-decreasing function, with $\phi(x) > 0$ for $x > 0$, then we obtain the **generalized Chebyshev's inequality**: for any random variable X and $c > 0$,

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}(\phi(|X|))}{\phi(c)}.$$

This follows in the same way by observing that $\mathbb{P}(|X| \geq c) \leq \mathbb{P}(\phi(|X|) \geq \phi(c))$, and using Markov's inequality with $Y = \phi(|X|)$. As an example, take $\phi(x) = x^4$ and we obtain $\mathbb{P}(|X| \geq c) \leq \mathbb{E}(X^4)/c^4$.

The next inequality involving random variables that we will consider is based on the concept of convexity. A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is **convex** if for all $x_1, x_2 \in \mathbb{R}$ and $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ with $\lambda_1 + \lambda_2 = 1$, we have

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2).$$



Thus a function is convex if the chord joining any two points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ on the graph of the function lies above the function between the points. It is easy to see that if $f(x)$ is a convex function then for $x_1 < x_2 < x_3$, the slope of the chord joining the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ is less than or equal to the slope of the chord joining the points $(x_2, f(x_2))$ and $(x_3, f(x_3))$; that is

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_2)}{x_3 - x_2}, \quad (5.1)$$

this follows from the definition of convexity, since

$$x_2 = \left(\frac{x_3 - x_2}{x_3 - x_1} \right) x_1 + \left(\frac{x_2 - x_1}{x_3 - x_1} \right) x_3,$$

so that

$$f(x_2) \leq \left(\frac{x_3 - x_2}{x_3 - x_1} \right) f(x_1) + \left(\frac{x_2 - x_1}{x_3 - x_1} \right) f(x_3), \quad (5.2)$$

and rearranging (5.2) gives (5.1). Furthermore, from (5.1), it is immediate that for points $x_1 < x_2 < x_3 < x_4$, we have

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_4) - f(x_3)}{x_4 - x_3}, \quad (5.3)$$

since we may apply (5.1) twice to obtain

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_2)}{x_3 - x_2} \leq \frac{f(x_4) - f(x_3)}{x_4 - x_3}. \quad (5.4)$$

It is the case that a function f is convex if and only if (5.1) holds for all choices of $x_1 < x_2 < x_3$.

Moreover, when f is differentiable then f being convex is equivalent to the derivative $f'(x)$ being non-decreasing (or $f''(x) \geq 0$); this may be seen by letting $x_2 \rightarrow x_1$ and $x_4 \rightarrow x_3$ in (5.3). With a similar argument, we may see that, when f is convex, then

$$f(y) - f(x) \geq (y - x)f'(x), \quad \text{for all } x, y. \quad (5.5)$$

Now if f is a convex function, for each $n \geq 2$, and for any points $x_1, \dots, x_n \in \mathbb{R}$ and any $\lambda_i \geq 0$, $1 \leq i \leq n$, with $\lambda_1 + \dots + \lambda_n = 1$, we have

$$f(\lambda_1 x_1 + \dots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \dots + \lambda_n f(x_n). \quad (5.6)$$

The proof of the inequality (5.6) is by induction on n . The case $n = 2$ is just the definition of convexity. So assume that (5.6) holds for any n points $x_1, \dots, x_n \in \mathbb{R}$ and any $\lambda_i \geq 0$, $1 \leq i \leq n$, with $\lambda_1 + \dots + \lambda_n = 1$, and suppose that we are given $x_1, \dots, x_{n+1} \in \mathbb{R}$ and $\lambda_i \geq 0$, $1 \leq i \leq n+1$, with $\lambda_1 + \dots + \lambda_{n+1} = 1$. We may assume that $\lambda_i > 0$ each i , otherwise the result follows by the inductive step immediately. Then, by first using the case $n = 2$ and then the inductive hypothesis, we have

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left((1 - \lambda_{n+1}) \sum_{i=1}^n \left(\frac{\lambda_i}{1 - \lambda_{n+1}}\right) x_i + \lambda_{n+1} x_{n+1}\right) \\ &\leq (1 - \lambda_{n+1}) f\left(\sum_{i=1}^n \left(\frac{\lambda_i}{1 - \lambda_{n+1}}\right) x_i\right) + \lambda_{n+1} f(x_{n+1}), \\ &\leq (1 - \lambda_{n+1}) \left(\sum_{i=1}^n \left(\frac{\lambda_i}{1 - \lambda_{n+1}}\right) f(x_i)\right) + \lambda_{n+1} f(x_{n+1}) = \sum_{i=1}^{n+1} \lambda_i f(x_i), \end{aligned}$$

completing the induction.

Jensen's Inequality For a random variable X and a convex function f ,

$$f(\mathbb{E} X) \leq \mathbb{E} f(X).$$

The proof of Jensen's Inequality in the case when X takes on just finitely many values $x_1, \dots, x_n$, with probabilities $p_i = \mathbb{P}(X = x_i)$, $1 \leq i \leq n$, with $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$, is just a restatement of (5.6), since

$$f(\mathbb{E} X) = f\left(\sum_{i=1}^n p_i x_i\right) \leq \sum_{i=1}^n p_i f(x_i) = \mathbb{E} f(X).$$

In general, for any random variable, we may use (5.5), to see that

$$f(X) - f(\mathbb{E} X) \geq (X - \mathbb{E} X) f'(\mathbb{E} X),$$

and taking the expectation of both sides we see that

$$\mathbb{E} f(X) - f(\mathbb{E} X) \geq \mathbb{E} (X - \mathbb{E} X) f'(\mathbb{E} X) = 0,$$

which gives the result.

Example 5.7 Arithmetic-Geometric Mean Inequality For positive real numbers $x_1, \dots, x_n$,

$$\left(\prod_{i=1}^n x_i\right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^n x_i.$$

This follows by Jensen's inequality applied to the convex function $f(x) = -\log x$, and the random variable X which takes the value x_i with probability $\frac{1}{n}$, so that

$$-\frac{1}{n} \sum_{i=1}^n \log x_i = \mathbb{E} (-\log(X)) \geq -\log(\mathbb{E} X) = -\log\left(\frac{1}{n} \sum_{i=1}^n x_i\right)$$

from which we see that

$$\log\left(\left(\prod_{i=1}^n x_i\right)^{1/n}\right) \leq \log\left(\frac{1}{n} \sum_{i=1}^n x_i\right),$$

which gives the result, since $\log$ is an increasing function. □

5.2 Weak Law of Large Numbers

Theorem 5.8 Weak Law of Large Numbers *Let $X_1, X_2, \dots$ be independent, identically distributed random variables with $\mathbb{E} X_1 = \mu$ and $\text{Var}(X_1) < \infty$. For any constant $\epsilon > 0$,*

$$\mathbb{P} \left(\left| \frac{X_1 + \dots + X_n}{n} - \mu \right| > \epsilon \right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Proof. By Chebyshev's inequality we have

$$\begin{aligned} \mathbb{P} \left(\left| \frac{X_1 + \dots + X_n}{n} - \mu \right| > \epsilon \right) &\leq \frac{1}{\epsilon^2} \mathbb{E} \left(\frac{X_1 + \dots + X_n}{n} - \mu \right)^2 \\ &= \frac{1}{n^2 \epsilon^2} \mathbb{E} (X_1 + \dots + X_n - n\mu)^2 \\ &= \frac{1}{n^2 \epsilon^2} \text{Var} (X_1 + \dots + X_n) = \frac{n}{n^2 \epsilon^2} \text{Var} (X_1) \\ &= \frac{1}{n \epsilon^2} \text{Var} (X_1) \rightarrow 0, \end{aligned}$$

as required. □

Notes 1. The statement in Theorem 5.8 is normally referred to by saying that the random variable $(X_1 + \dots + X_n)/n$ ‘converges in probability’ to μ , written

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{P} \mu, \quad \text{as } n \rightarrow \infty.$$

2. This result should be distinguished from the Strong Law of Large Numbers which states that

$$\mathbb{P} \left(\frac{X_1 + \dots + X_n}{n} \rightarrow \mu, \quad \text{as } n \rightarrow \infty \right) = 1.$$

As the name implies, the Strong Law of Large Numbers implies the Weak Law. The mode of convergence in the Strong Law is referred to as ‘convergence with probability one’ or ‘almost sure convergence’.

3. Notice that the requirement that the random variables in the Weak Law be independent is not one that we may dispense with. For example, suppose that $\Omega = \{\omega_1, \omega_2\}$ has just two points and let $X_n(\omega_1) = 1$ and $X_n(\omega_2) = 0$ for each n , so that the random

variables are identically distributed, but not of course independent. Let $p = \mathbb{P}(\{\omega_1\}) = 1 - \mathbb{P}(\{\omega_2\})$, where $0 < p < 1$, then $\mathbb{E} X_1 = p$, and we have

$$\frac{X_1(\omega_1) + \cdots + X_n(\omega_1)}{n} = 1 \quad \text{and} \quad \frac{X_1(\omega_2) + \cdots + X_n(\omega_2)}{n} = 0, \quad \text{for all } n,$$

so that the conclusion of Theorem 5.8 cannot hold.

4. By giving a more refined argument it is possible to dispense with the requirement in the statement of the Theorem that $\text{Var}(X_1) < \infty$. The conclusion still holds provided $\mathbb{E} |X_1| < \infty$.

5. It should be noticed that the Weak Law of Large Numbers is ‘distribution free’ in that the particular distribution of the summands $\{X_i\}$ only influences the result through the mean, μ , (and, in the form we have stated it, through the fact that the variance is finite) but otherwise the underlying distribution does not enter the conclusion of the Theorem.

6. The Weak Law of Large Numbers underlies the ‘frequentist’ interpretation of probability. Suppose that we have independent repetitions of an experiment, and we set $X_i = 1$ if a particular outcome occurs on the i th repetition (e.g., ‘Heads’), and $X_i = 0$, otherwise (e.g., ‘Tails’). Then $\mathbb{E} X_i = p$, say, where $p = \mathbb{P}(X_i = 1)$ is the probability of the outcome. Then $(X_1 + \cdots + X_n)/n$ is the average number of occurrences of the outcome in n repetitions and this converges in the above sense to $\mathbb{E} X_i = p$; thus the probability p is the long-run proportion of times that the outcome occurs.

5.3 Central Limit Theorem

Theorem 5.9 Central Limit Theorem *Let $X_1, X_2, \dots$ be independent, identically distributed random variables with $\mathbb{E} X_1 = \mu$ and $\sigma^2 = \text{Var}(X_1)$, where $0 < \sigma^2 < \infty$. For any x , $-\infty < x < \infty$,*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy = \Phi(x),$$

which is the distribution function of the standard $N(0, 1)$ distribution.

Sketch of Proof: We will illustrate why the result of the Theorem holds by using moment generating functions in the case when the moment generating function of the $\{X_i\}$, $m(\theta) =$

$\mathbb{E}(e^{\theta X_1})$, satisfies $m(\theta) < \infty$ for a non-trivial range of values of θ (that is, an open interval which necessarily contains the point $\theta = 0$). It should be noted that this condition on the moment generating function is not necessary for the result to hold. Let

$$Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}, \quad \text{and} \quad m_n(\theta) = \mathbb{E}(e^{\theta Y_n}),$$

then we will show that as $n \rightarrow \infty$, $m_n(\theta) \rightarrow e^{\theta^2/2}$, which is the moment generating function of the $N(0, 1)$ distribution. This is sufficient to establish the conclusion of the Theorem, although we will not prove that in this course. We now have

$$m_n(\theta) = \mathbb{E}\left(e^{\theta(X_1 + \cdots + X_n - n\mu)/(\sigma\sqrt{n})}\right) = e^{-\theta\mu\sqrt{n}/\sigma} \mathbb{E}\left(e^{\theta(X_1 + \cdots + X_n)/(\sigma\sqrt{n})}\right)$$

and since the $\{X_i\}$ are independent and identically distributed, this

$$= e^{-\theta\mu\sqrt{n}/\sigma} \left[\mathbb{E}\left(e^{\theta X_1/(\sigma\sqrt{n})}\right) \right]^n = \left[e^{-\theta\mu/(\sigma\sqrt{n})} m\left(\frac{\theta}{\sigma\sqrt{n}}\right) \right]^n.$$

Expand the two terms using Taylor's Theorem to see that $m_n(\theta)$ equals

$$\left[\left(1 - \frac{\theta\mu}{\sigma\sqrt{n}} + \frac{\theta^2\mu^2}{2\sigma^2n} + O\left(\frac{1}{n^{3/2}}\right) \right) \left(1 + \frac{\theta}{\sigma\sqrt{n}} m'(0) + \frac{\theta^2}{2\sigma^2n} m''(0) + O\left(\frac{1}{n^{3/2}}\right) \right) \right]^n;$$

now, using the fact that $m'(0) = \mathbb{E} X_1 = \mu$, and $m''(0) = \mathbb{E}(X_1^2) = \sigma^2 + \mu^2$, this shows that

$$\begin{aligned} m_n(\theta) &= \left[1 - \frac{\theta^2\mu^2}{\sigma^2n} + \frac{\theta^2(\sigma^2 + \mu^2)}{2\sigma^2n} + \frac{\theta^2\mu^2}{2\sigma^2n} + O\left(\frac{1}{n^{3/2}}\right) \right]^n \\ &= \left[1 + \frac{\theta^2}{2n} + O\left(\frac{1}{n^{3/2}}\right) \right]^n \rightarrow e^{\theta^2/2}, \quad \text{as required.} \end{aligned} \quad \square$$

Notes 1. The mode of convergence described in Theorem 5.9 for the random variables

$$Y_n = (X_1 + \cdots + X_n - n\mu)/(\sigma\sqrt{n})$$

is known as 'convergence in distribution' and the conclusion is written as $Y_n \xrightarrow{D} Z$, where $Z \sim N(0, 1)$.

2. Note that, like the Weak Law of Large Numbers, the Central Limit Theorem is distribution free, in that the underlying distribution of the $\{X_i\}$ influences the form of the result only through the mean $\mu = \mathbb{E} X_1$ and variance $\sigma^2 = \text{Var}(X_1)$.

3. Note that in the Central Limit Theorem, by subtracting off at each stage the mean of the sum of the random variables $X_1 + \cdots + X_n$, that is $n\mu$, and dividing by its standard deviation, $\sigma\sqrt{n}$, we are ensuring that the random variable Y_n has $\mathbb{E} Y_n = 0$ and $\text{Var}(Y_n) = 1$, for each n .

Example 5.10 *Normal approximation to the binomial distribution* If the random variable $Y \sim \text{Bin}(n, p)$, we may think of the distribution of Y as being the same as that of the sum of n i.i.d. random variables each of which has the Bernoulli distribution; thus the random variable $(Y - np)/\sqrt{np(1-p)}$ has approximately the $N(0, 1)$ distribution for large n . Note that here p is being held fixed and $n \rightarrow \infty$, unlike the situation we described in the Poisson approximation to the binomial where $n \rightarrow \infty$ and $p \rightarrow 0$ in such a way that the product $np \rightarrow \lambda > 0$. $\square$

Example 5.11 *Normal approximation to the Poisson distribution* When the random variable $Y \sim \text{Poiss}(n)$, where $n \geq 1$ is an integer, we may think of Y as having the same distribution as that of the sum of n i.i.d. random variables each with the $\text{Poiss}(1)$ distribution. Thus $(Y - n)/\sqrt{n}$ has approximately the $N(0, 1)$ distribution for large n . The same conclusion is true for $Y \sim \text{Poiss}(\lambda)$ for non-integer λ ; that is, $(Y - \lambda)/\sqrt{\lambda}$ is approximately $N(0, 1)$ for λ large. $\square$

Example 5.12 *Opinion polls* Suppose that the proportion of voters in the population who vote Labour is p , where p is unknown. A random sample of n voters is taken and it is found that S voters in the sample vote Labour and we estimate p by S/n . We want to ensure that $|S/n - p| < \epsilon$, for some small given ϵ , with high probability, ≥ 0.95 , say. How large must n be? Note that $S \sim \text{Bin}(n, p)$, so that $\mathbb{E} S = np$ and $\text{Var}(S) = np(1-p)$. Then by the Central Limit Theorem, we require

$$\begin{aligned} \mathbb{P}\left(\left|\frac{S}{n} - p\right| < \epsilon\right) &= \mathbb{P}\left(-\epsilon\sqrt{\frac{n}{p(1-p)}} < \frac{S - np}{\sqrt{np(1-p)}} < \epsilon\sqrt{\frac{n}{p(1-p)}}\right) \\ &\approx \Phi\left(\epsilon\sqrt{\frac{n}{p(1-p)}}\right) - \Phi\left(-\epsilon\sqrt{\frac{n}{p(1-p)}}\right) \\ &= 2\Phi\left(\epsilon\sqrt{\frac{n}{p(1-p)}}\right) - 1 \geq 0.95, \quad \text{since } \Phi(x) = 1 - \Phi(-x), \end{aligned}$$

so that we need $\epsilon\sqrt{n/p(1-p)} \geq 1.96$, that is

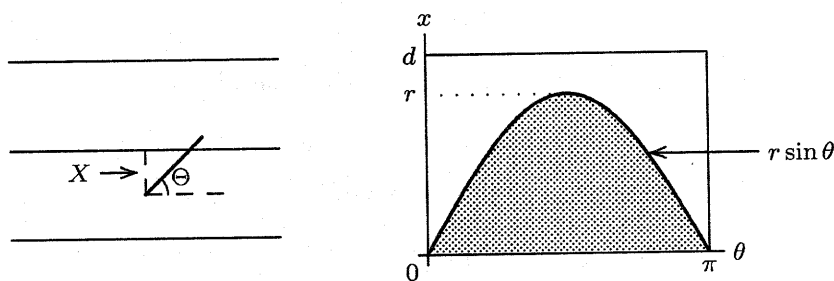
$$n \geq (1.96)^2 p(1-p)/\epsilon^2.$$

We do not know the value of p , but it is always the case that $p(1-p) \leq \frac{1}{4}$, with equality occurring when $p = \frac{1}{2}$, so to ensure that we have the required bound we need $n \geq (1.96/2\epsilon)^2$. For example, if we take $\epsilon = 0.02$, so that the estimate of the percentage of Labour voters is accurate to within 2 percentage points with 95% probability we would need to take a sample with $n \geq 2401$. The typical sample size in opinion polls is $n \approx 1000$ which corresponds to an error $\epsilon \approx 0.03$. $\square$

5.4 Geometric probability

Buffon's Needle Consider a needle of length r which is thrown at random on to a plane surface on which there are parallel straight lines at distance $d > r$ apart. What is the probability that the needle intersects one of the lines?

Think of the parallel lines running West-East and let X be the distance from the point representing the Southern end of the needle to the nearest line North of that point; if the needle is parallel to the lines, take the right-hand end point. Let Θ be the angle that the needle makes with the West-East lines. Then we will assume that X is uniformly distributed on $[0, d)$ and Θ is uniformly distributed on $[0, \pi]$, and that X and Θ are independent.



The joint probability density of (X, Θ) is

$$f(x, \theta) = \begin{cases} \frac{1}{\pi d}, & \text{for } 0 \leq x < d \text{ and } 0 \leq \theta \leq \pi, \\ 0 & \text{otherwise.} \end{cases}$$

Then if A is the shaded area in the $x - \theta$ plane illustrated, the probability that the needle intersects a line is

$$\mathbb{P}(X \leq r \sin \Theta) = \int_A \int f(x, \theta) dx d\theta = \int_0^\pi \int_0^{r \sin \theta} \frac{1}{\pi d} dx d\theta = \frac{r}{\pi d} \int_0^\pi \sin \theta d\theta = \frac{2r}{\pi d}.$$

This probability was derived in 1777 by the French mathematician and naturalist Georges Louis Leclerc, Comte de Buffon, who suggested a method of approximating the value of π by repeatedly dropping a needle and estimating the probability that a line is intersected (and hence the value of π) by recording the proportion of times that the needle crosses a line.

First note that if $X \sim N(\mu, \sigma^2)$, where σ^2 is small, and $g: \mathbb{R} \rightarrow \mathbb{R}$ then

$$g(X) = g(\mu) + (X - \mu)g'(\mu) + \dots \simeq N\left(g(\mu), (g'(\mu))^2 \sigma^2\right),$$

where the symbol $\simeq$ may be read as "approximately distributed as". Now, if $S_n = X_1 + \dots + X_n$ denotes the total number of times that the needle intersects a line in n drops of the needle, where X_i is the indicator of the event that a line is intersected on the i th drop, then $S_n \sim \text{Bin}(n, p)$ where $p = 2r/(\pi d)$. By the Central Limit Theorem we have that $S_n/n \simeq N(p, p(1-p)/n)$. Let $g(x) = 2r/(xd)$, so that $g(p) = \pi$ and $g'(p) = -\pi^2 d/(2r)$. We see that an estimate of π is given by $\hat{\pi} = g(S_n/n)$, where

$$\hat{\pi} \simeq N\left(\pi, \frac{\pi^2}{2rn} (\pi d - 2r)\right).$$

One small difficulty that arises when one tries to replicate Buffon's procedure for estimating π on a computer is how to do the simulation without using the value of π to take random samples of Θ from the uniform distribution on $(0, \pi]$. One way around this is to generate a sample (X, Y) which has the uniform distribution over the circle centre the origin and of unit radius as follows:

Step 1. generate independent X and Y each with the uniform distribution on $[-1, 1]$;

Step 2. if $X^2 + Y^2 > 1$ repeat Step 1, otherwise take (X, Y) .

Now set $\Theta = \frac{1}{2} \tan^{-1}(Y/X)$, which will be uniform on $(0, \pi)$. □

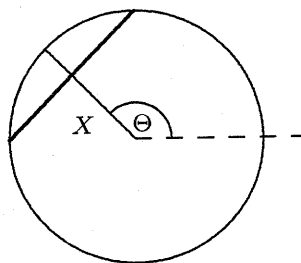
Bertrand's Paradox This results from the following question posed by Bertrand in 1889: What is the probability that a chord chosen at random joining two points of a circle of radius r has length $\leq r$?

The difficulty with the question is that there is not a unique interpretation of what it means for a chord to be chosen 'at random'; there are different ways to do this and they lead to different probabilities for the length, C , of the chord being less than r . We will consider two approaches.

Approach 1. Let X be a random variable having the uniform distribution on $(0, r)$ and let Θ be a random variable, independent of X , with the uniform distribution on $(0, 2\pi)$.

The length of the chord is

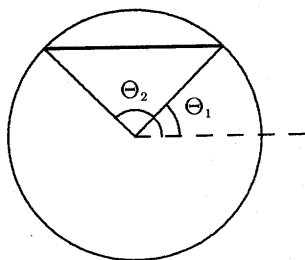
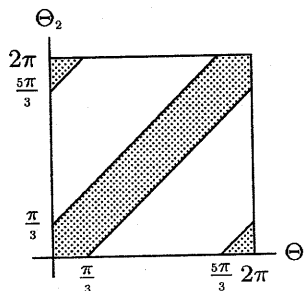
$$C = 2\sqrt{r^2 - X^2}$$



Construct the chord by taking a reference line (the x -axis, say) and drawing the radius at angle Θ with the line. Then take the chord at right angles to this radius at distance X from the centre of the circle. We then have

$$\mathbb{P}(C \leq r) = \mathbb{P}(4(r^2 - X^2) \leq r^2) = \mathbb{P}(\sqrt{3}r/2 \leq X) = 1 - \sqrt{3}/2 \approx 0.134.$$

Approach 2. Let Θ_1 and Θ_2 be independent random variables each with the uniform distribution on $(0, 2\pi)$. Take the end points of the chord as the points $(r \cos \Theta_i, r \sin \Theta_i)$, $i = 1, 2$, on the circumference of the circle, where the angles are measured from some reference line. The length of the chord is $C = 2r \sin \left(\frac{|\Theta_1 - \Theta_2|}{2} \right)$.



The probability is then

$$\mathbb{P}(C \leq r) = \mathbb{P}\left(\sin\left(\frac{|\Theta_1 - \Theta_2|}{2}\right) \leq \frac{1}{2}\right) = \mathbb{P}\left(|\Theta_1 - \Theta_2| \leq \frac{\pi}{3} \text{ or } |\Theta_1 - \Theta_2| \geq \frac{5\pi}{3}\right);$$

this probability is the area of the shaded region in the square divided by $(2\pi)^2$ which gives the probability to be $\frac{1}{3} \approx 0.3333$. This probability is the same as when you take one end of the chord as fixed (say on the reference line) and take the other end at a point at angle Θ uniformly distributed on $(0, 2\pi)$.

It should be noted that both probabilities may arise as the outcomes of physical experiments which are choosing the chord 'at random'. For example, the probability in Approach 1 would be found if a circular disc of radius r is thrown randomly onto a table on which parallel lines at distance $2r$ are drawn; the chord would be determined by the unique line intersecting the circle and the distribution of the distance of center of the circle to the nearest line would be uniform on $(0, r)$. By contrast, the probability in Approach 2 is obtained if the disc is pivoted on a point on its circumference, which is on a given straight line and the disc is spun around that point, then the chord would be determined by the intersection of the given line and the circumference of the disc.

9 January 2007