SOURAV CHATTERJEE

LECTURE NOTES FOR MATH 151

Contents

Basic concepts

Experiments, outcomes, events and probabilities

An **experiment** has many possible **outcomes**. The set of all outcomes is called the **sample space**. For example, if the experiment is 'roll a die', the possible outcomes are 1, 2, 3, 4, 5 and 6. The sample space for this experiment is $\Omega = \{1, \ldots, 6\}$.

Sample spaces may be finite or infinite. For now, let us consider only finite sample spaces. If the outcome of the experiment is supposed to be 'random', we assign a **probability**[1] $P(\omega)$ to each outcome ω in the sample space Ω. The constraints are that $P(\omega)$ has to be nonnegative for each ω, and

$$\sum_{\omega \in \Omega} P(\omega) = 1.$$

In the die-rolling example, if the die is a fair die, it is reasonable[2] to assign $P(\omega) = 1/6$ to each outcome ω.

An **event** is a set of outcomes. For example, in the above die-rolling example, the event 'The die turns up at least 4' is the set $\{4, 5, 6\}$. The probability of an event is the sum of the probabilities of the outcomes constituting that event[3]. For instance, if we name the above event A, then

$$P(A) = P(4) + P(5) + P(6) = \frac{1}{2}.$$

As a slightly more complicated experiment, consider rolling two fair dice. The sample space is the set of all ordered pairs (i, j) of numbers between 1 and 6, that is,

$$\Omega = \{(1,1), (1,2), \ldots, (1,6), (2,1), \ldots, (2,6), \ldots, (6,1), \ldots, (6,6)\}.$$

Again, a reasonable model is $P(\omega) = 1/36$ for each ω. Suppose we let A be the event 'The sum of the two numbers is 5'. Set theoretically,

$$A = \{(1,4), (2,3), (3,2), (4,1)\}.$$

Thus,

$$P(A) = \frac{4}{36} = \frac{1}{9}.$$

[1] There is a lot of philosophical debate about the meaning of 'probability'. The most intuitive notion is that $P(\omega)$ is the fraction of times that ω will occur if the experiment is repeated many times. This is the **frequentist** view of probability. The problem with this notion is that many experiments are not repeatable. (For example, what is the probability that a certain candidate will win an election?) Sometimes people think of probabilities as *beliefs* that are updated according to some set of rules as new knowledge is acquired. This is the **Bayesian** approach. For simplicity, we will adopt the frequentist viewpoint.

[2] One should view this as a kind of scientific theory. You cannot really 'prove' that the probability of each outcome is 1/6. But if you adopt the frequentist viewpoint, and you have an actual fair die in your hand, you can roll it many times and verify that each outcome indeed happens approximately one-sixth of the times, and this approximation becomes more and more accurate if the experiment is repeated more and more times. The model is the validated by the experiments.

[3] So, what is the meaning of the probability of an event according to the frequentist viewpoint?

On the other hand, the chance of getting the sum to be equal to 12 is 1/36, since that event contains only one outcome.

For our next experiment, consider tossing a fair coin three times. The sample space is

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Since there are 8 outcomes, and all of them are presumably equally likely[4], we can assign $P(\omega) = 1/8$ to each $\omega \in \Omega$. Let A be the event 'We get at least two heads'. Set theoretically,

$$A = \{HHH, HHT, HTH, THH\},$$

and so $P(A) = 4/8 = 1/2$.

Finally, let us generalize the above experiment to a sequence of n tosses of a fair coin. Obviously, the sample space Ω is the set of all sequences of heads and tails of length n. The first observation is that the size of Ω, which we will denote by $|\Omega|$, is 2^n. To see this, observe that a sequence of length n is obtained by appending either H or T to the end of a sequence of length $n-1$. Therefore the number sequences is multiplied by 2 each time we add a toss. Thus, we may assign $P(\omega) = 2^{-n}$ to each outcome ω.

Take any k between 0 and n. What is the probability of the event A = 'We get exactly k heads'? Clearly,

$$P(A) = \frac{\text{number of sequences with exactly } k \text{ heads}}{2^n}.$$

But the number in the numerator is not so easy to calculate, unless you know how to do it. This is what we will do in the next section: Learn to count!

The fundamental principle of counting

Suppose you have to do k tasks. Suppose that the first task can be done in n_1 ways. Suppose that after the first task has been executed, no matter how you have done it, the second task can be done in n_2 ways. Suppose that after the first two tasks have been carried out, no matter how, there are n_3 ways of doing the third task. And so on. *Then the number of ways of doing all of the tasks is $n_1 n_2 \cdots n_k$.* This is known as the **fundamental principle of counting**[5].

Let us now work out three very important applications of this principle. First, suppose that you have n cards, marked $1, 2, \ldots, n$. The task is that you have to arrange those cards as a deck. What is the number of ways of doing this?

This task can be broken up into the following sequence of n tasks: Put down the first card, then place a second card on top of it, and

[4] You are, of course, free to disbelieve this. You may say that after two heads, it is more likely to get a tail than a head; that is, HHT is more likely than HHH. Such a model, however, will not give correct predictions for a real experiment with a real coin. This is sometimes called the *gambler's fallacy.* It is not a fallacy in the mathematical or logical sense; it is a fallacy only because it leads to wrong predictions in the real world.

[5] This is not a theorem; there is no 'proof'. It is one of those things that we consider as obvious or self-evident. As a very simple example, suppose that you have to choose one of two routes to come to class, and on each route, there are three breakfast places from which you can pick up breakfast. You have two tasks: (1) Choose a route. (2) Choose a breakfast place. The first task can be done in 2 ways. *Having done the first task,* there are 3 ways of doing the second — even though there are 6 breakfast places in all. So here $n_1 = 2$ and $n_2 = 3$. Convince yourself that the total number of ways of doing both tasks is $n_1 n_2 = 6$.

then put a third one on top of the second, and so on. Clearly, the first task can be done in n ways, because we are free to choose any one of the n cards to place at the bottom of the deck. Having done the first task, we are left with $n-1$ ways of doing the second. Continuing in this way, and applying the fundamental principle of counting, we deduce that the total number of ways of arranging the n cards in a deck is

$$n(n-1)(n-2)\cdots 1.$$

This number is denoted by $n!$ (and pronounced *n-factorial*). This is the number of ways of arranging the number $1, 2, \ldots, n$ in a sequence. Any such arrangement is known as a **permutation** of $1, \ldots, n$. The set of all such permutations is usually denoted by S_n. For convenience, we define $0! = 1$.

Our next example is the problem of seating k guests in n chairs, where k is between 1 and n. The first guest can be seated in n ways. Having seated the first guest, the second guest can be seated in $n-1$ ways, and so on. Therefore the total number of ways of seating k guests in n chairs is

$$\underbrace{n(n-1)\cdots(n-k+1)}_{k \text{ terms}} = \frac{n!}{(n-k)!}. \tag{1}$$

Note that this makes sense even if $k = 0$, since the right side is 1.

Next, consider the problem of selecting a *set* of k chairs out of n. Note that this is different than the previous example in that we are not considering the *order in which the chairs are picked*[6]. What is the number of ways of doing this task?

Let us call it x. Now consider the task from the previous example: Seat k guests in n chairs. That task can be broken up into a sequence of two tasks — first, pick a set of k chairs out of n, and then, pick an arrangement of those k chairs to seat the k guests. The first task can be done in x ways. Having done the first task, the second task is the same as picking a permutation of $1, \ldots, k$, which, as we have learnt, can be done in $k!$ ways. Thus, k guests can be seated in n chairs in $xk!$ ways. But we already know that the number of ways is given by the formula (1). Therefore

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

which gives

$$x = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

This number is a **binomial coefficient**, usually denoted by

$$\binom{n}{k}.$$

[6] For example, if $k = 2$, and we pick chair 2 first and then chair 3, this would be considered to be the same as picking chair 3 first and then chair 2. We will not count these as different ways of doing the task of picking a set of 2 chairs out of n.

By the convention $0! = 1$, it follows that

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

The name 'binomial coefficient' comes from the **binomial theorem**, which says that[7]

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

Example: Back to coin tossing

Recall the experiment where we were tossing a fair coin n times. How many outcomes have exactly k heads? Note that this is the same as the number of ways of choosing a set of k chairs out of n, since specifying the locations of the heads completely determines the outcome. Thus, there are $\binom{n}{k}$ sequences of n tosses which have exactly k heads. Consequently, the probability of getting exactly k heads is

$$\binom{n}{k} 2^{-n}.$$

Example: The birthday paradox

In a class of n students, what is the chance that there is at least one pair of students with the same birthday (that is, day and month, not year)? For simplicity, let us ignore February 29 and assume that there are 365 days in a year. The sample space Ω is the set of all sequences of length n where each member of the sequence is a number between 1 and 365. By the fundamental principle of counting, $|\Omega| = 365^n$. Let us assume that model that all of the sequences are equally likely[8].

Again, by the fundamental principle of counting, the number of ways that all n birthdays can be different from each other is

$$365 \cdot 364 \cdot 363 \cdots (365 - n + 1).$$

Note that this holds even if $n > 365$, since introduces a 0 in the above product. Thus, the number of sequences with at least one pair of duplicated birthdays is 365^n minus the above product, which means that the probability of this event is

$$1 - \frac{365 \cdot 364 \cdot 363 \cdots (365 - n + 1)}{365^n}.$$

The actual values of this probability are surprising, which is why it is called a paradox. For n as small as 23, this is approximately 0.51. For $n = 40$, it is bigger than 0.89. So in a class of size 40, there is near certainty that there is a pair of students with the same birthday.

[7] The proof is easy: Just note that when you expand $(a+b)^n$ using the distributive law, the number of terms containing k a's and $n-k$ b's is the same as the number of ways of choosing k locations out of n, which is $\binom{n}{k}$. Any such term contributes $a^k b^{n-k}$.

[8] It is not clear that this model is fully accurate. For example, births may be less likely on national holidays due to the unavailability of medical staff. But we will go with it.

The reason why this feels so surprising is that very few of us know anyone with the same birthday as ourselves, even though each of us knows a lot of people. The resolution of this apparent contradiction is that for a *given person*, who knows the birthdays of n acquaintances, the probability that this person shares his or her birthday with one of these n acquaintances is

$$1 - \frac{364^n}{365^n}.$$

(Prove this.) For $n = 40$, this probability is less than 0.11. The smallest n for which this probability exceeds 0.5 is 253. Very few of us are aware of the birthdays of more than 30 or 40 acquaintances. This shows why most of us do not know of anyone with the same birthday as their own selves.

Operations with events

Since an event is just a subset of the sample space, we can perform set theoretic operations with events. If A is an event, the **complement** of A, denoted by A^c, is the set $\Omega \setminus A$. Since

$$P(A) = \sum_{\omega \in A} P(\omega),$$

and $P(\Omega) = 1$, it follows that

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

The **union** of two events A and B is the set $A \cup B$, which consists of all outcomes that are either in A or in B or in both A and B. In probability theory, $A \cup B$ is often referred to as 'A or B'.

The **intersection** of A and B, denoted by $A \cap B$, is the set of all outcomes that belong to both A and B. We often call this event 'A and B'.

Now suppose we add up $P(A)$ and $P(B)$. Then we are adding up $P(\omega)$ for each ω that belongs to the union of A and B, but we are double counting those in the intersection. So if we subtract off $P(A \cap B)$ from this sum, we end up with $P(A \cup B)$. In other words,

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

This is known as the **inclusion-exclusion formula**.

What if we have three events A, B and C? Writing $A \cup B \cup C$ as the union of $A \cup B$ and C, we can apply the inclusion-exclusion formula to get

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

By the distributive law for unions and intersections[9],

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C).$$

Therefore again by the inclusion-exclusion formula,

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

Combining the steps, we get

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

This is the inclusion-exclusion formula for three events. We can clearly see a pattern emerging here. In fact, there is a **generalized inclusion-exclusion formula** for the probability of the union of n events:

$$\begin{aligned} P(A_1 \cup \cdots \cup A_n) = \sum P(A_i) &- \sum P(A_i \cap A_j) \\ &+ \sum P(A_i \cap A_j \cap A_k) - \cdots \\ &+ (-1)^{n-1} P(A_1 \cap \cdots \cap A_n), \end{aligned}$$

where there is no double counting in any of the sums[10]. Note that there are $\binom{n}{r}$ terms in the r^{th} sum, since each term corresponds to the choice of a set of r indices out of n. We will use this observation later.

Example: The secretary problem

Suppose that a secretary needs to insert n letters into n marked envelopes. But due to a mishap, the letters get all mixed up, and the secretary just inserts the letters randomly into envelopes. The *secretary problem* is the problem of computing the chance the none of the letters go into the correct envelope[11].

A great surprise is that for large n, this probability is neither close to 0 nor close to 1. In fact, as $n \to \infty$, this probability converges to $1/e$, which is approximately 0.37. We will now prove this.

Let A_i be the event that letter i goes into the correct envelope. Then $A_1 \cup \cdots \cup A_n$ is the event that at least one letter goes into the correct envelope. We will now apply the generalized inclusion-exclusion formula to calculate the probability of this event.

Take any r between 1 and n. Then $A_1 \cap \cdots \cap A_r$ is the event that letters $1, \ldots, r$ all go into the correct envelopes. If this event happens, then the number of ways of inserting the remaining $n - r$ letters into

[9] Convince yourself using Venn diagrams.

[10] Try to prove this by induction; that is, assume that the formula is true for $n - 1$ events, and the prove it for n, the same way as we extended from 2 to 3 events.

[11] There is also another very different problem in probability that goes by the name of 'secretary problem'. See Wikipedia.

the remaining $n-r$ envelopes is $(n-r)!$. Since the total number of ways of inserting letters into envelopes in $n!$, this shows that

$$P(A_1 \cap \cdots \cap A_r) = \frac{(n-r)!}{n!}.$$

Now take any distinct $i_1, \ldots, i_r$ between 1 and n. Clearly, the same argument shows that $P(A_{i_1} \cap \cdots \cap A_{i_r})$ is also equal to $(n-r)!/n!$. Thus, each term in the r^{th} sum of the inclusion-exclusion formula equals $(n-r)!/n!$. Since there are $\binom{n}{r}$ terms, the r^{th} sum equals

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

Therefore,

$$P(A_1 \cup \cdots \cup A_n) = \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!}.$$

The probability that no letter goes into the correct envelope is 1 minus the above quantity, which converges[12] to e^{-1} as $n \to \infty$.

Disjoint events

Two events A and B are called **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$, that is, if there is no outcome that is in both A and B. Yet another way to put it is that the events A and B cannot happen simultaneously. A sequence of events $A_1, \ldots, A_n$ is called disjoint or mutually exclusive if $A_i \cap A_j = \emptyset$ for any $i \neq j$. That is, if no two events can happen simultaneously. The most important property of disjoint events is that if $A_1, \ldots, A_n$ are disjoint, then

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i).$$

This is an immediate consequence of the inclusion-exclusion formula, but it also follows easily from the definition of the probability of an event as the sum of the probabilities of outcomes, since no outcome can belong to two of the above events.

A collection of events $A_1, \ldots, A_n$ is called a **partition** of the sample space if the events are disjoint and their union is Ω.

Proposition 1. *If $A_1, \ldots, A_n$ is a partition of Ω, then for any event B,*

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i).$$

Proof. First, note that the events $B \cap A_1, \ldots, B \cap A_n$ are disjoint. This follows easily from the fact that $A_1, \ldots, A_n$ are disjoint. Next, we claim that

$$B = (B \cap A_1) \cup \cdots \cup (B \cap A_n). \tag{2}$$

[12] The convergence is very fast. Even for $n = 5$, it is 0.367 to three places of decimal, whereas $e^{-1} = 0.368$ to three places of decimal.

To prove this, we will now show that the two sets displayed above both contain the other[13]. First, take any $\omega \in B$. Since $A_1, \ldots, A_n$ is a partition of Ω, ω must be in some A_i. Then $\omega \in B \cap A_i$, and hence

$$\omega \in (B \cap A_1) \cup \cdots \cup (B \cap A_n).$$

This proves that $B \subseteq (B \cap A_1) \cup \cdots \cup (B \cap A_n)$. Conversely, take any $\omega \in (B \cap A_1) \cup \cdots \cup (B \cap A_n)$. Then $\omega \in B \cap A_i$ for some i, and so, $\omega \in B$. Thus, $B \supseteq (B \cap A_1) \cup \cdots \cup (B \cap A_n)$. This concludes the proof of (2). Combining this with the fact that $B \cap A_1, \ldots, B \cap A_n$ are disjoint completes the proof. □

The following corollary is often useful.

Corollary 1. *For any two events A and B,*

$$P(B) = P(B \cap A) + P(B \cap A^c).$$

Proof. Simply observe that A, A^c is a partition of Ω, and apply Proposition 1. □

Conditional probability

Let A and B be two events, with $P(A) > 0$. The **conditional probability of B given that A has happened** is defined as[14]

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

When $P(A) = 0$, we leave it undefined[15].

As an example, consider a single roll of a die. If A is the event that the number that turns up is at least 4, and B is the event that the number is 6, then

$$P(B|A) = \frac{1/6}{1/2} = \frac{1}{3}.$$

Observe that for any two events A and B, $P(A \cap B) = P(A)P(B|A)$. This holds even if $P(A) = 0$, irrespective of how we define $P(B|A)$ in that situation.

The law of total probability

The following proposition is sometimes called the law of total probability.

Proposition 2 (Law of total probability). *Let $A_1, \ldots, A_n$ be a partition of Ω. Then for any event B,*

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

[13] This is the standard way of showing that two sets are equal.

[14] From the frequentist viewpoint, this formula is justified as follows. Suppose that the experiment is repeated many times. Then $P(A)$ is the fraction of times that A happened, and $P(A \cap B)$ is the fraction of times that A and B both happened.Therefore $P(B|A)$ is the fraction of times that B happened *among those instances where A happened.*

[15] Actually, conditional probability given A can be defined even if $P(A) = 0$, and that is an important matter. We will talk about it later.

Proof. Apply Proposition 1, and use $P(B \cap A_i) = P(B|A_i)P(A_i)$. □

Considering the partition A, A^c, we get the following corollary.

Corollary 2. *For any two events A and B,*

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

Bayes rule

Let A and B be two events with $P(A) > 0$ and $P(B) > 0$. The following formula is an easy consequence of the definition of conditional probability:

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

This is known as **Bayes rule**.

Sometimes, Bayes rule can lead to surprising consequences. Consider the following example. Suppose that a rare disease afflicts 0.5% of the population. Suppose that there is a diagnostic test which is 99% accurate, which means that it gives the correct diagnosis with probability 0.99 if a person has the disease and also if a person does not have the disease. Now, if a random person tests positive, what is the conditional probability that the person has the disease?

We proceed systematically, as follows. Let D be the event that the person has the disease[16]. Let $+$ denote the event that the person tests positive. We are interested in evaluating $P(D|+)$. By Bayes rule,

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

By the law of total probability,

$$P(+) = P(+|D)P(D) + P(+|D^c)P(D^c).$$

By the given information, we know that $P(D) = 0.005$, $P(+|D) = 0.99$ and $P(+|D^c) = 0.01$. Thus,

$$\begin{aligned} P(D|+) &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)} \\ &= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} = 0.3322148. \end{aligned}$$

Thus, a person who tests positive is only about 33% likely to have the disease[17].

[16] Try to set this up as an experiment with a sample space, probability, etc.

[17] This can be demystified as follows. Suppose that 1000 randomly chosen people are tested. Since the disease afflicts 0.5% of the population, we may expect that 5 people out of these 1000 actually have the disease. The test will almost certainly diagnose these 5 people correctly. Among the remaining 995, approximately 1% — about 10 people — are misdiagnosed by the test. Thus, 15 people will get positive results but only 5 of them really have the disease.

Independent events

An event B is said to be independent of an event A if the information that A has happened does not change the likelihood of B; that is, $P(B|A) = P(B)$. This can be rewritten as $P(B \cap A) = P(B)P(A)$. But note that this can again be rewritten as $P(A|B) = P(A)$. Therefore, if B is independent of A, then A is independent of B. This slightly strange fact shows that independence is a symmetric relation: We say that two events A and B are **independent** if

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

For example, suppose we toss a fair coin twice. Let A be the event that the first toss turns up heads and let B be the event that the second toss turns up heads. Then $P(A) = P(B) = 2/4 = 1/2$, and $P(A \cap B) = 1/4 = P(A)P(B)$. Thus, A and B are independent events.

On the other hand, suppose a fair coin is tossed three times. Let A be the event that the first two tosses are heads, and let B be the event that the last two tosses are tails. Then $P(A) = P(B) = 2/8 = 1/4$, and $P(A \cap B) = 1/8 \neq P(A)P(B)$. Thus, A and B are not independent.

The concept of independence extends to more than two events. Events $A_1, \ldots, A_n$ are called independent (or **mutually independent**) if for any k between 1 and n, and any distinct indices $i_1, \ldots, i_k$ between 1 and n,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

This is equivalent to saying that for any distinct $i_1, \ldots, i_k, j_1, \ldots, j_l$,

$$P(A_{j_1} \cap \cdots \cap A_{j_l} | A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{j_1} \cap \cdots \cap A_{j_l}).$$

For an example of a sequence of independent events, consider n tosses of a fair coin. Let A_i be the event that toss i turns up heads. Then $A_1, \ldots, A_n$ are independent events, as shown by the following argument. Take any distinct $i_1, \ldots, i_k$. The number of outcomes where tosses $i_1, \ldots, i_k$ all turn up heads is 2^{n-k}, because the remaining tosses can be determined in 2^{n-k} ways (by the fundamental principle of counting). Thus,

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = \frac{2^{n-k}}{2^n} = 2^{-k}.$$

Applying this with $k = 1$, we get that $P(A_i) = 1/2$ for each i. Thus,

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

proving the independence of $A_1, \ldots, A_n$.

There is also a different concept called **pairwise independence**. Events $A_1, \ldots, A_n$ are called pairwise independent if $P(A_i \cap A_j) = P(A_i)P(A_j)$ for any $i \neq j$. If the events are independent, then they are automatically pairwise independent. Surprisingly, the converse is not true, as shown by the following counterexample. Toss a fair coin twice. Let A be the event that the first toss turns up heads. Let B be the event that the second toss turns up heads. Let C be the event that either both tosses are heads or both tosses are tails. Then it is easy to check that $P(A) = P(B) = P(C) = 1/2$ and $P(A \cap B) = P(A \cap C) = P(B \cap C) = 1/4$, which means that A, B and C are pairwise independent. However,

$$P(A \cap B \cap C) = \frac{1}{4} \neq P(A)P(B)P(C),$$

which shows that the three events are not mutually independent.

Independence of A and B implies that A^c and B are independent, as are A^c and B^c, and A and B^c. Let us only show this for A^c an B:

$$\begin{aligned} P(A^c \cap B) &= P(B) - P(A \cap B) \quad \text{(by Corollary 1)} \\ &= P(B) - P(A)P(B) \quad \text{(by independence of } A \text{ and } B\text{)} \\ &= (1 - P(A))P(B) = P(A^c)P(B). \end{aligned}$$

More generally, if $A_1, \ldots, A_n$ are independent events, and $B_1, \ldots, B_n$ are events such that each B_i is either A_i or A_i^c, then $B_1, \ldots, B_n$ are also independent (prove this).

Discrete random variables

Random variables as functions

We will continue with finite sample spaces for the time being. Let Ω be a finite sample space. A function $X : \Omega \to \mathbb{R}$ is called a **random variable**. In other words, a random variable assigns a number to each outcome. The numbers need not be all different.

For example, consider the experiment of rolling a fair die twice. We can define a number of random variables related to this experiment. For example, we can define X to be the number that turns up on the first roll, Y to be number that turns up on the second roll, and Z to be the sum of the two numbers. Thus, for instance, if $\omega = (1, 4)$, then $X(\omega) = 1$, $Y(\omega) = 4$, and $Z(\omega) = 5$. The function Z is the sum of the two functions X and Y. We write this simply as $Z = X + Y$.

Consider n tosses of a fair coin. Let X be the number of heads. Also, for each i, let

$$X_i = \begin{cases} 1 & \text{if toss } i \text{ turns up heads,} \\ 0 & \text{if toss } i \text{ turns up tails.} \end{cases}$$

It is easy to see how X and $X_1, \ldots, X_n$ are random variables (that is, functions from the sample space into the real line), and

$$X = \sum_{i=1}^{n} X_i.$$

If X is a random variable and x is a real number, the event

$$\{\omega : X(\omega) = x\}$$

is usually abbreviated as $\{X = x\}$, and its probability is denoted by $P(X = x)$. Similarly, for any subset A of the real line, the event

$$\{\omega : X(\omega) \in A\}$$

is abbreviated as $\{X \in A\}$ and its probability is written as $P(X \in A)$. Another convention is that if X and Y are two random variables and A and B are two subsets of real numbers, then the event $\{X \in A\} \cap \{Y \in B\}$ is often written as $\{X \in A, Y \in B\}$.

Probability mass function

The **probability mass function** (p.m.f.) of a random variable X is the function $f : \mathbb{R} \to \mathbb{R}$ defined as

$$f(x) = P(X = x),$$

and the **cumulative distribution function** (c.d.f.) is defined as

$$F(x) = P(X \leq x).$$

For example, let X be the number of heads in n tosses of a fair coin. Then, as we calculated earlier, the probability mass function of X is

$$f(k) = \binom{n}{k} 2^{-n}$$

when k is an integer between 0 and n, and 0 otherwise.

If the sample space is finite, a random variable can take only finitely many possible values. If, however, the sample space is infinite, the set of possible values of a random variable can be infinite. If the set of possible values of a random variable is finite or countably infinite, the random variable is called a **discrete random variable**. Note that for a discrete random variable with p.m.f. f, $f(x) \neq 0$ only for countably many x's, and

$$\sum_x f(x) = 1,$$

since the events $\{X = x\}$ form a partition of Ω as x ranges over all[18] possible values of X. In this chapter, we will only deal with discrete random variables.

We use the notation $X \sim f$ as an abbreviation of the sentence "The random variable X has probability mass function f."

Independence

A collection of discrete random variables $X_1, \ldots, X_n$ is called independent if for any $x_1, \ldots, x_n$,

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n). \end{aligned}$$

Note that unlike events, we did not require the product rule for all subcollections. This is because it's automatic.

Proposition 3. *If $X_1, \ldots, X_n$ is a collection of independent discrete random variables, then any subcollection is also independent.*

[18] This requires that the sum rule for probabilities of unions of disjoint events extends to countably infinite collections of disjoint events. A proper justification of this needs the measure theoretic framework of probability theory. We will just assume that this is true.

Proof. We will prove that X_1 and X_2 are independent. The general case is similar. Accordingly, note that for any x_1 and x_2, the event $\{X_1 = x_1, X_2 = x_2\}$ is the expressible as

$$\begin{aligned} &\{X_1 = x_1, X_2 = x_2\} \\ &= \bigcup_{x_3,\ldots,x_n} \{X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n\}, \end{aligned}$$

where the sum is taken over all possible values of $x_3, \ldots, x_n$, and the events on the right are disjoint[19]. Since the random variables are discrete, the union is countable. Thus,

$$P(X_1 = x_1, X_2 = x_2) = \sum_{x_3,\ldots,x_n} P(X_1 = x_1, \ldots, X_n = x_n).$$

By independence, $P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$. We can then bring the common factor $P(X_1 = x_1)P(X_2 = x_2)$ outside of the sum, getting

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2) \\ &= P(X_1 = x_1)P(X_2 = x_2) \sum_{x_3,\ldots,x_n} P(X_3 = x_3) \cdots P(X_n = x_n). \end{aligned}$$

By the distributive law[20],

$$\begin{aligned} &\sum_{x_3,\ldots,x_n} P(X_3 = x_3) \cdots P(X_n = x_n) \\ &= \left(\sum_{x_3} P(X_3 = x_3)\right) \cdots \left(\sum_{x_n} P(X_n = x_n)\right) = 1. \end{aligned}$$

Thus, $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$. Since this holds for any x_1 and x_2, X_1 and X_2 are independent. □

We will later encounter infinite sequences of independent random variables. An infinite sequence of random variables $X_1, X_2, \ldots$ is called independent if for each n, $X_1, \ldots, X_n$ are independent.

Bernoulli and binomial random variables

Until now, we have only considered tosses of fair coins. There is also the concept of a p-coin, which is a coin that turns up heads with probability p and tails with probability $1 - p$, where p is a number in the interval $[0, 1]$. A fair coin, then, is the same as a $1/2$-coin. Mathematically, the sample space is $\Omega = \{H, T\}$, with $P(T) = 1 - p$ and $P(H) = p$. Let $X : \Omega \to \mathbb{R}$ be defined as $X(T) = 0$ and $X(H) = 1$. Then $P(X = 0) = 1 - p$ and $P(X = 1) = p$. A random variable such as X is called a $\mathit{Bernoulli}(p)$ random variable. We abbreviate this by writing $X \sim \mathit{Bernoulli}(p)$. Sometimes, we say that the **distribution** of X is $\mathit{Bernoulli}(p)$.

[19] Prove this by showing that the two sets contain each other.

[20] For example, $\sum_{i,j} a_i b_j = (\sum_i a_i)(\sum_j b_j)$.

Now suppose that a p-coin is tossed n times. If we want to define a model that renders the outcomes of these n tosses independent, then we must define

$$P(\omega) = p^k(1-p)^{n-k}$$

for an outcome ω which has k heads and $n-k$ tails[21]. This generalizes the case of a fair coin ($p = 1/2$) that we have seen before. Let X be the number of heads. Then clearly[22],

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

for any integer k between 0 and n. Any random variable with this p.m.f. is called a $Bin(n, p)$ random variable.

Often, $Bernoulli(p)$ is abbreviated as $Ber(p)$ and $Binomial(n, p)$ is abbreviated as $Bin(n, p)$.

Infinite sequence of coin tosses

Suppose that we decide to keep tossing a p-coin until it turns up heads for the first time, and record the number of tosses required to get there as a random variable X. Unless there is something severely wrong with the coin (that is, it never turns up heads — in other words, $p = 0$), this is going to happen eventually[23]. So this is an experiment that one can conduct in real life. But what is the sample space for this experiment? A moment's thought will reveal that we cannot construct a finite sample space for this experiment. There is no upper bound on the number of tosses required to complete the experiment. The only recourse is to put it in the framework of an experiment where the outcomes are all possible *infinite* sequences of coin tosses, and for such an outcome ω, define $X(\omega)$ to be the location of the first head in ω.

The main challenge in the above setup is the definition of probability. Consider the case $p = 1/2$. Then all outcomes must be equally likely, but there are infinitely many outcomes. So the probability of any single outcome must be zero. And yet, $P(\Omega)$ needs to be 1. This makes it impossible to define the probability of an event as the sum of the probabilities of its constituent outcomes. Instead, we directly define probabilities of events so that they satisfy a certain set of axioms. The problem with this approach, again, is that it is generally impossible to define probabilities of *all* events in this manner without running into contradictions. In the rigorous mathematical definition of probability theory, this problem is resolved by defining the probabilities of a *subcollection* of events, known as a σ-algebra. We try

[21] If P is defined in this manner, and A_i is the event that toss i turns up heads, show that the events $A_1, \ldots, A_n$ are independent.

[22] Try to prove this if it is not obvious.

[23] Can you prove this?

to ensure that the σ-algebra contains all events that will ever be of interest to us.

In the above setting, let X_i be 1 if toss i turns up heads and 0 otherwise. Then $X_1, X_2, \ldots$ are random variables defined on Ω. Without going deeper into the measure theoretic foundation of probability, let us only be content with the following: *It is possible to define the probabilities of a certain subcollection of events such that any event involving finitely many of the X_i's belongs to this subcollection, and for any n, $X_1, \ldots, X_n$ are independent $Bernoulli(p)$ random variables.*

A sequence $X_1, X_2, \ldots$ as above is called an infinite sequence of independent $Bernoulli(p)$ random variables. Since these random variables all have the same distribution, we say that they are **independent and identically distributed (i.i.d.)**.

In general, an infinite sequence of random variables $X_1, X_2, \ldots$ is called independent if for every n, $X_1, \ldots, X_n$ are independent. If moreover the random variables have the same distribution, they are called i.i.d.

Geometric random variables

Consider an infinite sequence of tosses of a p-coin, as discussed above. Let X be the first time the coin turns up heads. Then we say that X has a $Geometric(p)$ distribution[24]. Often, $Geometric(p)$ is abbreviated as $Geo(p)$.

Let us now derive the p.m.f. of X. Let X_i be 1 if toss i turns up heads and 0 otherwise. Then for any $k \geq 1$,

$$\begin{aligned} P(X = k) &= P(X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1) \\ &= P(X_1 = 0) \cdots P(X_{k-1} = 0) P(X_k = 1) \\ &= (1-p)^{k-1} p. \end{aligned}$$

Notice that any positive integer is a possible value[25] of X.

Poisson random variables

Take any real number $\lambda > 0$. A random variable X is said to have a $Poisson(\lambda)$ distribution (often abbreviated as $Poi(\lambda)$) if its set of possible values is the set of nonnegative integers, and for any $k \geq 0$,

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

A simple way to construct such a random variable is to take Ω to be the set of nonnegative integers, let $P(k)$ be given by the above formula[26] for each $k \in \Omega$, and let $X(k) = k$ for each $k \in \Omega$. But this

[24] Sometimes, X is defined to be the number of tails before the first head. We will not use that definition.

[25] Can you show using the formula that the sum of $P(X = k)$ over all positive integers k equals 1?

[26] Show that $\sum_{k=0}^{\infty} P(k) = 1$.

feels like cheating, because X is not defined in terms of some natural experiment. Indeed, what is the origin of a Poisson random variable? Why are they important?

Typically, the number of occurrences of a certain type of event within a continuous time period — for example, the number of phone calls arriving at a call center in one hour — is modeled as a Poisson random variable.

What is the justification for such modeling, besides the fact that it is often quite successful? The mathematical reasoning comes from the following result.

Proposition 4. *Take any $\lambda > 0$. For each n, let $X_n \sim Bin(n, \lambda/n)$. Then for any $k \geq 0$,*

$$\lim_{n\to\infty} P(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Proof. Note that

$$\begin{aligned} P(X_n = k) &= \binom{n}{k}\frac{\lambda^k}{n^k}\left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k}. \end{aligned}$$

Now, for fixed k,

$$\lim_{n\to\infty}\frac{n(n-1)\cdots(n-k+1)}{n^k} = 1,$$

and

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k} = 1.$$

Moreover,

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n} = e^{-\lambda}.$$

The proof is completed by combining the above observations □

Let us now try to understand the meaning of the above result through an example. Consider the example of the call center. Suppose that in each time interval of one second, the call center receives a single call with probability 0.001, and no calls with probability 0.999. Assume that the calls are made independently. Then the number of calls received in one hour (call it X) has a $Bin(3600, 0.001)$ distribution. In other words, $X \sim Bin(n, \lambda/n)$, where $n = 3600$ and $\lambda = 3600 \times 0.001 = 3.6$. Since 3600 is a large number, we can apply Proposition 4 to conclude that X is approximately a $Poi(3.6)$ random variable. The advantage of doing so is that Poisson random variables are mathematically easier to handle than binomial random variables and have nicer properties.

Joint probability mass function

Let $X_1, \dots, X_n$ be discrete random variables defined on the same sample space. The function

$$f(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n)$$

is called the **joint probability mass function** (joint p.m.f.) of the random variables $X_1, \dots, X_n$. In this context, the **marginal p.m.f.** of X_i is

$$f_i(x) = P(X_i = x).$$

The marginal p.m.f.'s can be obtained from the joint p.m.f. using the following simple method:

$$f_i(x) = \sum_{x_1,\dots,x_{i-1},x_{i+1},\dots,x_n} f(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n). \tag{3}$$

This is a simple consequence of the observation that the event $\{X_i = x\}$ is the union of the disjoint events $\{X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = x, X_{i+1} = x_{i+1}, \dots, X_n = x_n\}$ as $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ range over all possible values of $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$.

As an example, consider the following. Let n be a positive integer. Choose (X, Y) **uniformly** from the set of A_n all pairs of positive integers (x, y) such that $x + y \le n$. What this means is that A_n is the set of all possible values of (X, Y), and $P(X = x, Y = y)$ is the same for every $(x, y) \in A_n$. For example, if $n = 3$, this set consists of the pairs $(1, 1)$, $(1, 2)$ and $(2, 1)$, and so in this case

$$P(X = 1, Y = 1) = P(X = 1, Y = 2) = P(X = 2, Y = 1) = \frac{1}{3}.$$

For general n, note that the number of (x, y) such that $x + y = k$ is exactly $k - 1$, which shows that[27]

$$|A_n| = \sum_{k=2}^{n} (k-1) = \sum_{j=1}^{n-1} j = \frac{n(n-1)}{2},$$

and so for each $(x, y) \in A_n$,

$$P(X = x, Y = y) = \frac{1}{|A_n|} = \frac{2}{n(n-1)}.$$

Thus, the joint p.m.f. of (X, Y) is the function

$$f(x, y) = \begin{cases} \frac{2}{n(n-1)} & \text{if } (x, y) \in A_n, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Let us now compute the marginal p.m.f.'s $f_1(x) = P(X_1 = x)$ and $f_2(y) = P(Y = y)$. Note that the possible values of X are $1, 2, \dots, n -$

[27] This is the standard arithmetic series summation formula.

1, and given that $X = x$, the possible values of Y are $1, 2, \dots, n - x$. Take any $1 \leq x \leq n - 1$. Then by (3) and (4),

$$f_1(x) = \sum_y f(x, y) = \sum_{y=1}^{n-x} \frac{2}{n(n-1)} = \frac{2(n-x)}{n(n-1)}.$$

Similarly,

$$f_2(y) = \frac{2(n-y)}{n(n-1)}$$

for $y = 1, 2, \dots, n - 1$. You may check that f_1 and f_2 are indeed probability mass functions.

We sometimes use joint p.m.f.'s to establish independence of random variables. The following result is used for this purpose.

Proposition 5. *Let $X_1, \dots, X_n$ be discrete random variables with joint probability mass function f. Suppose that*

$$f(x_1, \dots, x_n) = h_1(x_1) h_2(x_2) \cdots h_n(x_n)$$

for some probability mass functions $h_1, \dots, h_n$. Then $X_1, \dots, X_n$ are independent, and $X_i \sim f_i$ for $i = 1, \dots, n$.

Proof. Let $f_1, \dots, f_n$ be the marginals of f. Since $h_1, \dots, h_n$ are probability mass functions, equation (3) gives

$$\begin{aligned} f_1(x) &= \sum_{x_2, \dots, x_n} f(x, x_2, \dots, x_n) \\ &= \sum_{x_2, \dots, x_n} h_1(x) h_2(x_2) \cdots h_n(x_n) \\ &= h_1(x) \sum_{x_2, \dots, x_n} h_2(x_2) \cdots h_n(x_n) \\ &= h_1(x) \left(\sum_{x_2} h_2(x_2) \right) \cdots \left(\sum_{x_n} h_n(x_n) \right) \\ &= h_1(x). \end{aligned}$$

Thus, $f_1 = h_1$. Similarly, $f_i = h_i$ for every i, which shows that $X_i \sim f_i$. Moreover, this also shows that

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n) &= f(x_1, \dots, x_n) \\ &= h_1(x_1) \cdots h_n(x_n) \\ &= f_1(x_1) \cdots f_n(x_n) \\ &= P(X_1 = x_1) \cdots P(X_n = x_n). \end{aligned}$$

Therefore, $X_1, \dots, X_n$ are independent. □

As an illustrative application of Proposition 5, consider the following. In a sequence of tosses of a p-coin, let X_i be the number of tosses required to get the i^{th} head after getting the $(i-1)^{\text{th}}$ head. The claim is that $X_1, X_2, \ldots$ are i.i.d. $Geo(p)$ random variables. To see this, just note that for any n, and any positive integers $x_1, \ldots, x_n$, the event $\{X_1 = x_1, \ldots, X_n = x_n\}$ simply means that the first $x_1 - 1$ tosses are tails, the next one is head, the next $x_2 - 1$ tosses are tails, the next one is again head, and so on. Thus,

$$P(X_1 = x_1, \ldots, X_n = x_n) = (1-p)^{x_1-1}p(1-p)^{x_2-1}p\cdots(1-p)^{x_n-1}p.$$

But $f(x) = (1-p)^{x-1}p$ is the p.m.f. of $Geo(p)$. Therefore by Proposition 5, the claim is proved.

Conditional probability mass function

Let X and Y be discrete random variables with joint probability mass function f. Take any x such that $P(X = x) > 0$. Then the conditional probability mass function of Y given $X = x$ is the function g_x defined as

$$g_x(y) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{f(x,y)}{f_1(x)},$$

where f_1 is the marginal p.m.f. of X. The standard convention is to denote the joint p.m.f. of (X, Y) by $f_{X,Y}$, the marginal p.m.f.'s of X and Y by f_X and f_Y, and the conditional p.m.f. of Y given $X = x$ by $f_{Y|X=x}$. In this notation,

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

For example, let X be the number of heads in n tosses of a p-coin, and let Y be the number of heads in the first m tosses, where $1 \le m \le n$. Then the joint density of (X, Y) at a point (x, y), where $0 \le y \le m$, $0 \le x \le n$, and $0 \le x - y \le n - m$, is

$$\begin{aligned} f_{X,Y}(x,y) &= P(X = x, Y = y) \\ &= P(Y = y, X - Y = x - y) \\ &= \binom{m}{y}\binom{n-m}{x-y}p^x(1-p)^{n-x}, \end{aligned}$$

since the number of ways to get y heads in the first m tosses and $x - y$ heads in the next $n - m$ tosses is $\binom{m}{y}\binom{n-m}{x-y}$, and each such outcome has probability $p^x(1-p)^{n-x}$. For any other (x, y), $f_{X,Y}(x, y) = 0$. On the other hand, we know that $X \sim Bin(n, p)$, and so, for any $0 \le x \le n$,

$$f_X(x) = \binom{n}{x}p^x(1-p)^{n-x}.$$

Thus, for any $0 \le x \le n$, the conditional p.m.f. of Y given $X = x$ is

$$f_{Y|X=x}(y) = \frac{\binom{m}{y}\binom{n-m}{x-y}}{\binom{n}{x}},$$

provided that the constraints $0 \le y \le m$ and $0 \le x - y \le n - m$ are satisfied. Otherwise, $f_{Y|X=x}(y) = 0$. The constraints can be alternatively written as[28]

$$\max\{0, x + m - n\} \le y \le \min\{m, x\}.$$

This distribution is called the **hypergeometric distribution** with parameters n, m and x, and denoted by $Hypergeometric(n, m, x)$.

[28] Check that the two versions are equivalent.

Expectation and variance

Expected value

Let X be a discrete random variable. The **expected value** or **expectation** or **mean** of X is defined as

$$E(X) = \sum_x xP(X = x),$$

where the sum[29] is over all possible values of X, provided that the sum converges absolutely[30].

Let us now calculate the expected values of the various types of random variables considered earlier. First, let $X \sim Bernoulli(p)$. Then

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

Next, let $X \sim Bin(n, p)$. Then

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

Now note that the $k = 0$ term in the above sum is zero, and for $k \geq 1$,

$$\begin{aligned} k\binom{n}{k} &= k\frac{n(n-1)\cdots(n-k+1)}{1 \cdot 2 \cdots (k-1) \cdot k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{1 \cdot 2 \cdots (k-1)} = n\binom{n-1}{k-1}. \end{aligned}$$

Therefore

$$\begin{aligned} E(X) &= \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k(1-p)^{n-k} \\ &= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} \\ &= np\sum_{j=0}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j} \text{ (replacing } j = k-1) \\ &= np(p+1-p)^{n-1} \text{ (applying the binomial formula).} \\ &= np. \end{aligned}$$

[29] From the frequentist viewpoint, $E(X)$ is the average value of X if the experiment is repeated many times. This is justified as follows. If the experiment is repeated many times, then X takes the value x approximate $P(X = x)$ fraction of times. Therefore the sum defining $E(X)$ is approximately the overall average value of X.

[30] If the possible value are all nonnegative and the sum diverges, then $E(X) = \infty$. If the sum is convergent but not absolutely convergent, $E(X)$ is left undefined.

Thus, if $X \sim Bin(n, p)$, then $E(X) = np$. This makes sense intuitively, since a p-coin tossed n times is expected to turn up heads np times on average.

Next, suppose that $X \sim Geo(p)$. Then

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

Again, we use the differentiation trick, recalling that the derivative can be moved inside an infinite sum if the result is absolutely convergent. By the geometric series summation formula,

$$\sum_{k=1}^{\infty} x^k = \frac{1}{1-x} - 1.$$

when $|x| < 1$. Differentiating both sides, we get

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}.$$

Plugging in $x = 1 - p$, and multiplying both sides by p, we have

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

Again, this makes sense intuitively, because a p-coin turns up head p fraction of times, so the first head is expected to occur at toss number $1/p$, on average.

Finally, let us calculate the expected value of a $Poi(\lambda)$ random variable. Let $X \sim Poi(\lambda)$. Then

$$E(X) = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!}.$$

Note that the $k = 0$ term does not contribute, and for $k \geq 1$,

$$\frac{k}{k!} = \frac{1}{(k-1)!}.$$

This gives

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

In view of Proposition 4, this makes perfect sense, since a $Poi(\lambda)$ random variable is approximately a $Bin(n, \lambda/n)$ random variable for large n, and the expected value of a $Bin(n, \lambda/n)$ random variable is $n \cdot \lambda/n = \lambda$, as we know.

Expectation of a function of random variables

The following result is often useful.

Proposition 6. *Let $X_1, \ldots, X_n$ be discrete random variables and $Y = f(X_1, \ldots, X_n)$ for some function f. Then*

$$E(Y) = \sum_{x_1,\ldots,x_n} f(x_1, \ldots, x_n) P(X_1 = x_1, \ldots, X_n = x_n).$$

Proof. Note that

$$E(Y) = \sum_y y P(Y = y).$$

But for any y,

$$P(Y = y) = \sum_{\substack{x_1,\ldots,x_n: \\ f(x_1,\ldots,x_n)=y}} P(X_1 = x_1, \ldots, X_n = x_n).$$

Consequently,

$$\begin{aligned}
E(Y) &= \sum_y y\Bigg(\sum_{\substack{x_1,\ldots,x_n: \\ f(x_1,\ldots,x_n)=y}} P(X_1 = x_1, \ldots, X_n = x_n)\Bigg) \\
&= \sum_y \sum_{\substack{x_1,\ldots,x_n: \\ f(x_1,\ldots,x_n)=y}} y P(X_1 = x_1, \ldots, X_n = x_n) \\
&= \sum_y \sum_{\substack{x_1,\ldots,x_n: \\ f(x_1,\ldots,x_n)=y}} f(x_1, \ldots, x_n) P(X_1 = x_1, \ldots, X_n = x_n) \\
&= \sum_{x_1,\ldots,x_n} f(x_1, \ldots, x_n) P(X_1 = x_1, \ldots, X_n = x_n),
\end{aligned}$$

which completes the proof. □

Linearity of expectation

A very important corollary of Proposition 6 is the following result, known as **linearity of expectation**.

Corollary 3. *Let $X_1, \ldots, X_n$ be discrete random variables. Take any real numbers $a_0, a_1, \ldots, a_n$ and let $Y = a_0 + a_1 X_1 + \cdots + a_n X_n$. Then*

$$E(Y) = a_0 + a_1 E(X_1) + \cdots + a_n E(X_n).$$

Proof. By Proposition 6,

$$\begin{aligned} E(Y) &= \sum_{x_1,\ldots,x_n} (a_0 + a_1 x_1 + \cdots + a_n x_n) P(X_1 = x_1, \ldots, X_n = x_n) \\ &= a_0 \sum_{x_1,\ldots,x_n} P(X_1 = x_1, \ldots, X_n = x_n) \\ &\quad + a_1 \sum_{x_1,\ldots,x_n} x_1 P(X_1 = x_1, \ldots, X_n = x_n) \\ &\quad + \cdots + a_n \sum_{x_1,\ldots,x_n} x_n P(X_1 = x_1, \ldots, X_n = x_n). \end{aligned}$$

Now,

$$\sum_{x_1,\ldots,x_n} P(X_1 = x_1, \ldots, X_n = x_n) = 1,$$

since the events $\{X_1 = x_1, \ldots, X_n = x_n\}$ are disjoint as we vary $x_1, \ldots, x_n$, and their union is Ω. Thus, the first term in the above expression is simply a_0.

Next, notice that for any x_1, the event $\{X_1 = x_1\}$ is the disjoint union of the events $\{X_1 = x_1, \ldots, X_n = x_n\}$ over all $x_2, \ldots, x_n$. Therefore for any x_1,

$$\sum_{x_2,\ldots,x_n} P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1).$$

As a consequence, we have

$$\begin{aligned} &\sum_{x_1,\ldots,x_n} x_1 P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \sum_{x_1} \sum_{x_2,\ldots,x_n} x_1 P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \sum_{x_1} x_1 \biggl(\sum_{x_2,\ldots,x_n} P(X_1 = x_1, \ldots, X_n = x_n) \biggr) \\ &= \sum_{x_1} x_1 P(X_1 = x_1) = E(X_1). \end{aligned}$$

Similarly taking care of the other terms, we get the desired expression for $E(Y)$. □

To see how Corollary 3 greatly simplifies calculations, let us revisit the expected value of a binomial random variable. Let $X \sim Bin(n, p)$. We know that X can be written as $X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are $Bernoulli(p)$ random variables. Therefore

$$E(X) = E(X_1) + \cdots + E(X_n) = np,$$

which is quite a bit simpler than our previous calculation of $E(X)$.

The method of indicators

Let A be an event. The **indicator** of A is a random variable, denoted by 1_A, defined as follows:

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Clearly, 1_A is a $Bernoulli(p)$ random variable, where $p = P(A)$. In particular, $E(1_A) = P(A)$. The **method of indicators** is a technique for evaluating the expected value of a complicated random variable X by finding a way to write X as a sum of indicator variables $1_{A_1}, \ldots, 1_{A_n}$ for some events $A_1, \ldots, A_n$ (that is, X counts the number of i such that A_i happened), and then using linearity of expectation to write

$$E(X) = \sum_{i=1}^{n} E(1_{A_i}) = \sum_{i=1}^{n} P(A_i).$$

Let us now work out some examples to understand what's going on. We have already seen one example of this, namely, when X is the number of heads in n tosses of a p-coin. In that setting, we let A_i be the event that toss i turns up heads. Then $X = \sum_{i=1}^n 1_{A_i}$, and so

$$E(X) = \sum_{i=1}^{n} P(A_i) = np,$$

since $P(A_i) = p$ for each i.

Next, recall the example of inserting n letters randomly into n envelopes. Let X be the number of letters that go into correct envelopes. Previously, we calculated $P(X = 0)$. How can we calculate $E(X)$? To do that, let us define A_i to be the event that letter i goes into the correct envelope, so that $X = \sum_{i=1}^n 1_{A_i}$. If letter i is put into the correct envelope, then the remaining letters can be distributed in $(n-1)!$ ways, which shows that

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

Therefore,

$$E(X) = \sum_{i=1}^{n} P(A_i) = 1.$$

Thus, on average, the number of letters going into correct envelopes is 1.

As our third example, consider an experiment where n balls are dropped independently at random into n boxes (so that multiple balls are allowed to fall into the same box). Let X be the number of empty boxes. What is $E(X)$? The random variable X has no simple distribution, but it is fairly easy to compute $E(X)$ using the method of indicators. Let A_i be the event that box i is empty. Then $X = \sum_{i=1}^{n} 1_{A_i}$.

Now note that, by the fundamental principle of counting, the total number of ways of distributing the balls in the boxes is n^n. On the other hand if box i has to be kept empty, then the number of ways of distributing the balls is $(n-1)^n$. Therefore

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

Thus, we get

$$E(X) = \sum_{i=1}^{n} P(A_i) = n\left(1 - \frac{1}{n}\right)^n.$$

Note that this is approximately ne^{-1} when n is large. In other words, the expected fraction[31] of empty boxes is approximately e^{-1}.

As our last example, let us consider consider the number X of *head runs* in a sequence of n tosses of a p-coin. A head run is simply a continuous sequence of heads. For example, the sequence

$$HHHTTHTHTTTHH$$

has 4 head runs, starting at tosses 1, 6, 8 and 12.

Again, the distribution of X is not easy to determine, but $E(X)$ can be evaluated using the method of indicators. The solution is a bit less obvious in this example than the previous ones. Notice that to count head runs, we simply have to count the number of tosses where a head run began. A head run can begin at toss 1, which is identified by the occurrence of a head on toss 1. Let us call this event A_1. Else, a head run can begin at a toss $i \geq 2$, which is identified by the occurrence of a head on toss i and a tail on toss $i-1$. Let us call this event A_i. This shows that $X = \sum_{i=1}^{n} 1_{A_i}$.

Now, $P(A_1) = p$, and $P(A_i) = p(1-p)$ for $i \geq 2$. Therefore[32]

$$E(X) = \sum_{i=1}^{n} P(A_i) = p + (n-1)p(1-p).$$

When $p = 1/2$, this gives $E(X) = (n+1)/4$.

Variance

The variance of a random variable X is defined as

$$Var(X) = E(X^2) - (E(X))^2.$$

The variance measured the *average squared deviation from the mean*. The meaning of this statement is made clear by the following simple result.

[31] We will see later, when we will discuss laws of large numbers, that the actual fraction X/n is also close to e^{-1} with high probability.

[32] Again, we will see later that the random variable X — and not just $E(X)$ — is close to $np(1-p)$ with high probability when n is large.

Proposition 7. *Let X be a random variable and let $Y = X - E(X)$. Then $Var(X) = E(Y^2)$.*

Proof. Let $a = E(X)$. Then

$$E(Y^2) = E(X^2 - 2aX + a^2) = E(X^2) - 2aE(X) + a^2.$$

But $E(X) = a$, and so $2aE(X) = 2a^2$. Plugging this into the above expression, we get $E(Y^2) = E(X^2) - a^2 = Var(X)$. □

Due to the above alternative expression of $Var(X)$, the square root of the variance is called the **standard deviation** of X. It is a measure of how much X deviates from its expected value on average.

An immediate consequence of Proposition 7 is that if X is a random variable and b, c are two real numbers, then

$$Var(bX + c) = b^2 Var(X).$$

To see this, just note that $bX + c - E(bX + c) = bY$, where $Y = X - E(X)$, and apply Proposition 7.

Let us now work out some examples. First, let $X \sim Bernoulli(p)$. Then we know that $E(X) = p$. But $X^2 = X$ since the only possible values of X are 0 and 1. Therefore $E(X^2)$ is also p. Thus,

$$Var(X) = p - p^2 = p(1 - p).$$

The variance of a $Bin(n, p)$ random variable is $np(1 - p)$. Although it is not very difficult to prove this directly, a much simpler derivation is possible by a method that we will discuss later. So let us postpone this computation.

Next let $X \sim Geo(p)$. A little trick helps in the computation of $E(X^2)$. Observe that

$$\begin{aligned} E(X(X-1)) &= \sum_{k=1}^{\infty} k(k-1)(1-p)^{k-1}p \\ &= p(1-p)\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2}. \end{aligned}$$

But for $|x| < 1$,

$$\begin{aligned} \sum_{k=2}^{\infty} k(k-1)x^{k-2} &= \frac{d^2}{dx^2}\left(\sum_{k=0}^{\infty} x^k\right) \\ &= \frac{d^2}{dx^2}\frac{1}{1-x} = \frac{2}{(1-x)^3}. \end{aligned}$$

Combining the above two observations, we get

$$E(X(X-1)) = p(1-p)\frac{2}{(1-(1-p))^3} = \frac{2(1-p)}{p^2}.$$

But we know that $E(X) = 1/p$. Thus,

$$E(X^2) = E(X(X-1) + X) = \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2}{p^2} - \frac{1}{p}.$$

Finally, we get

$$Var(X) = E(X^2) - (E(X))^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

We can employ the same trick for Poisson. Let $X \sim Poi(\lambda)$. Then

$$\begin{aligned} E(X(X-1)) &= \sum_{k=0}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^k}{(k-2)!} \\ &= e^{-\lambda}\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = e^{-\lambda}\lambda^2 e^{\lambda} = \lambda^2. \end{aligned}$$

Recall that $E(X) = \lambda$. Therefore, we get

$$E(X^2) = E(X(X-1) + X) = \lambda^2 + \lambda,$$

and so

$$Var(X) = E(X^2) - (E(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Thus, the expectation and the variance of a $Poi(\lambda)$ random variable are both equal to λ.

Covariance

The **covariance** of two random variables X and Y is defined as

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

Notice that $Var(X) = Cov(X,X)$. An equivalent expression for covariance is[33]

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))].$$

Given a real number a, let us also denote by a the random variable that always takes the value a. Then $E(a) = a$, and $E(Xa) = aE(X)$ by linearity of expectation. Therefore $Cov(X,a) = 0$.

Note that $Cov(X,Y) = Cov(Y,X)$. Another very important property of covariance is that it is *bilinear*, as shown by the following result.

Proposition 8. *Let $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ be random variables and $a_1, \ldots, a_m, b_1, \ldots, b_n$ be real numbers. Let $U = a_1X_1 + \cdots + a_mX_m$ and $V = b_1Y_1 + \cdots + b_nY_n$. Then*

$$Cov(U,V) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_ib_jCov(X_i,Y_j).$$

[33] Prove the equivalence.

Proof. By the distributive law and the linearity of expectation,

$$E(UV) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j E(X_i Y_j).$$

Similarly,

$$\begin{aligned} E(U)E(V) &= \left(\sum_{i=1}^{m} a_i E(X_i)\right)\left(\sum_{j=1}^{n} b_j E(Y_j)\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j E(X_i)E(Y_j). \end{aligned}$$

Therefore

$$\begin{aligned} Cov(U, V) &= E(UV) - E(U)E(V) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j (E(X_i Y_j) - E(X_i)E(Y_j)) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j Cov(X_i, Y_j), \end{aligned}$$

which completes the proof. □

The following corollary is often useful.

Corollary 4. *Let $X_1, \ldots, X_n$ be random variables and $a_1, \ldots, a_n$ be real numbers. Let $U = a_1 X_1 + \cdots + a_n X_n$. Then*

$$Var(U) = \sum_{i,j=1}^{n} a_i a_j Cov(X_i, X_j).$$

Proof. Recall that $Var(U) = Cov(U, U)$, and apply Proposition 8. □

The above corollary, combined with the method of indicators, allows us to calculate variances of fairly complicated random variables. We will do this a little later, after taking care of one other important matter.

Expectation of a product of independent random variables

When computing covariances, the following fact is often useful.

Proposition 9. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$E(X_1 X_2 \cdots X_n) = E(X_1)E(X_2)\cdots E(X_n).$$

Proof. By Proposition 6,

$$E(X_1 \cdots X_n) = \sum_{x_1,\ldots,x_n} x_1 \cdots x_n P(X_1 = x_1, \ldots, X_n = x_n).$$

But by independence,

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n).$$

Therefore

$$\begin{aligned} E(X_1 \cdots X_n) &= \sum_{x_1,\ldots,x_n} x_1 \cdots x_n P(X_1 = x_1) \cdots P(X_n = x_n) \\ &= \sum_{x_1,\ldots,x_n} (x_1 P(X_1 = x_1)(x_2 P(X_2 = x_2)) \cdots (x_n P(X_n = x_n)) \\ &= \left(\sum_{x_1} x_1 P(X_1 = x_1)\right) \cdots \left(\sum_{x_n} x_n P(X_n = x_n)\right) \\ &= E(X_1) \cdots E(X_n), \end{aligned}$$

which completes the argument. □

The most important corollary of the above result is the following.

Corollary 5. *If X and Y are independent, then $Cov(X, Y) = 0$.*

Proof. By Proposition 9, $E(XY) = E(X)E(Y)$, which immediately implies that $Cov(X, Y) = 0$. □

A corollary of the above corollary is the following.

Corollary 6. *Let $X_1, \ldots, X_n$ be independent random variables and let $a_1, \ldots, a_n$ be real numbers. Let $U = a_1 X_1 + \cdots + a_n X_n$. Then*

$$Var(U) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

Proof. By Corollary 4,

$$Var(U) = \sum_{i,j=1}^{n} a_i a_j Cov(X_i, X_j).$$

But by independence and Corollary 5, $Cov(X_i, X_j) = 0$ whenever $i \neq j$. Since $Cov(X_i, X_i) = Var(X_i)$, this completes the proof. □

As an application of Corollary 6, let us now calculate the variance of a binomial random variable. Let $X \sim Bin(n, p)$, and let us write $X = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are independent $Bernoulli(p)$ random variables. Then by Corollary 6,

$$Var(X) = \sum_{i=1}^{n} Var(X_i).$$

But we know that $Var(X_i) = p(1 - p)$ for each i. Therefore

$$Var(X) = np(1 - p).$$

The method of indicators for variance

The method of indicators, in conjunction with Corollary 4, gives a powerful method for calculating the variances of complicated random variables. As an illustrative example, let us revisit the problem of dropping n balls independently at random into n boxes. Let X be the number of empty boxes. Let A_i be the event that box i is empty, so that $X = \sum_{i=1}^n 1_{A_i}$. Then by Corollary 4,

$$Var(X) = \sum_{i,j=1}^{n} Cov(1_{A_i}, 1_{A_j}). \tag{5}$$

Recall that 1_{A_i} is a $Bernoulli(p)$ random variable where

$$p = \left(1 - \frac{1}{n}\right)^n.$$

So for each i,

$$\begin{aligned} Cov(1_{A_i}, 1_{A_i}) &= Var(1_{A_i}) = p(1-p) \\ &= \left(1 - \frac{1}{n}\right)^n \left[1 - \left(1 - \frac{1}{n}\right)^n\right]. \end{aligned}$$

Next, note that for any $i \neq j$,

$$1_{A_i} 1_{A_j} = 1_{A_i \cap A_j}$$

because the left side is 1 if both A_i and A_j happen, and 0 otherwise. Therefore

$$E(1_{A_i} 1_{A_j}) = E(1_{A_i \cap A_j}) = P(A_i \cap A_j).$$

Now, $A_i \cap A_j$ is the event that boxes i and j are both empty. Therefore by counting[34],

$$P(A_i \cap A_j) = \frac{(n-2)^n}{n^n} = \left(1 - \frac{2}{n}\right)^n.$$

Thus, for $i \neq j$,

$$Cov(1_{A_i}, 1_{A_j}) = \left(1 - \frac{2}{n}\right)^n - \left(1 - \frac{1}{n}\right)^{2n}.$$

Now, in (5), there are n terms where $i = j$, and $n(n-1)$ terms where $i \neq j$. Thus,

$$\begin{aligned} Var(X) = n&\left(1 - \frac{1}{n}\right)^n \left[1 - \left(1 - \frac{1}{n}\right)^n\right] \\ &+ n(n-1)\left[\left(1 - \frac{2}{n}\right)^n - \left(1 - \frac{1}{n}\right)^{2n}\right]. \end{aligned}$$

This is the exact value of the variance of X. It is interesting to figure out the asymptotic behavior of this as $n \to \infty$. If you like a challenge, you can try to show that this behaves like[35] $n(e-2)/e^2$ for large n.

[34] Figure this out on your own if it is not clear.

[35] More precisely, show that

$$\lim_{n\to\infty} \frac{Var(X)}{n(e-2)/e^2} = 1.$$

Laws of large numbers

Elementary inequalities

Until now, we have been calculating exact values of probabilities, expectations, etc. Often, however, it is hard to calculate exactly. If that is the case, it helps to give upper and lower bounds. These are known as *inequalities*. The simplest inequality in probability theory is the following. Let A and B be events such that A implies B. Then set theoretically, $A \subseteq B$. A consequence of this is that B is the disjoint union of A and $B \setminus A$, which shows that

$$P(A) \leq P(B).$$

Another very simple and useful inequality is that if X is a nonnegative random variable (which means that $X(\omega) \geq 0$ for all ω), then

$$E(X) \geq 0.$$

For discrete random variables (which is the only kind discussed so far), this is obvious from the definition of $E(X)$.

Let X and Y be two random variables. We say $X \leq Y$ if $X(\omega) \leq Y(\omega)$ for all ω. If $X \leq Y$, then the random variable $Y - X$ is nonnegative, and therefore $E(Y - X) \geq 0$. By linearity of expectation, this means that

$$E(X) \leq E(Y). \tag{6}$$

An important consequence of this inequality is the **union bound**, also known as **Boole's inequality** or **Bonferroni's inequality**, which says that for any events $A_1, \ldots, A_n$,

$$P(A_1 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} P(A_i).$$

To prove this, let $X = 1_{A_1 \cup \cdots \cup A_n}$ and $Y = \sum_{i=1}^{n} 1_{A_i}$. Then clearly, $X \leq Y$, because $X(\omega)$ is either 0 or 1 for each ω, $Y(\omega)$ is nonnegative, and if $X(\omega) = 1$, then $1_{A_i}(\omega) = 1$ for some i. Thus

$$P(A_1 \cup \cdots \cup A_n) = E(X) \leq E(Y) = \sum_{i=1}^{n} P(A_i).$$

Note that if the sample space is finite or countable, this has an easy proof from the observation that $\sum P(A_i)$ sums $P(\omega)$ for each ω in the union at least once.

Another important consequence of (6) is the following. For any random variable X, $X \le |X|$ always, so we have $E(X) \le E|X|$. Since $-X$ is also $\le |X|$, we have $-E(X) = E(-X) \le E|X|$. Thus,

$$|E(X)| \le E|X|.$$

A different upper bound on $|E(X)|$ is obtained by using the fact that $Var(X) \ge 0$. Note that this inequality can be rewritten as $E(X^2) \ge (E(X))^2$, which gives

$$|E(X)| \le \sqrt{E(X^2)}.$$

The above inequality has the following generalization.

Proposition 10. *For any random variable X and any real number $p \ge 1$,*

$$|E(X)| \le (E|X|^p)^{1/p}.$$

Proof. For any $y \ge 0$, we claim that

$$y \le \frac{y^p}{p} + 1 - \frac{1}{p}.$$

To prove this, observe that the two sides are equal when $y = 1$, and the derivative of the right side is greater than that of the left side when $y \ge 1$, and less than that of the left side when $y \le 1$. Thus, the right side must always lie above the left side.

This inequality implies that if Y is a nonnegative random variable with $E(Y^p) = 1$, then

$$E(Y) \le \frac{E(Y^p)}{p} + 1 - \frac{1}{p} = 1.$$

Now take any random variable X and apply the above inequality to the nonnegative random variable $Y = |X|/(E|X|^p)^{1/p}$, which satisfies $E(Y^p) = 1$. Finally, apply $|E(X)| \le E|X|$. □

Markov's inequality

The following important result is the first nontrivial inequality in probability theory, commonly known as **Markov's inequality**.

Theorem 1 (Markov's inequality)**.** *Let X be a nonnegative random variable. Then for any $t > 0$,*

$$P(X \ge t) \le \frac{E(X)}{t}.$$

Proof. Let $Y = 1_{\{X \geq t\}}$. Let $Z = X/t$. Then note that if $Y(\omega) = 0$, then $Y(\omega) \leq Z(\omega)$ since Z is always nonnegative. On the other hand, if $Y(\omega) = 1$, then $X(\omega) \geq t$, and hence $Z(\omega) \geq 1 = Y(\omega)$. Thus, $Y \leq Z$ always, and therefore $E(Y) \leq E(Z)$. But $E(Y) = P(X \geq t)$ and $E(Z) = E(X)/t$ by linearity of expectation. □

As a simple application of Markov's inequality, let us revisit the secretary problem. Let n letters be inserted randomly into n marked envelopes, and let X be the number of letters that go into the correct envelopes. We showed that $E(X) = 1$ in the previous chapter. Therefore by Markov's inequality[36],

$$P(X \geq k) \leq \frac{1}{k}$$

for each k. For example, the chance that 10 or more letters get inserted correctly is ≤ 0.1.

Chebyshev's inequality

The following result, known as **Chebyshev's inequality**, shows that the difference between the value of a random variable and its expected value is unlikely to be a large multiple of its standard deviation. The stunning generality of this result makes it extremely useful in theory and practice.

Theorem 2 (Chebyshev's inequality)**.** *Let X be any random variable. Then for any $t > 0$,*

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

Proof. Let $Y := (X - E(X))^2$. Then by Markov's inequality,

$$P(|X - E(X)| \geq t) = P((X - E(X))^2 \geq t^2) = P(Y \geq t^2) \leq \frac{E(Y)}{t^2}.$$

But by Proposition 7, $E(Y) = Var(X)$. This completes the proof. □

To understand the meaning of Chebyshev's inequality, consider the following. Let s be the standard deviation of X. Then Chebyshev's inequality gives us that for any $L > 0$,

$$P(|X - E(X)| \geq Ls) \leq \frac{Var(X)}{L^2 s^2} = \frac{1}{L^2}.$$

Thus, for example, the chance that X deviates from $E(X)$ by more than 5 times its standard deviation is $\leq 1/25$. Note that this is true for any X. For a specific X, the probability may be much smaller.

[36] The actual probability of this event is much less than $1/k$, but Markov's inequality gives a simple way of getting a meaningful bound.

As a concrete example, consider a fair coin tossed one million times. Let X be the number of heads. Then $E(X) = 500000$ and $Var(X) = 250000$. Therefore the standard deviation of X is the square-root of 250000, which is 500. By Chebyshev's inequality, the probability of deviating from the mean by more than 10 standard deviations is ≤ 0.01. In other words,

$$P(495000 \le X \le 505000) \ge 0.99.$$

Convergence in probability

Suppose that a random variable X lies within the interval $[1.99, 2.01]$ with probability 0.99. Then we may say that X is approximately equal to 2 with high probability. Generalizing this, suppose that we have a sequence of random variables $X_1, X_2, \ldots$ such that for each n, X_n belongs to the interval $[2 - 1/n, 2 + 1/n]$ with probability $1 - 1/n$. Then, as n increases, it becomes more and more likely that X_n is close to 2. We may reasonably declare that X_n 'converges' to 2. However, this notion of convergence is quite different than the convergence of real numbers. In fact, it is so different that a new definition is necessary.

Definition 1. Let $X_1, X_2, \ldots$ be a sequence of random variables, and c be a real number. We say that $X_n \to c$ **in probability** if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - c| \ge \varepsilon) = 0.$$

For example, if $X_n \sim \mathit{Bernoulli}(1/n)$, then $X_n \to 0$ in probability, because for any given $\varepsilon > 0$, eventually $1/n$ becomes smaller than ε, and so $P(|X_n - 0| \ge \varepsilon) = 1/n$ for all large enough n.

The following result gives a very useful method for proving convergence in probability.

Theorem 3. *Let $X_1, X_2, \ldots$ be a sequence of random variables. If $E(X_n)$ converges to a real number c, and $Var(X_n) \to 0$, then $X_n \to c$ in probability.*

Proof. Take any $\varepsilon > 0$. By Markov's inequality,

$$P(|X_n - c| \ge \varepsilon) = P((X_n - c)^2 \ge \varepsilon^2) \le \frac{E(X_n - c)^2}{\varepsilon^2}.$$

Let $a_n = E(X_n)$. Then

$$\begin{aligned} E(X_n - c)^2 &= E(X_n - a_n + a_n - c)^2 \\ &= E(X_n - a_n)^2 + 2(a_n - c)E(X_n - a_n) + (a_n - c)^2. \end{aligned}$$

But $E(X_n - a_n)^2 = Var(X_n)$ and $E(X_n - a_n) = 0$. Thus,

$$E(X_n - c)^2 = Var(X_n) + (a_n - c)^2,$$

and so

$$P(|X_n - c| \geq \varepsilon) \leq \frac{Var(X_n) + (a_n - c)^2}{\varepsilon^2}.$$

But $Var(X_n) \to 0$ and $a_n \to c$. Thus,

$$\lim_{n\to\infty} P(|X_n - c| \geq \varepsilon) = 0.$$

But this holds for any $\varepsilon > 0$. Thus, $X_n \to c$ in probability. □

The weak law of large numbers

We have learnt that the expected value of a random variable X should be thought of as the 'long term average value' of X in many repeated experiments. However, we have not yet seen a mathematical justification of this claim. The following result, known as the **weak law of large numbers**, gives such a justification[37]. Recall that a sequence of random variables is called independent and identically distributed (i.i.d.) if the variables are independent and they all have the same distribution.

Theorem 4 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with expected value μ and finite variance. For each n, let*

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

be the average of the first n of these variables. Then as $n \to \infty$, $\overline{X}_n \to \mu$ in probability.

Proof. Since $X_1, X_2, \ldots$ are independent, Corollary 6 from the previous chapter gives

$$Var(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i).$$

Since $X_1, X_2, \ldots$ are identically distributed, $Var(X_i) = Var(X_1)$ for all i. Thus,

$$Var(\overline{X}_n) = \frac{Var(X_1)}{n},$$

which tends to 0 as $n \to \infty$. Also, by linearity of expectation, $E(\overline{X}_n) = \mu$ for each n. Therefore by Theorem 3, $\overline{X}_n \to \mu$ in probability. □

[37] We have to be careful about calling this a justification. The frequentist view of probability *assumes* that probabilities and expectations are long term averages, whereas the law large numbers seems to be *proving* this claim. Have we then been able to prove what we assumed earlier? Not really. Consider, for example, a fair coin tossed n times. The weak law of large numbers says that with high probability, the number of heads is close to $n/2$. But this is not a sequence of repeated experiments — we are now in the framework of a *single* experiment of n tosses, where all possible sequences are equally likely. In other words, we now have a *new* assumption: If this experiment of n tosses is repeated many times, then each sequence will occur 2^{-n} fraction of those times. Under this assumption, the theorem implies that most of the times this experiment is conducted, we will observe approximately $n/2$ heads. We do not really have a framework for repeated experiments in probability theory; we always work with a single experiment. So we can never mathematically justify the assumption that the probability of an outcome is the fraction of times that outcome will occur if the experiment is repeated many times.

Two applications of the weak law

The simplest application of the weak law of large numbers is in coin tossing. Suppose a p-coin is tossed n times. It is a direct consequence of the weak law that the fraction of heads converges in probability to p as $n \to \infty$. To see this, just let X_i be 1 if toss i turns up heads and 0 otherwise. Then $X_1, X_2, \ldots$ are i.i.d. $Bernoulli(p)$ random variables, and $\overline{X}_n$ is the fraction of heads in the first n tosses.

Next, consider the number of tosses of a p-coin that are required to get n heads. This can be seen as a sum of independent $Geo(p)$ random variables $Y_1, Y_2, \ldots, Y_n$, where Y_i is the number of tosses required to get the i^{th} head after getting the $(i-1)^{\text{th}}$ head. (This was established as an application of Proposition 5 in the chapter on discrete random variables.) Thus, if Z_n is the time to get the n^{th} head, then the weak law tells us that $Z_n/n \to 1/p$ in probability as $n \to \infty$.

Example: Number of empty boxes

The weak law of large numbers applies only to averages of i.i.d. random variables. Many other kinds of random variables converge in probability. All of these examples can be called 'laws of large numbers'. For one such, let us consider a familiar example from the previous chapter. Let n balls be dropped independently at random into n boxes, and let X_n be the number of empty boxes[38]. We calculated that

$$E(X_n) = n\left(1 - \frac{1}{n}\right)^n$$

and

$$\begin{aligned}Var(X_n) &= n\left(1-\frac{1}{n}\right)^n\left[1-\left(1-\frac{1}{n}\right)^n\right] \\ &\qquad + n(n-1)\left[\left(1-\frac{2}{n}\right)^n - \left(1-\frac{1}{n}\right)^{2n}\right].\end{aligned}$$

Let $Y_n = X_n/n$ be the fraction of empty boxes. Then by the above formulas,

$$E(Y_n) = \left(1-\frac{1}{n}\right)^n \to e^{-1} \text{ as } n \to \infty,$$

and

$$\begin{aligned}Var(Y_n) &= \frac{1}{n^2}Var(X_n) \\ &= \frac{1}{n}\left(1-\frac{1}{n}\right)^n\left[1-\left(1-\frac{1}{n}\right)^n\right] \\ &\qquad + \left(1-\frac{1}{n}\right)\left[\left(1-\frac{2}{n}\right)^n - \left(1-\frac{1}{n}\right)^{2n}\right].\end{aligned}$$

[38] Previously, we denoted this by X without the subscript, but since we will be taking $n \to \infty$ here, adding a subscript is necessary.

Since $(1-1/n)^n \to e^{-1}$, $(1-2/n)^n \to e^{-2}$ and $(1-1/n)^{2n} \to e^{-2}$, it follows that $Var(Y_n) \to 0$ as $n \to \infty$. Thus, by Theorem 3, $Y_n \to e^{-1}$ in probability.

Example: Number of head runs

Let X_n be the number of head runs in a sequence of n tosses of a p-coin. From the previous chapter, we know that

$$E(X_n) = p + (n-1)p(1-p).$$

Let A_1 be the event that the first toss comes up heads, and for each $i \geq 2$, let A_i be the event that the toss i is heads and toss $i-1$ is tails. Then we know that $X_n = \sum_{i=1}^n 1_{A_i}$. Thus, by Corollary 4,

$$Var(X_n) = \sum_{i=1}^n \sum_{j=1}^n Cov(1_{A_i}, 1_{A_j}).$$

Now note that the events A_i and A_j are independent if $|i-j| \geq 2$, and therefore[39] the covariance is zero under this condition (by Corollary 5). Thus, for each i, $Cov(1_{A_i}, 1_{A_j})$ may be nonzero for at most 3 values of j. Even if the covariance is nonzero, it is bounded by 1, since

$$\begin{aligned} Cov(1_{A_i}, 1_{A_j}) &= E(1_{A_i}1_{A_j}) - E(1_{A_i})E(1_{A_j}) \\ &\leq E(1_{A_i}1_{A_j}) = P(A_i \cap A_j) \leq 1. \end{aligned}$$

Combining these observations, we see that

$$Var(X_n) \leq 3n.$$

Therefore, if $Y_n = X_n/n$, then $E(Y_n) \to p(1-p)$ and $Var(Y_n) \to 0$ as $n \to \infty$. By Theorem 3, this tells us that $Y_n \to p(1-p)$ in probability.

Example: The coupon collector's problem

Suppose that there are n types of coupons, and each time you buy an item, you get a randomly selected type of coupon. In particular, you may get the same type of coupon multiple times. Let n be the number of times you have to buy before you acquire all n types of coupons. Figuring out the behavior of T_n for large n is a classical problem in probability theory. We will now show that T_n is approximately $n \log n$ with high probability, in the sense that

$$\frac{T_n}{n \log n} \to 1 \text{ in probability as } n \to \infty. \tag{7}$$

[39] We also need the fact that if A and B are independent events, the 1_A and 1_B are independent random variables. Try to prove this.

Let X_i be the number of trials required to get the i^{th} new type of coupon after having obtained $i-1$ distinct types of coupons. Note that X_1 is always equal to 1, but $X_2, X_3, \ldots, X_n$ are not deterministic. We claim that $X_1, \ldots, X_n$ are independent random variables, with $X_i \sim Geo((n-i+1)/n)$ for each i. To prove this, consider the probability

$$P(X_1 = x_1, \ldots, X_n = x_n)$$

for some positive integers $x_1, \ldots, x_n$. Let $m_0 = 0$ and $m_i = x_1 + \cdots + x_i$ for $i = 1, \ldots, n$, so that m_i is the time at which the i^{th} new coupon is obtained. The event $\{X_1 = x_1, \ldots, X_n = x_n\}$ is an event involving the first m_n trials. By assumption, all possible outcomes of the first m_n trials are equally likely. The number of all possible outcomes is n^{m_n}. Therefore the above probability can be obtained simply by counting the number of outcomes that result in the event $\{X_1 = x_1, \ldots, X_n = x_n\}$ and dividing by n^{m_n}.

To construct an outcome belonging to this event, we can first choose the order in which the n types of coupons are obtained, which gives us the coupons obtained at times $m_1, \ldots, m_n$. This can be done in $n!$ ways. Once this is determined, there are $i-1$ ways of choosing coupons for each time point $j \in \{m_{i-1}+1, \ldots, m_i - 1\}$, since each such coupon has to belong to the set of $i-1$ coupons chosen by time m_{i-1}. Thus,

$$P(X_1 = x_1, \ldots, X_n = x_n) = \frac{n! \prod_{i=1}^n (i-1)^{x_i - 1}}{n^{m_n}}.$$

(Recall the $\prod$ is the sign for product, just as $\sum$ is the sign for sum. The term for $i = 1$ is 0 if $x_i > 1$ and 1 otherwise.) An easy inspection reveals that the right side can be written as

$$\prod_{i=1}^n \left(1 - \frac{n-i+1}{n}\right)^{x_i - 1} \frac{n-i+1}{n}. \tag{8}$$

By Proposition 5, this shows that $X_1, \ldots, X_n$ are independent random variables, with $X_i \sim Geo((n-i+1)/n)$.

There is also a more intuitive way to arrive at the above conclusion. Note that

$$\begin{aligned} &P(X_1 = x_1, \ldots, X_n = x_n) \\ &= P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}). \end{aligned}$$

Take any i. Given that $X_1 = x_1, \ldots, X_{i-1} = x_{i-1}$, we know that $i-1$ distinct coupons have been found up to time $x_1 + \cdots + x_{i-1}$. Irrespective of the identities of these coupons, each subsequent trial is likely to yield a new coupon with probability $(n-i+1)/n$. Thus, the waiting time for the next new coupon should follow a $Geo((n-i+1)/n)$ distribution. This gives[40]

$$P(X_i = x_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}) = \left(1 - \frac{n-i+1}{n}\right)^{x_i - 1} \frac{n-i+1}{n},$$

which again yields the formula (8).

It is now easy to prove (7). By the independence of the X_i's and the formulas for the expected value and variance of geometric random variables,

$$\begin{aligned} E(T_n) &= \sum_{i=1}^n E(X_i) \\ &= \sum_{i=1}^n \frac{n}{n-i+1} = n\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right), \end{aligned}$$

and

$$\begin{aligned} Var(T_n) &= Var\left(\sum_{i=1}^n X_i\right) \\ &= \sum_{i=1}^n Var(X_i) = \sum_{i=1}^n \frac{n(i-1)}{(n-i+1)^2} \\ &\le n^2\left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2}\right). \end{aligned}$$

Now recall from basic calculus and real analysis[41] that

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=1}^n \frac{1}{i} = 1,$$

and

$$\sum_{i=1}^\infty \frac{1}{i^2} < \infty.$$

These results, when combined with the above formulas for $E(T_n)$ and $Var(T_n)$, yield that

$$\lim_{n\to\infty} E\left(\frac{T_n}{n\log n}\right) = 1$$

and

$$\lim_{n\to\infty} Var\left(\frac{T_n}{n\log n}\right) = 0.$$

By Theorem 3, this completes the proof of (7).

[40] This kind of argument is acceptable only if the writer and the reader are both experienced in probability theory. Unless backed by experience, this style of proof can lead to erroneous conclusions. The style of the first argument is less prone to errors.

[41] To see this quickly, note that the step function which equals $1/i$ in the interval $[i, i+1]$ is lower bounded by the function $1/x$ in $[1,\infty)$ and upper bounded by $1/(x-1)$ in $[2,\infty)$. This gives

$$\int_1^n \frac{dx}{x} \le \sum_{i=1}^n \frac{1}{i} \le 1 + \int_2^n \frac{dx}{x-1}.$$

The integrals on both sides are asymptotic to $\log n$. One can argue similarly for $\sum i^{-2}$ using the function x^{-2}.

Continuous random variables

Probability density function

Recall that a random variable is simply a function from the sample space into the real line. A random variable X is called *continuous* if there is a function $f : \mathbb{R} \to [0, \infty)$ such that for any $-\infty \le a \le b \le \infty$,

$$P(a < X \le b) = \int_a^b f(x)dx. \tag{9}$$

The function f is called the **probability density function (p.d.f.)** of X. Note that a p.d.f. is by definition nonnegative, and

$$\int_{-\infty}^{\infty} f(x)dx = P(-\infty < X \le \infty) = 1.$$

Observe that at any x where f is continuous,

$$f(x) = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} f(y)dy = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} P(x - \varepsilon < X \le x + \varepsilon). \tag{10}$$

Thus, if x is a continuity point of f, then $f(x)$ can be interpreted as the chance that X belongs to a small interval centered at x, divided by the length of the interval[42].

Let X be a continuous random variable with p.d.f. f. It follows from equation (9) is that for any subset[43] A of the real line,

$$P(X \in A) = \int_A f(x)dx. \tag{11}$$

The derivation of (11) from (9) needs measure theory. The rough idea is that any subset of $\mathbb{R}$ that is not too weird can be approximately represented as a union of disjoint intervals, and then (9) can be applied to each interval and the results added up to get (11).

Another important fact[44] is that for any real number x,

$$P(X = x) = \int_x^x f(y)dy = 0.$$

A consequence of this is that for any a and b, the probabilities of X belonging to the intervals $[a, b]$, $(a, b]$, $[a, b)$ and (a, b) are all the same. For example,

$$P(X \in [a, b]) = P(X = a) + P(X \in (a, b]) = P(X \in (a, b]).$$

[42] Without continuity, however, we cannot make this claim. For example, if f is the p.d.f. of X, and we change the value of f at a single point z to get a new function g, then equation (9) is also satisfied with g. Therefore g can also be considered to be a p.d.f. of X, although f and g differ at z.

[43] Actually, this holds only for **Borel measurable** subsets, but a discussion of that requires measure theory. Fortunately, non-Borel sets are extremely rare and strange.

[44] This may look strange, because it says that the chance of X being exactly equal to x is zero for any x. But if the experiment is carried out, X will take on *some* value. So if the chance of X being equal to any given value is zero, how can it take on some value when the experiment is conducted? The reason is that for an *uncountable* collection of disjoint events, the probability of the union need not be equal to the sum of the probabilities. This law holds only for unions of a finite or countably infinite number of disjoint events.

Construction of continuous random variables

Any random variable that arises in practice is discrete, because any measurement is only up to a certain number of places of decimal. Continuous random variables are idealized mathematical objects. Typically, they arise as limits or infinite sums of discrete random variables. The following is an example.

Let $X_1, X_2, \ldots$ be an infinite sequence of i.i.d. $Ber(1/2)$ random variables. Define

$$X = \sum_{i=1}^{\infty} 2^{-i} X_i. \tag{12}$$

Note that the series on the right is convergent (which implies that X is well-defined), and the limit always lies in the interval $[0, 1]$. Moreover, note that the decimal expansion of X is $0.X_1X_2\ldots$, which implies that X can take on any value in $[0, 1]$. It is possible to show[45] that for any $0 \le a \le b \le 1$,

$$P(a < X \le b) = b - a. \tag{13}$$

This implies that X is a continuous random variable with p.d.f.

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

A random variable X with above p.d.f. is said to be *uniformly distributed* over $[0, 1]$. We will revisit uniform random variables later.

Given a probability density function, there is a standard mathematical construction of a random variable with that p.d.f. This goes as follows. Let $f : \mathbb{R} \to [0, \infty)$ be a function whose integral is well-defined[46]. Suppose further that

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

Then it is possible to construct a random variable X with p.d.f. f. Take $\Omega = \mathbb{R}$. For each $A \subseteq \Omega$, define

$$P(A) = \int_A f(x)dx, \tag{14}$$

provided that the integral makes sense[47]. Then $P(A) \ge 0$ for every A, $P(\emptyset) = 0$, and $P(\Omega) = 1$. Furthermore, for any disjoint $A_1, \ldots, A_n$,

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i),$$

and this identity holds even for countably infinite collections of disjoint events[48]. You can check that this property is sufficient to derive

[45] The proof goes as follows. Take any positive integers n and k such that $k \le 2^n$. Then note that the event $(k-1)/2^n < X \le k/2^n$ happens if and only if $X_1, \ldots, X_n$ take certain specific values, and so

$$P((k-1)/2^n < X \le k/2^n) = 2^{-n}.$$

An element of $[0, 1]$ is called a *dyadic rational* if it is of the form $k/2^n$ for some integers n and k. If a and b are dyadic rationals, then the interval $(a, b]$ is the union of disjoint intervals of the form $((k-1)/2^n, k/2^n]$. Therefore by the above equation,

$$P(a < X \le b) = b - a.$$

Approximating any $a, b \in [0, 1]$ by sequences of dyadic rationals approaching a and b from above and below, it is now easy to establish (13).

[46] The real meaning of this involves measure theory. For now, you may think of it as saying that the Riemann integral of f over any interval is well-defined.

[47] This makes sense only if A is a Borel set, but that is beyond the scope of this discussion.

[48] Again, this requires measure theory.

all that we have proved until now about probabilities of events. Thus it is mathematically consistent to imagine that there is an experiment whose set of outcomes is $\Omega = \mathbb{R}$, and the probability of any event A is given by (14). Now define $X : \Omega \to \mathbb{R}$ and $X(\omega) = \omega$. Then for any event A,

$$P(X \in A) = P(\{\omega : X(\omega) \in A\}) = P(A) = \int_A f(x)dx.$$

This shows that f is the p.d.f. of X.

Here are three types of continuous random variables that commonly arise in applications.

- **Uniform random variables:** Given $-\infty < a < b < \infty$, we say that $X \sim Uniform[a,b]$ (or $Unif[a,b]$) if the p.d.f. of X is

  $$f(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a,b], \\ 0 & \text{otherwise.} \end{cases}$$

- **Exponential random variables:** Given $\lambda > 0$, we say that $X \sim Exponential(\lambda)$ (or $Exp(\lambda)$) if the p.d.f. of X is

  $$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

  The $Exp(1)$ distribution is sometimes called the *standard exponential distribution.*

- **Normal random variables:** Given $\mu \in \mathbb{R}$ and $\sigma > 0$, we say that $X \sim Normal(\mu, \sigma^2)$ (or $N(\mu, \sigma^2)$) if the p.d.f. of X is

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

  The $N(0,1)$ distribution is sometimes called the *standard normal distribution.* Normal random variables are also called **Gaussian random variables**. The normal p.d.f. is famously known as the **bell-shaped curve**, because, as the name suggests, its graph is shaped like a bell. A normal random variable with parameters μ and σ^2 is said to be **centered** if $\mu = 0$.

It is easy to verify that the probability density functions for exponential and uniform random variables indeed integrate to 1 when integrated over the whole real line. This is not obvious for the normal p.d.f. The proof goes as follows. First, observe that by the change of variable $y = (x - \mu)/\sigma$,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

Let I denote the integral on the right. Then

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx\right)\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy\right) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dxdy. \end{aligned}$$

Now recall the use of polar coordinates to evaluate integrals over $\mathbb{R}^2$: For any $f : \mathbb{R}^2 \to \mathbb{R}$,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_0^{\infty}\int_0^{2\pi} f(r\cos\theta, r\sin\theta) r d\theta dr.$$

Applying this to the function $f(x,y) = e^{-(x^2+y^2)/2}$, we get

$$\begin{aligned} I^2 &= \frac{1}{2\pi}\int_0^{\infty}\int_0^{2\pi} e^{-r^2/2} r d\theta dr \\ &= \int_0^{\infty} e^{-r^2/2} r dr = 1. \end{aligned}$$

Thus, the p.d.f. of a normal random variable indeed integrates to 1.

Cumulative distribution function

The cumulative distribution function (c.d.f.) of a continuous random variable X is defined just like that of a discrete random variable, that is,

$$F(x) = P(X \le x).$$

Note that

$$F(x) = P(-\infty < X \le x) = \int_{-\infty}^{x} f(y)dy.$$

By the fundamental theorem of calculus, this implies that

$$f(x) = F'(x),$$

where F' is the derivative of F.

Conversely, if the c.d.f. F of a random variable X is differentiable, then for any $-\infty \le a \le b \le \infty$,

$$\begin{aligned} P(a < X \le b) &= P(X \le b) - P(X \le a) \\ &= F(b) - F(a) \\ &= \int_a^b F'(x)dx, \end{aligned}$$

which shows that X is a continuous random variable with p.d.f. F'. This fact is sometimes useful for computing probability density functions, such as in the following example. Let $X \sim Unif[0,1]$ and $Y = X^2$. What is the p.d.f. of Y? First, note that Y can only take values in $[0,1]$, which implies that the p.d.f. must be 0 outside this interval. Take any $y \in [0,1]$. Then the c.d.f. of Y at y is

$$\begin{aligned} P(Y \le y) &= P(X^2 \le y) = P(X \le \sqrt{y}) \\ &= \int_0^{\sqrt{y}} dx = \sqrt{y}. \end{aligned}$$

Therefore the p.d.f. of Y at y is

$$\frac{d}{dy}\sqrt{y} = \frac{1}{2\sqrt{y}}.$$

The above method generalizes to any monotone function of a continuous random variable. This is the topic of the next section.

Change of variable formula

Let X be a continuous random variable with p.d.f. f. Let $u : \mathbb{R} \to \mathbb{R}$ be a strictly increasing or decreasing function. Let $Y = u(X)$. We will now calculate the p.d.f. of Y. Since u is a strictly increasing or decreasing function, it has an inverse[49]. Let us call it v, so that $X = v(Y)$. Moreover, if u is increasing then v is also increasing, and if u is decreasing then v is decreasing. Suppose that u and v are increasing, and also that they are differentiable. Then for any y,

$$P(Y \le y) = P(v(Y) \le v(y)) = P(X \le v(y)) = F(v(y)),$$

where F is the c.d.f. of X. Therefore the p.d.f. of Y is

$$\frac{d}{dy}F(v(y)) = F'(v(y))v'(y) = f(v(y))v'(y).$$

If u and v are decreasing, then a similar argument shows that the p.d.f. of Y is

$$-f(v(y))v'(y).$$

Combining the two scenarios, we see that whenever u is strictly monotone and $v = u^{-1}$, and v is differentiable, the p.d.f. of Y is

$$g(y) = f(v(y))|v'(y)|. \tag{15}$$

This is known as the change of variable formula for probability density functions. We have seen one example of this earlier. Here is another. Let $X \sim Unif[0,1]$, and $Y = -\log X$. The range of possible values of Y is $[0,\infty)$. Since $X = e^{-Y}$ and the p.d.f. of X is 1 everywhere on $[0,1]$, the change of variable formula implies that the p.d.f. of Y on $[0,\infty)$ is

$$g(y) = \left|\frac{d}{dy}e^{-y}\right| = e^{-y}.$$

[49] The inverse function, u^{-1}, has the property that $u^{-1}(u(x)) = x$.

In other words, $Y \sim Exp(1)$.

As a second example, let $X \sim Exp(\lambda)$ and $Y = \alpha X$. Then $X = v(Y)$, where $v(y) = y/\alpha$. Note that $v'(y) = 1/\alpha$ for all y. Therefore the p.d.f. of Y is

$$g(y) = \frac{\lambda}{\alpha} e^{-\lambda y/\alpha}.$$

Thus, $Y \sim Exp(\lambda/\alpha)$.

Similarly, if $X \sim N(\mu, \sigma^2)$, and $Y \sim \alpha + \beta X$, the change of variable formula implies that the p.d.f. of Y is

$$g(y) = \frac{1}{\sqrt{2\pi}\sigma|\beta|} e^{-(y-\alpha-\beta\mu)^2/2\sigma^2\beta^2},$$

which shows that

$$Y \sim N(\alpha + \beta\mu, \beta^2\sigma^2). \tag{16}$$

In particular, this shows that if $X \sim N(0, 1)$, then $\alpha + \beta X \sim N(\alpha, \beta^2)$.

The argument used to derive (15) can be extended to functions that are not monotone. For example, let $X \sim N(0, 1)$, and $Y = |X|$. Since Y is a nonnegative random variable, its p.d.f. is zero at any negative y. For any $y \geq 0$,

$$\begin{aligned} P(Y \leq y) &= P(|X| \leq y) = P(-y \leq X \leq y) \\ &= \int_{-y}^{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \int_{0}^{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx + \int_{-y}^{0} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \end{aligned}$$

If we make the change of variable $z = -x$ in the second integral, it becomes

$$\int_{0}^{y} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Thus,

$$P(Y \leq y) = 2 \int_{0}^{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Differentiating this, we see that the p.d.f. of Y at y is

$$\frac{2}{\sqrt{2\pi}} e^{-y^2/2}.$$

If we had naively applied (15) to this example, we would have missed the factor 2.

Joint probability density function

Suppose that $X_1, \ldots, X_n$ are continuous random variables defined on the same sample space. The n-tuple $X = (X_1, \ldots, X_n)$ is called a

random vector. Note that X is a function from the sample space Ω into $\mathbb{R}^n$. The probability density function of X, also called the **joint p.d.f.** of $X_1, \ldots, X_n$, is a function $f : \mathbb{R}^n \to [0, \infty)$ such that for any[50] subset $A \subseteq \mathbb{R}^n$,

$$P(X \in A) = \int_A f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

Note that the integral of f over the whole of $\mathbb{R}^n$ must be 1. Conversely, given any function $f : \mathbb{R}^n \to [0, \infty)$ such that

$$\int_{\mathbb{R}^n} f(x_1, \ldots, x_n) dx_1 \cdots dx_n = 1,$$

it is possible to construct a random vector X with p.d.f. f. This is achieved by following exactly the same argument as for random variables.

If we are given the joint density of $X_1, \ldots, X_n$, it is possible to calculate the probability density functions of the individual random variables $X_1, \ldots, X_n$ using the following simple method[51]. If f is the joint p.d.f. of $X_1, \ldots, X_n$, then for each i, the p.d.f. of X_i is given by

$$f_i(x) = \int_{\mathbb{R}^{n-1}} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

Let us verify this for $i = 1$. Take any $B \subseteq \mathbb{R}$. Let $A = B \times \mathbb{R}^{n-1}$. Then

$$\begin{aligned} P(X_1 \in B) &= P(X \in A) \quad (\text{where } X = (X_1, \ldots, X_n)) \\ &= \int_A f(x_1, \ldots, x_n) dx_1 \cdots dx_n \\ &= \int_B \left(\int_{\mathbb{R}^{n-1}} f(x_1, \ldots, x_n) dx_2 \cdots dx_n \right) dx_1 \\ &= \int_B f_1(x_1) dx_1. \end{aligned}$$

Thus, f_1 must be the p.d.f. of X_1.

For example, let $X = (X_1, X_2)$ be a random vector that is uniformly distributed over the unit disk in $\mathbb{R}^2$, which means that for any subset A of the unit disk, $P(X \in A)$ is proportional to the area of A. The proportionality constant must be $1/\pi$, since the area of the unit disk is π. Therefore, the p.d.f. f of X must be equal to $1/\pi$ everywhere inside the unit disk, and 0 outside. Let us now calculate the p.d.f. f_1 of X_1. Note that X_1 takes values in the interval $[-1, 1]$. For any x in this interval,

$$f_1(x) = \int_{-\infty}^{\infty} f(x, x_2) dx_2.$$

Since f is zero outside the unit disk, the integrand in the above integral is zero if $x_2 > \sqrt{1 - x^2}$ or $x_2 < -\sqrt{1 - x^2}$. On the other hand, inside this range, it is equal to $1/\pi$. Thus[52],

$$f_1(x) = \frac{2}{\pi} \sqrt{1 - x^2}.$$

[50] Again, we need A to be a Borel subset of $\mathbb{R}^n$.

[51] This is the continuous analog of the formula (3) for marginal probability mass functions.

[52] Since f_1 must integrate to 1, this gives a simple proof of the fact that

$$\int_{-1}^{1} \sqrt{1 - x^2} dx = \frac{\pi}{2},$$

which is not entirely trivial to prove directly.

Independence

Let $X_1, \ldots, X_n$ be continuous random variables with joint p.d.f. f. Let f_i be the marginal p.d.f. of X_i, for $i = 1, \ldots, n$. We say that $X_1, \ldots, X_n$ are independent if

$$f(x_1, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n).$$

Let $X_1, \ldots, X_n$ be independent, with joint density f and marginal densities $f_1, \ldots, f_n$ as above. Take any subsets $B_1, \ldots, B_n$ of the real line. Let $A = B_1 \times \cdots \times B_n$. Then

$$\begin{aligned}
&P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X \in A) \ \text{ (where } X = (X_1, \ldots, X_n)) \\
&= \int_A f(x_1, \ldots, x_n) dx_1 \cdots dx_n \\
&= \int_{B_n} \int_{B_{n-1}} \cdots \int_{B_1} f_1(x_1) \cdots f_n(x_n) dx_1 \cdots dx_n \\
&= P(X_1 \in B_1) P(X_2 \in B_2) \cdots P(X_n \in B_n),
\end{aligned}$$

where the last line was obtained by integrating the variables one by one.

Just as for joint p.m.f.'s of discrete random variables, if the joint density of a random vector $(X_1, \ldots, X_n)$ has the form

$$f(x_1, \ldots, x_n) = h_1(x_1) \cdots h_n(x_n)$$

where $h_1, \ldots, h_n$ are probability density functions on the real line, then $X_1, \ldots, X_n$ are independent and h_i is the p.d.f. of X_i for each i. The proof is similar to that of Proposition 5.

For example, consider a random vector $X = (X_1, X_2)$ that is uniformly distributed on the square $[0, 1]^2$. Then its p.d.f. is 1 inside the square and 0 outside. So we can write the p.d.f. as

$$f(x_1, x_2) = f_1(x_1) f_2(x_2)$$

where both f_1 and f_2 are functions of one variable that are 1 inside the interval $[0, 1]$ and 0 outside. But this is the p.d.f. of a $Unif[0, 1]$ random variable. Thus, X_1 and X_2 are independent $Unif[0, 1]$ random variables.

On the other hand, consider the example of a random vector $X = (X_1, X_2)$ that is uniformly distributed in the unit disk. As noted in the previous section, the p.d.f. f of X is $1/\pi$ inside the unit disk and 0 outside. On the other hand, we computed that the p.d.f. of X_1 is

$$f_1(x) = \frac{2}{\pi} \sqrt{1 - x^2}$$

in the interval $[-1, 1]$ and 0 outside. By a similar argument, the p.d.f. f_2 of X_2 is also the same. Thus, $f \neq f_1 f_2$, and so X_1 and X_2 are not independent.

Conditional probability density function

Let X and Y be two random variables with joint probability density function f. Take any $x \in \mathbb{R}$ where $f(x) > 0$. The **conditional probability density function** of Y given $X = x$, which we denote by g_x, is defined as

$$g_x(y) = \frac{f(x,y)}{f_1(x)},$$

where f_1 is the marginal p.d.f. of X. The intuition is that if we take a small ε, then

$$\begin{aligned}
&P(Y \in [y-\varepsilon, y+\varepsilon] \mid X \in [x-\varepsilon, x+\varepsilon]) \\
&= \frac{P(Y \in [y-\varepsilon, y+\varepsilon], X \in [x-\varepsilon, x+\varepsilon])}{P(X \in [x-\varepsilon, x+\varepsilon])} \\
&\approx \frac{(2\varepsilon)^2 f(x,y)}{2\varepsilon f_1(x)} = 2\varepsilon g_x(y).
\end{aligned}$$

Just as for discrete random variables, the standard convention is to denote the joint density of (X, Y) by $f_{X,Y}$, the marginal densities of X and Y by f_X and f_Y, and the conditional density of Y given $X = x$ by $f_{Y|X=x}$. With this notation,

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

Consider the familiar example of a random vector (X, Y) distributed uniformly on the unit disk $D = \{(x,y) : x^2 + y^2 \le 1\}$. We have seen that

$$f_{X,Y}(x,y) = \begin{cases} 1/\pi & \text{if } (x,y) \in D, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & \text{if } -1 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Take any $x \in (-1, 1)$. The above formulas show that the conditional density of Y given $X = x$ is given by

$$f_{Y|X=x}(y) = \begin{cases} \frac{1}{2\sqrt{1-x^2}} & \text{if } |y| \le \sqrt{1-x^2}, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, given $X = x$, Y is uniformly distributed on the interval $[-\sqrt{1-x^2}, \sqrt{1-x^2}]$.

Given two random variables X and Y with joint density $f_{X,Y}$ and a set $A \subseteq \mathbb{R}^2$, we can evaluate the conditional probability of the event $(X, Y) \in A$ given $X = x$ using the formula

$$P((X,Y) \in A | X = x) = \int_{\{y:(x,y) \in A\}} f_{Y|X=x}(y) dy.$$

This is consistent with a version of the law of total probability for continuous variables:

$$\begin{aligned}
P((X,Y) \in A) &= \int_A f_{X,Y}(x,y)dydx \\
&= \int_A f_{Y|X=x}(x,y)f_X(x)dydx \\
&= \int_{-\infty}^{\infty} \int_{\{y:(x,y)\in A\}} f_{Y|X=x}(x,y)f_X(x)dydx \\
&= \int_{-\infty}^{\infty} P((X,Y) \in A|X=x)f_X(x)dx.
\end{aligned}$$

There are obvious generalizations of this formula to larger numbers of variables and vectors.

Multivariate change of variable formula

Let $X = (X_1, \ldots, X_n)$ be a continuous random vector with p.d.f. f, and let $Y = u(X)$, where $u : \mathbb{R}^n \to \mathbb{R}^n$ is a smooth[53] injective[54] function whose inverse is also smooth. Let $v = u^{-1}$ be the inverse of u. The multivariate change of variable formula is a formula for the p.d.f. of Y in terms of the p.d.f. of X and the Jacobian of the map v.

Let $v_1, \ldots, v_n$ denote the n components of the map v. Recall that the Jacobian matrix of v is the matrix-valued function

$$J(y) = \begin{pmatrix} \frac{\partial v_1}{\partial y_1} & \cdots & \frac{\partial v_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial v_n}{\partial y_1} & \cdots & \frac{\partial v_n}{\partial y_n} \end{pmatrix}.$$

The *Jacobian determinant*, or simply the Jacobian, of v is the determinant of the Jacobian matrix. The multivariate change of variable formula says that the p.d.f. of Y at a point $y \in \mathbb{R}^n$ is

$$g(y) = f(v(y))|\det J(y)|.$$

Note that this is a generalization of the univariate formula (15). We proved the univariate formula using the cumulative distribution function. The multivariate formula has a different proof (which is also applicable for the univariate case). It goes as follows. Take any $A \subseteq \mathbb{R}^n$. Let

$$B = v(A) = \{v(x) : x \in A\}.$$

Then

$$\begin{aligned}
P(Y \in A) &= P(v(Y) \in v(A)) = P(X \in B) \\
&= \int_B f(x)dx \ \ \text{(where } dx = dx_1 \cdots dx_n\text{)}.
\end{aligned}$$

[53] Infinitely differentiable.

[54] One-to-one.

Now let us apply the change of variable $y = u(x)$ to the above integral, which is the same as $x = v(y)$. The change of variable formula for multiple integrals[55] gives us

$$\int_B f(x)dx = \int_A f(v(y))|\det J(y)|dy.$$

Since this holds for any A, we conclude that the p.d.f. of Y must be $f(v(y))|\det J(y)|$.

Applications of the change of variable formula

Let us now work a number of important applications of the multivariate change of variable formula. Let X and Y be a pair of continuous random variables defined on the same sample space, with joint p.d.f. f. The change of variable formula allows us to calculate the probability density functions of $X + Y$, XY and X/Y. There is a general method of solving such problems, which will become clear soon.

First, let $u(x, y) = (x + y, y)$. Let $(Z, W) = u(X, Y)$. That is, $Z = X + Y$ and $W = Y$. The plan is to first calculate the p.d.f. of (Z, W) using the change of variable formula, and then compute the marginal p.d.f. of Z.

If $(z, w) = (x + y, y)$, then $x = z - w$ and $y = w$. Thus, $v(z, w) = (z - w, w)$ is the inverse of u. The Jacobian matrix of v is

$$J(z, w) = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

Thus, the p.d.f. of (Z, W) is

$$g(z, w) = f(v(z, w))|\det J(z, w)| = f(z - w, w).$$

Consequently, the p.d.f. of $Z = X + Y$ is

$$h(z) = \int_{-\infty}^{\infty} g(z, w)dw = \int_{-\infty}^{\infty} f(z - w, w)dw. \qquad (17)$$

Note that if X and Y were discrete random variables, then the formula for the p.m.f. of Z would be[56]

$$P(Z = z) = \sum_w P(X = z - w, Y = w),$$

which is a discrete version of (17).

Let us now turn our attention to XY. Let $Z = XY$. If X and Y were discrete, the p.m.f. of Z would be given by

$$P(Z = z) = \sum_w P(X = z/w, Y = w).$$

[55] If you do not want to apply that or do not remember what it is, here is an alternative sketch of the proof. Let A be a small ball containing y. Then $P(Y \in A) \approx \mathrm{vol}(A)g(y)$, where $\mathrm{vol}(A)$ denotes the volume of A. Let $B = v(A)$, so that $P(Y \in A) = P(X \in B)$. But in the small ball A, v behaves like the linear map $w(z) = v(y) + J(y)(z - y)$, by Taylor approximation. It is a standard fact from linear algebra (see below) that multiplication by a matrix M multiplies the volume of a region by $|\det M|$. Thus, $\mathrm{vol}(B) \approx |\det J(y)|\mathrm{vol}(A)$. To complete the proof, notice that $P(X \in B) \approx \mathrm{vol}(B)f(x)$, since B is a small region containing x.

For the linear algebra fact, let UDV^T be the singular value decomposition of M, where U and V are orthogonal, and D is diagonal. The actions of V^T and U preserve volumes, and it is easy to see that multiplication by D changes volume by $\det D$. But $\det D = |\det M|$.

[56] By the law of total probability.

This may lead us to guess that in the continuous case, the p.d.f. of Z at a point z is given by the formula

$$\int_{-\infty}^{\infty} f(z/w, w)dw.$$

However, this is not the case! The p.d.f. of Z is actually

$$h(z) = \int_{-\infty}^{\infty} \frac{1}{|w|} f(z/w, w)dw. \tag{18}$$

This shows that one should be careful about guessing results about continuous random variables from analogous results for discrete random variables. Let us now prove that h is indeed the p.d.f. of Z. The procedure is as before. Let $u(x,y) = (xy, y)$ and $(Z, W) = u(X,Y)$. If $u(x,y) = (z,w)$, then $x = z/w$ and $y = w$. Thus, $v(z,w) = (z/w, w)$ is the inverse of u. The Jacobian matrix of v is

$$J(z,w) = \begin{pmatrix} 1/w & -z/w^2 \\ 0 & 1 \end{pmatrix}.$$

The main difference now is that $\det J$ is not constant. By the change of variable formula, we get that the p.d.f. of (Z, W) is

$$g(z,w) = f(v(z,w))|\det J(z,w)| = \frac{1}{|w|} f(z/w, w).$$

This shows that the p.d.f. of $Z = XY$ is given by the function h displayed in (18).

As a final example, let us compute the p.d.f. of X/Y. We proceed exactly as in the previous two examples. Let $u(x,y) = (x/y, y)$ and $(Z, W) = u(X,Y)$. The inverse function is $v(z,w) = (zw, w)$, and its Jacobian matrix is

$$J(z,w) = \begin{pmatrix} w & z \\ 0 & 1 \end{pmatrix}.$$

Thus, the p.d.f. of (Z, W) is $g(z,w) = |w|f(zw, w)$, and so the p.d.f. of $Z = X/Y$ is

$$h(z) = \int_{-\infty}^{\infty} |w| f(zw, w)dw. \tag{19}$$

Now let X and Y be two *independent* random variables with p.d.f.'s f and g. Let $Z = X + Y$. Then by the formula (17), the p.d.f. of Z is

$$h(z) = \int_{-\infty}^{\infty} f(z-w)g(w)dw.$$

The p.d.f. h is called the **convolution** of f and g, and often denoted by $f \star g$.

Example: Sum of two independent centered normals

Let $X \sim N(0, a^2)$ and $Y \sim N(0, b^2)$ be independent centered normal random variables. Then by (17), the p.d.f. of $Z = X + Y$ is

$$\begin{aligned} h(z) &= \frac{1}{2\pi ab} \int_{-\infty}^{\infty} e^{-(z-w)^2/2a^2} e^{-w^2/2b^2} dw \\ &= \frac{1}{2\pi ab} \int_{-\infty}^{\infty} e^{-(b^2z^2 - 2b^2zw + (b^2+a^2)w^2)/2a^2b^2} dw \\ &= \frac{e^{-z^2/2a^2}}{2\pi ab} \int_{-\infty}^{\infty} e^{-((b^2+a^2)w^2 - 2b^2zw)/2a^2b^2} dw. \end{aligned}$$

Now note that[57]

$$(b^2 + a^2)w^2 - 2b^2zw = (b^2 + a^2)\left(w - \frac{b^2z}{b^2 + a^2}\right)^2 - \frac{b^4z^2}{b^2 + a^2}.$$

Plugging this into the previous expression, we get

$$h(z) = \frac{e^{-z^2/2(a^2+b^2)}}{2\pi ab} \int_{-\infty}^{\infty} e^{-(b^2+a^2)(w - b^2z/(b^2+a^2))^2/2a^2b^2} dw.$$

Substituting

$$x = \frac{\sqrt{b^2 + a^2}}{ab}\left(w - \frac{b^2z}{b^2 + a^2}\right)$$

in the integral, we get

$$h(z) = \frac{e^{-z^2/2(a^2+b^2)}}{2\pi ab} \frac{ab}{\sqrt{b^2 + a^2}} \int_{-\infty}^{\infty} e^{-x^2/2} dx.$$

But we know that the integral on the right equals $\sqrt{2\pi}$. Thus,

$$h(z) = \frac{1}{\sqrt{2\pi(a^2 + b^2)}} e^{-z^2/2(a^2+b^2)}.$$

But this is the p.d.f. of $N(0, a^2 + b^2)$. Thus, $Z \sim N(0, a^2 + b^2)$.

The above result can be extended by induction to arbitrary linear combinations of independent normal random variables.

Proposition 11. *Let $X_1, \ldots, X_n$ be independent normal random variables, with $X_i \sim N(\mu_i, \sigma_i^2)$. Take any real numbers $a_0, \ldots, a_n$ and let $Y = a_0 + a_1X_1 + \cdots + a_nX_n$. Then*

$$Y \sim N(a_0 + a_1\mu_1 + \cdots + a_n\mu_n, a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2).$$

Proof. We have already seen the case $n = 1$ in equation (16). Let us now prove it for $n = 2$. By equation (16), $a_i(X_i - \mu_i) \sim N(0, a_i^2\sigma_i^2)$ for each i. Thus, by the result proved above,

$$a_1(X_1 - \mu_1) + a_2(X_2 - \mu_2) \sim N(0, a_1^2\sigma_1^2 + a_2^2\sigma_2^2).$$

[57] This is 'completing the square'.

Therefore again by equation (16),

$$\begin{aligned} a_0 + a_1X_1 + a_2X_2 &= a_0 + a_1\mu_1 + a_2\mu_2 + a_1(X_1 - \mu_1) + a_2(X_2 - \mu_2) \\ &\sim N(a_0 + a_1\mu_1 + a_2\mu_2, a_1^2\sigma^2 + a_2^2\sigma_2^2). \end{aligned}$$

Now suppose that this claim holds for $n-1$ variables. Under this assumption, we will now prove it for n variables[58]. Let $Z = a_0 + a_1X_1 + \cdots + a_{n-1}X_{n-1}$ and $W = a_nX_n$, so that $Y = Z + W$. Since the result holds for $n-1$ variables (by assumption),

$$Z \sim N(a_0 + a_1\mu_1 + \cdots + a_{n-1}\mu_{n-1}, a_1^2\sigma_1^2 + \cdots + a_{n-1}^2\sigma_{n-1}^2).$$

Also, by equation (16), $W \sim N(a_n\mu_n, a_n^2\sigma_n^2)$. Therefore, applying the case $n = 2$, we get the desired conclusion for Y. □

Example: Ratio of two independent centered normals

Let $X \sim N(0, a^2)$ and $Y \sim N(0, b^2)$ be independent centered normal random variables. Let $Z = X/Y$. By equation (19), the p.d.f. of Z is

$$h(z) = \frac{1}{2\pi ab}\int_{-\infty}^{\infty} |w| e^{-(zw)^2/2a^2} e^{-w^2/2b^2} dw.$$

Since the integrand is an even function of w, the integral from $-\infty$ to ∞ equals two times the integral from 0 to ∞. Thus,

$$\begin{aligned} h(z) &= \frac{1}{\pi ab}\int_0^\infty w e^{-(z^2b^2+a^2)w^2/2a^2b^2} dw \\ &= \frac{ab}{\pi(z^2b^2+a^2)}\int_0^\infty \frac{d}{dw}(e^{-(z^2b^2+a^2)w^2/2a^2b^2}) dw \\ &= \frac{ab}{\pi(z^2b^2+a^2)} = \frac{(a/b)}{\pi(z^2 + (a/b)^2)}. \end{aligned}$$

A random variable with p.d.f.

$$\frac{\gamma}{\pi(x^2+\gamma^2)}$$

is called a *Cauchy*(γ) random variable. Thus, $X/Y \sim$ *Cauchy*(a/b). The *Cauchy*(1) distribution is called the **standard Cauchy distribution**. The above calculation shows that the ratio of two standard normal random variables is a standard Cauchy random variable.

Example: Gamma random variables

Let $X_1, \ldots, X_n$ be i.i.d. $Exp(\lambda)$ random variables. We will now show that the p.d.f. of $X_1 + \cdots + X_n$ is the function that is zero on the negative axis and equals

$$\frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}$$

[58] In case you have not seen it before, this is a general proof technique, known as **proof by induction**.

at $x \geq 0$. A random variable with this p.d.f. is called a $Gamma(n, \lambda)$ random variable[59].

The above claim is obviously true for $n = 1$. Let us assume that it holds for a sum of $n-1$ variables. Let f_n denote the p.d.f. of $X_1 + \cdots + X_n$. Then obviously $f_n(x) = 0$ if $x < 0$, since the random variables are nonnegative. For $x \geq 0$, we have[60]

$$f_n(x) = \int_{-\infty}^{\infty} f_1(x-y) f_{n-1}(y) dy.$$

Now note that $f_1(x-y) = 0$ if $y > x$, and $f_{n-1}(y) = 0$ if $y < 0$. Thus,

$$\begin{aligned}
f_n(x) &= \int_0^x f_1(x-y) f_{n-1}(y) dy \\
&= \int_0^x \lambda e^{-\lambda(x-y)} \frac{\lambda^{n-1} y^{n-2} e^{-\lambda y}}{(n-2)!} dy \\
&= \frac{\lambda^n e^{-\lambda x}}{(n-2)!} \int_0^x y^{n-2} dy \\
&= \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}.
\end{aligned}$$

This completes the induction step and proves the claim.

[59] Note that it is not obvious that the above function integrates to 1. One way to show that is by using the fact that it is the p.d.f. of a random variable, as we will show soon. Another way is to use repeated integrations by parts.

[60] Here we are implicitly using the fact that $X_1 + \cdots + X_{n-1}$ and X_n are independent. To see this, let $Y = X_1 + \cdots + X_n$. Take any $A, B \subseteq \mathbb{R}$, and write $P(Y \in A, X_n \in B)$ as $P((X_1, \ldots, X_n) \in D)$ for some suitable set $D \subseteq \mathbb{R}^n$, and then evaluate the latter as an integral. The integral will factor into a product of two integrals, one of which will equal $P(Y \in A)$ and the other will equal $P(X \in B)$.

More about continuous random variables

Expected value

The expected value (or expectation, or mean) of a continuous random variable X with p.d.f. f is defined as[61]

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

provided the integral is absolutely convergent[62]. For example, if $X \sim Unif[a,b]$, then

$$E(X) = \frac{1}{b-a}\int_a^b xdx = \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{a+b}{2}.$$

This makes sense, because a random variable which is uniformly distributed on the interval $[a,b]$ should have the midpoint of the interval as its long-term average value.

If $X \sim Exp(\lambda)$, then using integration by parts,

$$\begin{aligned} E(X) &= \int_0^{\infty} \lambda x e^{-\lambda x} dx \\ &= \left[-xe^{-\lambda x}\right]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = \frac{1}{\lambda}. \end{aligned}$$

The parameter λ is sometimes called the *rate*. The above calculation shows that the reciprocal of the rate is the mean of an exponential random variable.

As our third example, let $X \sim N(\mu, \sigma^2)$. Then

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} \frac{xe^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dx \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dx + \int_{-\infty}^{\infty} \frac{\mu e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dx \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dx + \mu \\ &= \int_{-\infty}^{\infty} \frac{ye^{-y^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dy + \mu. \end{aligned}$$

[61] Roughly, the idea behind this definition is as follows. Take some small ε, and *discretize* X by defining a new random variable X_ε which takes value $k\varepsilon$ if $X \in ((k-1)\varepsilon, k\varepsilon]$ for some $k \in \mathbb{Z}$. Then

$$\begin{aligned} E(X_\varepsilon) &= \sum_{k=-\infty}^{\infty} k\varepsilon P(X_\varepsilon = k\varepsilon) \\ &= \sum_{k=-\infty}^{\infty} k\varepsilon \int_{(k-1)\varepsilon}^{k\varepsilon} f(x)dx. \end{aligned}$$

As $\varepsilon \to 0$, $X_\varepsilon \to X$ and the last line tends to $\int_{-\infty}^{\infty} xf(x)dx$.

[62] That is,

$$\int_{-\infty}^{\infty} |x| f(x)dx < \infty.$$

In other words, we need $E|X| < \infty$.

But the integrand in the last line is an odd function of y, and the integral converges absolutely, So the value of the integral must be zero. Thus, $E(X) = \mu$. For this reason, the parameter μ is called the mean of the $N(\mu, \sigma^2)$ distribution. We will see later that σ^2 is the variance.

As our final example, let us consider $X \sim Cauchy(\gamma)$. Then

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

Since the integrand is an odd function of x, you may think that the expected value is zero. However, note that the integral is not absolutely convergent. For this reason, $E(X)$ is considered to be *undefined* for a Cauchy random variable.

Properties of expectation

Let $X_1, \ldots, X_n$ be continuous random variables with joint probability density function f. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a function and let $Y = g(X_1, \ldots, X_n)$. Then, just as for discrete random variables (Proposition 6), we have the following simple method for calculating $E(Y)$.

Proposition 12. *Let Y be as above. Then*

$$E(Y) = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

Proof. Let h be the p.d.f. of Y. Take any $\varepsilon > 0$. Let

$$A_\varepsilon = \sum_{k\in\mathbb{Z}} k\varepsilon P((k-1)\varepsilon < Y \le k\varepsilon).$$

Then

$$\begin{aligned} E(Y) - A_\varepsilon &= \int_{-\infty}^{\infty} xh(x)dx - \sum_{k\in\mathbb{Z}} k\varepsilon \int_{(k-1)\varepsilon}^{k\varepsilon} h(x)dx \\ &= \sum_{k\in\mathbb{Z}} \int_{(k-1)\varepsilon}^{k\varepsilon} (x - k\varepsilon) h(x) dx. \end{aligned}$$

We now compute an upper bound on the absolute value of the above difference, applying the triangle inequality[63] and the fact that the absolute value of an integral is less than or equal to the integral of the absolute value[64]:

$$\begin{aligned} |E(Y) - A_\varepsilon| &\le \sum_{k\in\mathbb{Z}} \int_{(k-1)\varepsilon}^{k\varepsilon} |x - k\varepsilon| h(x) dx \\ &\le \varepsilon \sum_{k\in\mathbb{Z}} \int_{(k-1)\varepsilon}^{k\varepsilon} h(x)dx = \varepsilon \int_{-\infty}^{\infty} h(x)dx = \varepsilon. \qquad (20) \end{aligned}$$

[63] That is, the inequality

$$|a_1 + a_2 + \cdots| \le |a_1| + |a_2| + \cdots.$$

[64] Proved by applying the triangle inequality to Riemann sums in the definition of integral.

For each $k \in \mathbb{Z}$, let

$$S_k = \{(x_1, \ldots, x_n) : g(x_1, \ldots, x_n) \in ((k-1)\varepsilon, k\varepsilon]\}.$$

Then $Y \in ((k-1)\varepsilon, k\varepsilon]$ if and only if $(X_1, \ldots, X_n) \in S_k$. Thus,

$$A_\varepsilon = \sum_{k \in \mathbb{Z}} k\varepsilon \int_{S_k} f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

So if we put

$$I = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n,$$

then

$$I - A_\varepsilon = \sum_{k \in \mathbb{Z}} \int_{S_k} (g(x_1, \ldots, x_n) - k\varepsilon) f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

Proceeding as before using triangle inequality, we get

$$|I - A_\varepsilon| \leq \varepsilon. \tag{21}$$

Combining (20) and (21), we get $|E(Y) - I| \leq 2\varepsilon$. But ε is arbitrary. So taking $\varepsilon \to 0$, we get $E(Y) = I$. □

An immediate consequence of the above proposition is that expectation is linear for continuous random variables, just as it is in the discrete case[65]. Similarly, another consequence is that the expectation of a product of independent continuous random variables is the product of the expectations.

As an application, let us compute the mean of a $Gamma(n, \lambda)$ random variable. Recall that if $X_1, \ldots, X_n$ are i.i.d. $Exp(\lambda)$ random variables, then $X_1 + \cdots + X_n \sim Gamma(n, \lambda)$. Since we calculated that the expected value of an $Exp(\lambda)$ random variable is $1/\lambda$, it follows by linearity of expectation that the expected value of a $Gamma(n, \lambda)$ random variable is n/λ.

Variance

The variance of a continuous random variable is defined just like the variance of a discrete random variable:

$$Var(X) = E(X^2) - (E(X))^2.$$

Similarly, covariance of two continuous random variables X and Y is defined as

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

All the properties of variance and covariance discussed previously continue to hold for continuous random variables. In particular:

[65] Prove this, if in doubt.

- $Var(X) = E[(X - E(X))^2]$.
- $Var(bX + c) = b^2 Var(X)$.
- Covariance is bilinear.
- If X and Y are independent, then $Cov(X, Y) = 0$.
- $Var(\sum_i X_i) = \sum_i \sum_j Cov(X_i, X_j)$.
- If $X_1, \ldots, X_n$ are independent, $Var(\sum_i X_i) = \sum_i Var(X_i)$.

Let us now calculate the variances of the continuous random variables introduced earlier. We will be using the expected values computed earlier, so please revisit those if you do not remember. First, let $X \sim Unif[a, b]$. Then

$$E(X^2) = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^3 - a^3}{3(b-a)} = \frac{1}{3}(b^2 + ab + a^2).$$

Thus,

$$\begin{aligned} Var(X) &= \frac{1}{3}(b^2 + ab + a^2) - \frac{b^2 + 2ab + a^2}{4} \\ &= \frac{(b-a)^2}{12}. \end{aligned}$$

Next, let $X \sim Exp(\lambda)$. Then

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} \lambda x^2 e^{-\lambda x} dx \\ &= \left[-x^2 e^{-\lambda x}\right]_0^{\infty} + \int_0^{\infty} 2x e^{-\lambda x} dx \\ &= \frac{2}{\lambda} E(X) = \frac{2}{\lambda^2}. \end{aligned}$$

Thus,

$$Var(X) = \frac{2}{\lambda^2} - (E(X))^2 = \frac{1}{\lambda^2}.$$

Since the variance of a sum of independent random variables is the sum of the variances, this also shows that the variance of a $Gamma(n, \lambda)$ random variable is n/λ^2.

Finally, let $X \sim N(\mu, \sigma^2)$. Then

$$\begin{aligned} Var(X) &= E[(X - E(X))^2] = E[(X - \mu)^2] \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dx. \end{aligned}$$

Putting $y = (x - \mu)/\sigma$, we get

$$Var(X) = \sigma^2 \int_{-\infty}^{\infty} \frac{y^2 e^{-y^2/2}}{\sqrt{2\pi}} dy.$$

It is an easy exercise to show using integration by parts that the integral equals 1. Thus, $Var(X) = \sigma^2$.

Inequalities and laws of large numbers

The results from the chapter of laws of large numbers are all valid for continuous random variables. The proofs are exactly the same, with probability mass functions replaced by probability density functions at the appropriate places. In particular:

- If X is a nonnegative continuous random variable, then $E(X) \geq 0$.
- If X and Y are continuous random variables such that $X \geq Y$ always, then $E(X) \geq E(Y)$. (This result also holds if one variable is continuous and the other discrete.)
- Similarly, when $p \geq 1$, the inequality $|E(X)| \leq (E|X|^p)^{1/p}$ holds for continuous random variables.
- Markov's and Chebyshev's inequalities hold for continuous random variables.
- The weak law of large numbers holds for continuous random variables, as well as the result that $E(X_n) \to c$ and $Var(X_n) \to 0$ implies $X_n \to c$ in probability.

The tail integral formula for expectation

Let X be a nonnegative continuous random variable with finite expected value. The **tail integral formula** for $E(X)$ says that

$$E(X) = \int_0^\infty P(X \geq t)dt.$$

To prove this[66], let F be the c.d.f. and $f = F'$ be the p.d.f. of X. Let us assume that $F(t)$ approaches 1 so fast as $t \to \infty$ that $(1 - F(t))t \to 0$. Since X is a continuous random variable, $P(X \geq t) = P(X > t) = 1 - F(t)$. Therefore using integration by parts, we get

$$\begin{aligned}\int_0^\infty P(X \geq t)dt &= \int_0^\infty (1 - F(t))dt \\ &= \big[(1 - F(t))t\big]_0^\infty + \int_0^\infty tf(f)dt = E(X).\end{aligned}$$

There is a generalization of this identity, which says that if $g : [0, \infty) \to \mathbb{R}$ is a differentiable function, then under mild conditions on g,

$$E(g(X)) = g(0) + \int_0^\infty g'(t)P(X \geq t)dt.$$

[66] There is a measure-theoretic proof of this identity that requires no assumptions on X other than that $X \geq 0$. In particular, it is not required that X is continuous. The proof goes as follows. Observe that

$$X = \int_0^X dt = \int_0^\infty 1_{\{X \geq t\}}dt.$$

Then there is a measure-theoretic result, known as the **monotone convergence theorem**, which says that we can switch the order of expectation and integration below, to get

$$\begin{aligned}E(X) &= E\Big(\int_0^\infty 1_{\{X\geq t\}}dt\Big) \\ &= \int_0^\infty E(\{X \geq t\})dt \\ &= \int_0^\infty P(X \geq t)dt.\end{aligned}$$

Mean vector and covariance matrix

Let $X = (X_1, \ldots, X_n)$ be an n-dimensional random vector (discrete or continuous). The **mean vector** of X, denoted by $E(X)$, is the n-dimensional vector whose i^{th} component is $E(X_i)$. If A is an $m \times n$ matrix and $Y = AX$ (treating X as a column vector), then Y is an m-dimensional random vector, and linearity of expectation implies that $E(Y) = AE(X)$.

The **covariance matrix** of X, sometimes denoted by $Cov(X)$, is the $n \times n$ matrix Σ with $(i, j)^{\text{th}}$ entry is

$$\sigma_{ij} = Cov(X_i, X_j).$$

Note that Σ is a symmetric matrix, since $Cov(X_i, X_j) = Cov(X_j, X_i)$. We claim that Σ is a positive semidefinite (p.s.d.) matrix, meaning that $u^T \Sigma u \geq 0$ for every $u \in \mathbb{R}^n$. (Here and later, we treat vectors as column vectors, and u^T denotes the transpose of u.)

To see this, let $u_1, \ldots, u_n$ denote the components of a vector u, and note that

$$\begin{aligned} u^T \Sigma u &= \sum_{i,j=1}^{n} u_i u_j \sigma_{ij} \\ &= \sum_{i,j=1}^{n} u_i u_j Cov(X_i, X_j) = Var\left(\sum_{i=1}^{n} u_i X_i\right). \end{aligned}$$

Since the variance of any random variable is nonnegative, this proves the claim.

Now take any $m \times n$ matrix A, and let $Y = AX$. Then Y is an m-dimensional random vector. We claim that the covariance matrix of Y is $A \Sigma A^T$. To see this, let a_{ij} denote the $(i, j)^{\text{th}}$ entry of A, and note that for any i and j,

$$\begin{aligned} Cov(Y_i, Y_j) &= Cov\left(\sum_{k=1}^{n} a_{ik} X_k, \sum_{l=1}^{n} a_{jl} X_l\right) \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n} a_{ik} a_{jl} Cov(X_k, X_l) \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n} a_{ik} a_{jl} \sigma_{kl}. \end{aligned}$$

The double sum in the last line is the $(i, j)^{\text{th}}$ entry of $A \Sigma A^T$.

Normal random vectors

Let $X_1, \ldots, X_n$ be i.i.d. standard normal random variables. The random vector $X = (X_1, \ldots, X_n)$ is called a standard normal (or

Gaussian) random vector. Note that the p.d.f. of X at a point $x = (x_1, \ldots, x_n)$ is

$$f(x) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\|x\|^2},$$

where $\|x\|$ is the Euclidean norm of x. Now take any nonsingular $n \times n$ matrix A, and let $Y = AX$. Since the covariance matrix of X is simply the identity matrix I, the covariance matrix of Y is $\Sigma = AA^T$. Let us now derive the p.d.f. of Y. Note that $X = A^{-1}Y$. This is a linear transformation, whose Jacobian matrix is A^{-1}. Therefore by the change of variable formula, the p.d.f. of Y at a point y is

$$g(y) = f(A^{-1}y)|\det A^{-1}|.$$

Now note that

$$\begin{aligned} f(A^{-1}y) &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(A^{-1}y)^T(A^{-1}y)} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}y^T(A^{-1})^T A^{-1}y} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}y^T\Sigma^{-1}y}. \end{aligned}$$

Also, note that

$$\det \Sigma = \det(AA^T) = \det(A)\det(A^T) = (\det A)^2,$$

and so

$$|\det A^{-1}| = |(\det A)^{-1}| = (\det \Sigma)^{-1/2}.$$

Thus,

$$g(y) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} e^{-\frac{1}{2}y^T\Sigma^{-1}y}.$$

Next, take any $\mu \in \mathbb{R}^n$ and let $Z = \mu + Y = \mu + AX$. Then again by the change of variable formula, it is easy to show that the p.d.f. of Z at a point z is given by

$$h(z) = g(z - \mu) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} e^{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)}.$$

The random variable Z has covariance matrix Σ and mean vector μ, and its p.d.f. is given by the above formula, which involves only μ and Σ as parameters. We say that Z is a **normal random vector** with mean μ and covariance matrix Σ, and write $Z \sim N(\mu, \Sigma)$. If $Z_1, \ldots, Z_n$ are the components of Z, we say that $Z_1, \ldots, Z_n$ are **jointly normal**. The distribution of Z is called the **multivariate normal distribution** with mean μ and covariance matrix Σ.

Recall that any positive definite matrix Σ can always be written as AA^T for some nonsingular matrix A. From this, it follows that given any positive definite matrix Σ and any vector μ (of the same dimension), the $N(\mu, \Sigma)$ distribution is well-defined since it arises as the distribution of $\mu + AX$, where X is a standard normal random vector.

Since a normal random vector is by definition a linear transformation of a standard normal random vector, it follows that any linear transformation of a normal random vector is again a normal random vector (with appropriate mean and covariance matrix). To be precise, if $Z \sim N(\mu, \Sigma)$ and $Z' = \nu + BZ$ for some $\nu \in \mathbb{R}^n$ and $n \times n$ nonsingular matrix B, then $Z' \sim N(\nu + B\mu, B\Sigma B^T)$.

What if we multiply a standard normal random vector by a *rectangular*, instead of square, matrix? We claim that the result is still a normal random vector. Let X be an n-dimensional standard normal random vector, and let A be an $m \times n$ matrix of full rank, where $m < n$. Let H be the subspace of $\mathbb{R}^n$ spanned by the rows of A, so that $\dim H = m$. Let $u_1, \ldots, u_{n-m}$ be an orthonormal basis of the orthogonal complement of H. Produce an $n \times n$ matrix B by adding $u_1, \ldots, u_{n-m}$ as row vectors below the rows of A. Let $Y = AX$ and $Z = BX$, so that Y consists of the first m elements of Z.

Note that $Z \sim N(0, BB^T)$. But by construction of B,

$$BB^T = \begin{pmatrix} \Sigma & 0 \\ 0 & I \end{pmatrix},$$

where $\Sigma = AA^T$, and I is the identity matrix of order $n - m$. The determinant of the above matrix is equal to $\det \Sigma$, and its inverse is

$$\begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & I \end{pmatrix}.$$

Thus, the p.d.f. of Z at a point $z \in \mathbb{R}^n$ is

$$f(z) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} e^{-\frac{1}{2}y^T\Sigma^{-1}y - \frac{1}{2}w^Tw},$$

where $y^T = (z_1, \ldots, z_m)$ and $w^T = (z_{m+1}, \ldots, z_n)$. To obtain the p.d.f. of Y, we have to integrate out w from the above density. But this is easily accomplished, since the variables y and w are separated in the exponent. Integrating out, we get that $Y \sim N(0, \Sigma)$.

An important consequence of the above result is that if a collection of random variables is jointly normal, any subcollection is also jointly normal.

Another very important property of jointly normal random variables is that independence can be easily verified by checking whether covariances are zero.

Proposition 13. *Let $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ be jointly normal random variables. Suppose that $Cov(X_i, Y_j) = 0$ for each i and j. Then the random vectors $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$ are independent*[67].

[67] In the sense that the joint density of $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is the product of the densities of $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$.

Proof. Let Σ be the covariance matrix of the random vector $Z = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$. By the given condition, Σ has the form

$$\begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix},$$

where Σ_1 is an $m \times m$ matrix and Σ_2 is an $n \times n$ matrix. Then note that

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_1^{-1} & 0 \\ 0 & \Sigma_2^{-1} \end{pmatrix},$$

and $\det \Sigma = \det \Sigma_1 \det \Sigma_2$. Thus, if μ_1 and μ_2 are the mean vectors of $(X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_n)$, then the p.d.f. of Z at a point $z = (x, y)$ (where $z \in \mathbb{R}^{m+n}$, $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$) is

$$\begin{aligned} &\frac{e^{-\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu)}}{(2\pi)^{(m+n)/2} (\det \Sigma)^{1/2}} \\ &= \frac{e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(y-\mu_2)^T \Sigma_2^{-1}(y-\mu_2)}}{(2\pi)^{(m+n)/2} (\det \Sigma_1 \det \Sigma_2)^{1/2}}. \end{aligned}$$

The above expression is the product of the probability density functions of $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ at x and y. This shows that X and Y are independent random vectors, with $X \sim N(\mu_1, \Sigma_1)$ and $Y \sim N(\mu_2, \Sigma_2)$. □

The central limit theorem

Convergence in distribution

Let $X_1, X_2, \ldots$ be a sequence of random variables and X be another random variable. Let F_n be the c.d.f. of X_n and F be the c.d.f. of X. We say that X_n **converges in distribution** to X as $n \to \infty$ if for every x where F is continuous[68],

$$\lim_{n\to\infty} F_n(x) = F(x). \tag{22}$$

Sometimes we also write $X_n \to F$ in distribution, or $F_n \to F$ in distribution. Note that when F is continuous (e.g. when X is a continuous random variable), we need (22) to hold for all x.

When F is continuous, a consequence of convergence in distribution is that for every $-\infty < a \le b < \infty$,

$$\lim_{n\to\infty} P(a \le X_n \le b) = P(a \le X \le b).$$

To see this, first note that

$$\begin{aligned} P(a < X_n \le b) &= P(X_n \le b) - P(X_n \le a) \\ &\to P(X \le b) - P(X \le a) = P(a < X \le b). \end{aligned}$$

On the other hand, for any $\varepsilon > 0$,

$$\begin{aligned} \limsup_{n\to\infty} P(X_n = a) &\le \limsup_{n\to\infty} P(a - \varepsilon < X_n \le a) \\ &= P(a - \varepsilon < X \le a) = F(a) - F(a - \varepsilon). \end{aligned}$$

The left side does not depend on ε. So we can take $\varepsilon \to 0$ on the right, and use the continuity of F to assert that $P(X_n = a) \to 0$. Thus,

$$\begin{aligned} P(a \le X_n \le b) &= P(X_n = a) + P(a < X_n \le b) \\ &\to P(a < X \le b) = P(a \le X \le b), \end{aligned}$$

where the last identity follows, again, by the continuity of F.

We have already seen one example of convergence in distribution. In Proposition 4, we showed that if $X_n \sim Bin(n, \lambda/n)$ and $X \sim Poi(\lambda)$, then for any integer $k \ge 0$,

$$\lim_{n\to\infty} P(X_n = k) = P(X = k).$$

[68] The condition that x has to be a continuity point of F is needed to ensure that we do not miss out 'obvious' examples. For instance, suppose that X_n is 0 with probability $1/2$ and $1 + 1/n$ with probability $1/2$. Then, as $n \to \infty$, it is clear that the distribution of X_n should converge to $Ber(1/2)$. However, $F_n(1) = 1/2$ for every n, whereas $F(1) = 1$ (where F is the c.d.f. of $Ber(1/2)$). With the given definition of convergence in distribution, we can avoid this problem, because 1 not a continuity point of F. Indeed, in this example, $F_n(x)$ does converge to $F(x)$ for every continuity point of F.

Since X_n and X are both nonnegative integer-valued random variables, this shows that for any real number $x \geq 0$,

$$\begin{aligned}\lim_{n\to\infty} P(X_n \leq x) &= \lim_{n\to\infty} \sum_{0\leq k\leq x} P(X_n = k) \\ &= \sum_{0\leq k\leq x} \lim_{n\to\infty} P(X_n = k) \\ &= \sum_{0\leq k\leq x} P(X = k) = P(X \leq x).\end{aligned}$$

Thus, $X_n \to X$ in distribution.

Statement of the central limit theorem

The goal of this chapter is to prove the following important result[69].

Theorem 5 (Central limit theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. For each n, let $S_n = X_1 + \cdots + X_n$. Then, as $n \to \infty$, then random variable*

$$\frac{S_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to a standard normal random variable. In particular, for any $-\infty < a \leq b < \infty$,

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

The theorem says that when n is large, $(S_n - n\mu)/\sqrt{n}\sigma$ behaves like a standard normal random variable, irrespective of the original distribution of the X_i's. Since a linear transformation of a normal random variable is again normal, this is the same as saying that S_n behaves approximately like a normal random variable with mean $n\mu$ and variance $n\sigma^2$. Indeed, a different way to write the conclusion of the theorem is

$$\lim_{n\to\infty} P(n\mu + a\sqrt{n}\sigma \leq S_n \leq n\mu + b\sqrt{n}\sigma) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx,$$

and the right side equals

$$P(n\mu + a\sqrt{n}\sigma \leq Z_n \leq n\mu + b\sqrt{n}\sigma)$$

for any n, where $Z_n \sim N(n\mu, n\sigma^2)$.

For a quick application, consider a fair coin tossed one million times. Let X be the number of heads. Note that X is the sum of a million i.i.d. random variables with mean $1/2$ and variance $1/4$. So in this case $n\mu = 500000$ and $\sqrt{n}\sigma = 500$. Therefore by the

[69] The proof will be fully rigorous to the extent possible without using measure theory. To avoid certain complications, we will work under the additional assumption that $E|X_i - \mu|^3 < \infty$.

central limit theorem, $(X - 500000)/500$ behaves approximately like a standard normal random variable. Numerical evaluation shows that

$$\int_{-2.576}^{2.576} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \approx 0.99.$$

Now $n\mu - 2.576\sqrt{n}\sigma = 498712$ and $n\mu + 2.576\sqrt{n}\sigma = 501288$. Therefore

$$P(498712 \le X \le 501288) \approx 0.99.$$

Recall that we had previously shown using Chebyshev's inequality that

$$P(495000 \le X \le 505000) \ge 0.99.$$

Therefore the central limit theorem not only gives a much narrower interval, but it also ensures that the chance of X belonging to the interval is actually close[70] to 0.99 instead of just being lower bounded by 0.99.

More generally, if $X_n \sim Bin(n, p)$, then $(X_n - np)/\sqrt{np(1-p)}$ converges in distribution to a standard normal random variable as $n \to \infty$. Note that there is no contradiction between this and our previous result that if n is large and p is small, then a $Bin(n, p)$ random variable behaves approximately like a $Poi(\lambda)$ random variable with $\lambda = np$. In that case, we implicitly took $n \to \infty$ and also $p \to 0$ such that $np \to \lambda$. Here, we are fixing p and letting $n \to \infty$.

Preparation for the proof

Let $g : \mathbb{R} \to \mathbb{R}$ be the function

$$g(x) = \begin{cases} e^{-1/x} & \text{if } x > 0, \\ 0 & \text{if } x \le 0. \end{cases}$$

Let $g^{(k)}$ denote the k^{th} derivative of g. It is easy to see that for any k, $g^{(k)}(x) = P_k(1/x)e^{-1/x}$, where P_k is a polynomial of degree k. Thus,

$$\lim_{x \downarrow 0} g^{(k)}(x) = 0,$$

because $e^{-1/x}$ approaches zero must faster than $P_k(1/x)$ blows up to ∞ as $x \downarrow 0$. Also, $g^{(k)}(x)$ remains bounded as $x \to \infty$. Since $g^{(k)}$ is continuous on $(0, \infty)$, these two facts imply that $g^{(k)}$ is uniformly bounded[71] on $(0, \infty)$.

Moreover, observe that since $g^{(k)}(x) = 0$ for any $x < 0$, we get that $g^{(k)}$ is well-defined and equals zero at $x = 0$.

To summarize, g belongs to the class C_b^∞ of functions on the real line that are infinitely differentiable with bounded derivatives (including the zeroth derivative, which is the function itself).

[70] How close? There is a quantitative bound on the rate of convergence in the central limit theorem, known as the **Berry–Esséen theorem**, which gives an answer to this question. We will not discuss this here. For this particular example, the best version of the Berry–Esséen theorem says that the actual probability is within ±0.00094 of 0.99.

[71] Try to give a complete proof using techniques from real analysis.

Now take any $-\infty < a < b < \infty$, and let

$$g_{a,b}(x) = g(x-a)g(b-x).$$

Then $g_{a,b}$ is also in $\mathcal{C}_b^\infty$, and it is zero everywhere except in (a,b), where it is strictly positive. Thus, if

$$C_{a,b} = \int_{-\infty}^{\infty} g_{a,b}(x)dx,$$

then $C_{a,b} > 0$, and if we define

$$h_{a,b}(x) = \frac{1}{C_{a,b}} g_{a,b}(x),$$

then $h_{a,b}$ is a probability density function. Let $H_{a,b}$ be the corresponding c.d.f. Since $h_{a,b}$ is zero outside (a,b) and strictly positive inside (a,b), the function $H_{a,b}$ equals 0 in $(-\infty, a]$, equals 1 in $[b,\infty)$ and is strictly increasing from 0 to 1 in (a,b). Moreover, $H_{a,b} \in \mathcal{C}_b^\infty$. These functions allow us to prove the following lemma.

Lemma 1. *Let X be a random variable, and suppose that $X_1, X_2, \ldots$ is a sequence of random variables (discrete or continuous) such that for each $f \in \mathcal{C}_b^\infty$, $E(f(X_n)) \to E(f(X))$ as $n \to \infty$. Then $X_n \to X$ in distribution.*

Proof. Let F_n be the c.d.f. of X_n and F be the c.d.f. of X. Take any $t \in \mathbb{R}$ where F is continuous, and some $s < t$. The function $1 - H_{s,t}$ defined above is everywhere less than or equal to the indicator function $1_{(-\infty,t]}$. Thus,

$$E(1 - H_{s,t}(X_n)) \le E(1_{(-\infty,t]}(X_n)) = F_n(t).$$

Since $H_{s,t} \in \mathcal{C}_b^\infty$, this gives

$$\liminf_{n\to\infty} F_n(t) \ge \lim_{n\to\infty} E(1 - H_{s,t}(X_n)) = E(1 - H_{s,t}(X)).$$

On the other hand, $1 - H_{s,t}$ is everywhere bigger than or equal to the function $1_{(-\infty,s]}$. This gives

$$E(1 - H_{s,t}(X)) \ge E(1_{(-\infty,s]}(X)) = F(s).$$

Combining, we get

$$\liminf_{n\to\infty} F_n(t) \ge F(s).$$

Note that this holds for any $s < t$. Since F is continuous at t, we can now take $s \uparrow t$ and get

$$\liminf_{n\to\infty} F_n(t) \ge F(t). \tag{23}$$

Working similarly with $H_{t,s}$ for $s > t$, we get

$$\begin{aligned}\limsup_{n\to\infty} F_n(t) &\le \lim_{n\to\infty} E(1 - H_{t,s}(X_n)) \\ &= E(1 - H_{t,s}(X)) \le F(s).\end{aligned}$$

Taking $s \downarrow t$ gives

$$\limsup_{n\to\infty} F_n(t) \le F(t). \tag{24}$$

Combining (23) and (24), we get[72]

$$\lim_{n\to\infty} F_n(t) = F(t),$$

which proves that $X_n \to X$ in distribution. □

The Lindeberg method

We will now prove the central limit theorem, under the additional assumption that $E|X_i - \mu|^3 < \infty$. The proof technique is known as **Lindeberg's method**, which is also useful for various other problems.

Fix n. Define

$$Y_i = \frac{X_i - \mu}{\sqrt{n}\sigma}.$$

Then $Y_1, Y_2, \ldots, Y_n$ are also i.i.d., with $E(Y_i) = 0$ and $E(Y_i^2) = 1/n$. Let

$$T_n = \sum_{i=1}^n Y_i = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

We have to show that $T_n \to Z$ in distribution, where $Z \sim N(0,1)$. By Lemma 1, it suffices to prove that for every $f \in \mathcal{C}_b^\infty$,

$$\lim_{n\to\infty} E(f(T_n)) = E(f(Z)). \tag{25}$$

Accordingly, take any $f \in \mathcal{C}_b^\infty$. Fix some $n \ge 1$. Let $Z_1, \ldots, Z_n$ be i.i.d. $N(0, 1/n)$ random variables, so that we can write[73]

$$Z = \sum_{i=1}^n Z_i.$$

For $i = 0, \ldots, n$, let

$$A_i = Y_1 + \cdots + Y_{i-1} + Y_i + Z_{i+1} + \cdots + Z_n.$$

Note that $A_0 = Z$ and $A_n = T_n$. Thus[74]

$$\begin{aligned}f(T_n) - f(Z) &= f(A_n) - f(A_0) \\ &= \sum_{i=1}^n (f(A_i) - f(A_{i-1})).\end{aligned} \tag{26}$$

[72] Recall that it is a familiar technique from real analysis to show that a sequence of numbers x_n converges to a limit x by proving that $\limsup_{n\to\infty} x_n \le x$ and $\liminf_{n\to\infty} x_n \ge x$.

[73] Recall that a linear combination of normal random variables is again normal, with appropriate mean and variance.

[74] This is known as a *telescoping sum*.

Take any i. Let

$$B_i = Y_1 + \cdots + Y_{i-1} + Z_{i+1} + \cdots + Z_n,$$

so that

$$A_i = B_i + Y_i, \quad A_{i-1} = B_i + Z_i. \tag{27}$$

Let C be a uniform upper bound on $|f'''(x)|$. Then by Taylor series expansion,

$$\left| f(A_i) - f(B_i) - Y_i f'(B_i) - \frac{Y_i^2}{2} f''(B_i) \right| \le \frac{C|Y_i|^3}{6}.$$

Therefore, by the inequality $|E(X)| \le E|X|$,

$$\begin{aligned} &\left| E\left(f(A_i) - f(B_i) - Y_i f'(B_i) - \frac{Y_i^2}{2} f''(B_i) \right) \right| \\ &\le E\left| f(A_i) - f(B_i) - Y_i f'(B_i) - \frac{Y_i^2}{2} f''(B_i) \right| \le \frac{CE|Y_i|^3}{6}. \end{aligned}$$

Now note that Y_i and B_i are independent. Thus[75],

$$E(Y_i f'(B_i)) = E(Y_i)E(f'(B_i)) = 0$$

and

$$E(Y_i^2 f''(B_i)) = E(Y_i^2)E(f''(B_i)) = \frac{E(f''(B_i))}{n}.$$

Therefore,

$$\left| E\left(f(A_i) - f(B_i) - \frac{f''(B_i)}{2n} \right) \right| \le \frac{CE|Y_i|^3}{6}. \tag{28}$$

By the same argument applied to A_{i-1} instead of A_i, and using the second equation in (27), we get

$$\left| E\left(f(A_{i-1}) - f(B_i) - \frac{f''(B_i)}{2n} \right) \right| \le \frac{CE|Z_i|^3}{6}. \tag{29}$$

From (28) and (29), it follows that

$$|E(f(A_i) - f(A_{i-1}))| \le \frac{C}{6}(E|Y_i|^3 + E|Z_i|^3).$$

But $E|Y_i|^3 = E|Y_1|^3$ and $E|Z_i|^3 = E|Z_1|^3$. Therefore by the above inequality and the telescoping sum (26), we get

$$\begin{aligned} |E(f(T_n)) - E(f(Z))| &\le \frac{C}{6} \sum_{i=1}^{n} (E|Y_i|^3 + E|Z_i|^3) \\ &= \frac{Cn}{6}(E|Y_1|^3 + E|Z_1|^3). \end{aligned}$$

But

$$E|Y_1|^3 = \frac{1}{n^{3/2}\sigma^3} E|X_1 - \mu|^3, \quad E|Z_1|^3 = \frac{1}{n^{3/2}} E|Z|^3.$$

Thus, taking $n \to \infty$, we get (25), which completes the proof of the central limit theorem under the assumption that $E|X_1 - \mu|^3$ is finite.

[75] Here we are implicitly using the result that if X and Y are independent random variables, then $f(X)$ and $g(Y)$ are also independent for any functions f and g. To see this, note that for any sets A and B,

$$\begin{aligned} &P(f(X) \in A, g(Y) \in B) \\ &= P(X \in f^{-1}(A), Y \in g^{-1}(B)) \\ &= P(X \in f^{-1}(A))P(Y \in g^{-1}(B)) \\ &= P(f(X) \in A)P(g(Y) \in B). \end{aligned}$$

The multivariate central limit theorem

The notion of convergence in distribution for random vectors is a bit more complicated than that for random variables. There are many equivalent definitions, one of which is the following. We say that a sequence of random vectors $X_1, X_2, \ldots$ converges in distribution to a random vector X if

$$\lim_{n\to\infty} P(X_n \in A) = P(X \in A) \tag{30}$$

for any[76] set A such that $P(X \in \partial A) = 0$, where ∂A denotes the boundary[77] of A.

Theorem 6 (Multivariate central limit theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. d-dimensional random vectors with mean vector μ and covariance matrix Σ. For each n, let $S_n = X_1 + \cdots + X_n$. Then the random vector $n^{-1/2}(S_n - n\mu)$ converges in distribution to $N(0, \Sigma)$.*

The proof of this theorem goes exactly as the proof of the univariate CLT via Lindeberg's method. The first step is to show that for any $f \in C_b^\infty(\mathbb{R}^d)$ (where $C_b^\infty(\mathbb{R}^d)$ is the set of all infinitely differentiable maps from $\mathbb{R}^d$ into $\mathbb{R}$ with bounded derivatives of all orders),

$$\lim_{n\to\infty} E[f(n^{-1/2}(S_n - n\mu))] = E[f(Z)],$$

where $Z \sim N(0.\Sigma)$ The proof of this follows by Lindeberg's method using multivariate Taylor expansion[78]. To complete the proof, we approximate the function 1_A (where A is as in (30)) from above and below by smooth functions, just as we approximated indicators of intervals by the functions $h_{a,b}$ in the univariate case. It is a more complicated here because A can be any set such that $P(Z \in \partial A) = 0$. This is beyond the scope of this discussion. However, in the special case when A is of the form $[a_1, b_1] \times \cdots \times [a_d, b_d]$ for some intervals $[a_1, b_1], \ldots, [a_d, b_d]$, the approximations to 1_A can be easily constructed by considering functions like $f(x_1, \ldots, x_d) = f_1(x_1) \cdots f_d(x_d)$, where f_i is a smooth approximation of $1_{[a_i, b_i]}$.

Example: Number of points in a region

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random vectors distributed uniformly on the unit disk $D = \{(x, y) : x^2 + y^2 \le 1\}$. For each n, you can view $X_1, \ldots, X_n$ are a set of n points distributed independently and uniformly in D.

Let T be the top half of D, that is

$$T = \{(x, y) \in D : y \ge 0\}.$$

[76] As usual, 'any set' means 'any Borel set'.

[77] That is, the set of points which have sequences converging to them from both A and A^c.

[78] Try to fill in the details.

Similarly, let Q be the top-right quarter of D, that is

$$Q = \{(x, y) \in D : x \geq 0, y \geq 0\}.$$

For each n, let A_n be the number of points among $X_1, \ldots, X_n$ that fall in T, and let B_n be the number points that fall in Q. Clearly, $A_n \sim Bin(n, 1/2)$ and $B_n \sim Bin(n, 1/4)$. So, for large n, the central limit theorem tells us that A_n behaves like a $N(n/2, n/4)$ random variable, and B_n behaves like a $N(n/4, 3n/16)$ random variable. What about their joint distribution? To understand this, let us define for each i,

$$Y_i = 1_{\{X_i \in T\}}, \quad Z_i = 1_{\{X_i \in Q\}}.$$

Then the pairs $(Y_1, Z_1), (Y_2, Z_2), \ldots$ are i.i.d. random vectors. A simple calculation shows that $E(Y_i) = E(Y_i^2) = 1/2$, $E(Z_i) = E(Z_i^2) = 1/4$, and $E(Y_i Z_i) = 1/4$. Thus, $Var(Y_i) = 1/4$, $Var(Z_i) = 3/16$, and $Cov(Y_i, Z_i) = 1/8$. Since $A_n = Y_1 + \cdots + Y_n$ and $B_n = Z_1 + \cdots + Z_n$, the multivariate central limit theorem shows that for large n, the pair (A_n, B_n) behaves like a bivariate normal random variable with mean vector $(n/2, n/4)$, and covariance matrix

$$\begin{pmatrix} n/4 & n/8 \\ n/8 & 3n/16 \end{pmatrix}.$$

More precisely, $n^{-1/2}(A_n - n/2, B_n - n/4)$ converges in distribution to a bivariate normal random vector with mean zero and covariance matrix

$$\begin{pmatrix} 1/4 & 1/8 \\ 1/8 & 3/16 \end{pmatrix}.$$

Central limit theorem for sums of dependent random variables

Often in practice we encounter random variables that are sums of dependent (instead of independent) random variables. For example, consider the number of heads runs in a sequence of coin tosses. We showed that it can be expressed as a sum of indicator random variables, but those variables were not independent. There are many ways to prove central limit theorems for sums of dependent random variables. The following result, in combination with Lemma 1, is applicable in a wide variety of problems[79].

Theorem 7. *Let $X_1, \ldots, X_n$ be random variables defined on the same sample space. Suppose that for each i, there is a set of indices N_i containing i such that X_i and the random vector $(X_j)_{j \notin N_i}$ are independent. Suppose moreover that there is a set of indices $M_i \supseteq N_i$ such that the random vectors $(X_j)_{j \in N_i}$ and $(X_j)_{j \notin M_i}$ are indepedent. Take any $f \in C_b^\infty$. Let*

[79] This is a special case of the so-called **dependency graph approach** for proving central limit theorems. Usually such results are proved using **Stein's method**. But since that is beyond the scope of this discussion, a different — and arguably simpler — proof is given here.

C be a number such that $|f''(x)|$ and $|f'''(x)|$ are $\le$ C for all x. Let K be a number such that the size of M_i is $\le$ K for all i. Let L be a number such that $E|X_i - E(X_i)|^3 \le L$ for all i. Let $S = X_1 + \cdots + X_n$ and $\sigma^2 = Var(S)$, and define

$$T = \frac{S - E(S)}{\sigma}.$$

Let $Z \sim N(0,1)$. Then

$$|E(f(T)) - E(f(Z))| \le \frac{9\pi n C K^2 L}{4\sigma^3}.$$

To quickly see how this result implies the CLT for i.i.d. sums, note that if $X_1, \ldots, X_n$ are i.i.d., we can take $N_i = \{i\}$, which gives $K = 1$. Also, in this case $\sigma^2 = nVar(X_1)$. Plugging in these quantities into the above bound, we get a bound of order $n^{-1/2}$, which is the same as what we got using Lindeberg's method. We will see a more nontrivial example in the next section.

The proof of Theorem 7 is divided into a number of steps. First, define

$$Y_i = \frac{X_i - E(X_i)}{\sigma},$$

so that $T = \sum_{i=1}^n Y_i$. Note that $E(Y_i) = 0$ for each i. Let $Z = (Z_1, \ldots, Z_n)$ be a normal random vector with mean zero and covariance matrix equal[80] to the covariance matrix of $Y = (Y_1, \ldots, Y_n)$.

Note that the vector Z also has the properties that for each i, Z_i and $(Z_j)_{j \notin N_i}$ are independent, and $(Z_j)_{j \in N_i}$ and $(Z_j)_{j \notin M_i}$ are independent. This is because independence of normal random vectors is guaranteed when covariances are zero, as we noted before.

Another important observation is that

$$E|Z_i|^3 \le \frac{3L}{\sigma^3}. \tag{31}$$

To see this, first note that Z_i is normal with mean zero and

$$Var(Z_i) = Var(Y_i) = E(Y_i^2).$$

But by Proposition 10,

$$E(Y_i^2) \le (E|Y_i|^3)^{2/3} \le \frac{L^{2/3}}{\sigma^2}. \tag{32}$$

An easy calculation shows that if $X \sim N(0, a^2)$, then $E|X|^3 \le 3a^3$. This gives (31).

Let $V = Z_1 + \cdots + Z_n$. Then $V \sim N(0,1)$. For each $a, b \in [0,1]$, define

$$\phi(a,b) = E(f(aT + bV)).$$

Our goal is to obtain an upper bound on $|\phi(1,0) - \phi(0,1)|$. We will start by producing an upper bound for $|\phi(a,b) - \phi(c,d)|$ for arbitrary $a, b, c, d \in [0,1]$. The following lemma is the first step in that direction.

[80] We know that this is possible since we can construct a normal random vector with covariance matrix equal to any given positive definite matrix, and the covariance matrix of any random vector is positive definite. A small issue we are skirting here is that $Cov(Y)$ may be positive semidefinite instead of positive definite (that is, may have a zero eigenvalue). This is actually not a problem, because normal random vectors are allowed to have singular covariance matrices (although we did not discuss that).

Lemma 2. *Take any $a, b, c, d \in [0,1]$. Let $U = (a-c)T + (b-d)V$. Then*

$$|\phi(a,b) - \phi(c,d) - E(Uf'(cT+dV))| \leq \frac{C\alpha}{2}.$$

where $\alpha = (a-c)^2 + (b-d)^2$.

Proof. By Taylor approximation and the fact that $|f''|$ is uniformly bounded by C, we get

$$|f(aT+bV) - f(cT+dV) - Uf'(cT+dV)| \leq \frac{CU^2}{2}.$$

Since T and V are independent, $E(T) = E(V) = 0$, and $E(T^2) = E(V^2) = 1$, we get $E(U^2) = \alpha$. This completes the proof of the lemma. □

Our next goal is to get an upper bound for the first order Taylor approximation of $\phi(a,b) - \phi(c,d)$ that we got from Lemma 2. First, note that

$$E(Uf'(cT+dV)) = \sum_{i=1}^{n} E(U_i f'(cT+dV)),$$

where

$$U_i = (a-c)Y_i + (b-d)Z_i.$$

For each i, let

$$A_i = \sum_{j \in N_i} (cY_j + dZ_j), \quad B_i = \sum_{j \notin N_i} (cY_j + dZ_j),$$

so that $A_i + B_i = cT + dV$. The following lemma gives a first approximation for $E(U_i f'(cT+dV))$.

Lemma 3. *For any i,*

$$|E(U_i f'(cT+dV)) - E(U_i A_i f''(B_i))| \leq \frac{3\beta C K^2 L}{2\sigma^3},$$

where $\beta = \max\{|a-c|, |b-d|\}$.

Proof. Note that by the given conditions, U_i is independent of B_i. Since $E(Y_i) = E(Z_i) = 0$, this gives

$$E(U_i f'(B_i)) = 0.$$

Thus,

$$E(U_i f'(cT+dV)) = E[U_i(f'(cT+dV) - f'(B_i))].$$

Again by Taylor approximation,

$$|f'(cT+dV) - f'(B_i) - A_i f''(B_i)| \le \frac{CA_i^2}{2}.$$

Combining, we get

$$|E(U_i f'(cT+dV)) - E(U_i A_i f''(B_i))| \le \frac{C}{2} E|U_i A_i^2|.$$

Now, if we expand $U_i A_i^2$ using the distributive law, then we get a sum of at most K^2 terms, each of which is of the form $\theta Q_1 Q_2 Q_3$, where each Q_i is either Y_j or Z_j for some j, and θ is one of the numbers $(a-c)c$, $(a-c)d$, $(b-d)c$ and $(b-d)d$. By the arithmetic-mean-geometric-mean (AM-GM) inequality,

$$|Q_1 Q_2 Q_3| \le \frac{|Q_1|^3 + |Q_2|^3 + |Q_3|^3}{3}.$$

By the assumption that $E|X_j - E(X_j)|^3 \le L$ and the inequality (31), we get that $E|Q_i|^3 \le 3L/\sigma^3$ for each i. Combining this with the fact that $|\theta| \le \max\{|a-c|, |b-d|\}$ (because $c, d \in [0,1]$), we get the desired bound. □

Next, define

$$C_i = \sum_{j \in M_i \setminus N_i} (cY_j + dZ_j), \quad D_i = \sum_{j \notin M_i} (cY_j + dZ_j),$$

so that $B_i = C_i + D_i$. The following lemma gives a second approximation for $E(U_i f'(cT+dV))$.

Lemma 4. *For any i,*

$$|E(U_i A_i f''(B_i)) - E(U_i A_i) E(f''(D_i))| \le \frac{3\beta C K^2 L}{\sigma^3},$$

where $\beta = \max\{|a-c|, |b-d|\}$.

Proof. By the given assumptions, $U_i A_i$ and D_i are independent. Therefore by Taylor approximation,

$$\begin{aligned}
&|E(U_i A_i f''(B_i)) - E(U_i A_i) E(f''(D_i))| \\
&= |E(U_i A_i (f''(B_i) - f''(D_i))| \\
&\le CE|U_i A_i C_i|.
\end{aligned}$$

We now expand $U_i A_i C_i$ using the distributive law. We end up with at most K^2 terms, where each term is of the form $\theta Q_1 Q_2 Q_3$, as in the proof of the previous lemma. Bounding the terms in the same way, we get the required bound. □

Finally, we obtain a bound on $|E(U_iA_i)|$.

Lemma 5. *For any i,*

$$|E(U_iA_i)| \leq \frac{|\gamma|KL^{2/3}}{\sigma^2}$$

where $\gamma = (a-c)c + (b-d)d$.

Proof. Since the Y_j's and Z_j's are independent and have mean zero,

$$E(U_iA_i) = (a-c)c \sum_{j\in N_i} E(Y_iY_j) + (b-d)d \sum_{j \in N_i} E(Z_iZ_j).$$

But the $E(Y_iY_j) = E(Z_iZ_j)$ by construction of Z. Therefore

$$E(U_iA_i) = \gamma \sum_{j\in N_i} E(Y_iY_j).$$

By the AM-GM inequality and the bound (32), we get

$$|E(Y_iY_j)| \leq E|Y_iY_j| \leq \frac{E(Y_i^2)+E(Y_j^2)}{2} \leq \frac{L^{2/3}}{\sigma^2}.$$

Since the size of N_i is at most K, this gives the required bound. □

Combining Lemmas 2, 3, 4 and 5, we get

$$\begin{aligned}
&|\phi(a,b)-\phi(c,d)| \\
&\leq |\phi(a,b)-\phi(c,d) - E(Uf'(cT+dV))| \\
&\quad + \sum_{i=1}^n |E(U_if'(cT+dV)) - E(U_iA_if''(B_i))| \\
&\quad + \sum_{i=1}^n |E(U_iA_if''(B_i)) - E(U_iA_i)E(f''(D_i))| \\
&\quad + \sum_{i=1}^n |E(U_iA_i)E(f''(D_i))| \\
&\leq \frac{C\alpha}{2} + \frac{9n\beta CK^2L}{2\sigma^3} + \frac{n|\gamma|CKL^{2/3}}{\sigma^2},
\end{aligned}$$

where $\alpha = (a-c)^2 + (b-d)^2$, $\beta = \max\{|a-c|, |b-d|\}$, an $\gamma = (a-c)c + (b-d)d$. We will be interested in the special situation where $a^2+b^2 = c^2+d^2$. The following lemma reduces the above bound to a more convenient form in this scenario.

Lemma 6. *Let $a,b,c,d \in [0,1]$ be such that $a^2+b^2 = c^2+d^2$. Then*

$$|\phi(a,b)-\phi(c,d)| \leq \frac{C\alpha}{2} + \frac{9n\beta CK^2L}{2\sigma^3} + \frac{n\alpha CKL^{2/3}}{2\sigma^2},$$

where $\alpha = (a-c)^2+(b-d)^2$ and $\beta = \max\{|a-c|, |b-d|\}$.

Proof. Let α and γ be as above. Since $a^2 + b^2 = c^2 + d^2$,

$$\begin{aligned}\gamma &= (a-c)\left(\frac{c+a}{2} + \frac{c-a}{2}\right) + (b-d)\left(\frac{d+b}{2} + \frac{d-b}{2}\right) \\ &= \frac{1}{2}(a^2 - c^2 - (a-c)^2 + b^2 - d^2 - (b-d)^2) \\ &= -\frac{1}{2}((a-c)^2 + (b-d)^2) = -\frac{\alpha}{2}.\end{aligned}$$

The proof is completed by plugging this into the bound obtained above. □

We are now ready to complete the proof of Theorem 7.

Proof of Theorem 7. Take any integer $m \geq 1$. Let $a_k = \cos(\pi k/2m)$ and $b_k = \sin(\pi k/2m)$ for $k = 0, 1, \ldots, m$. Then

$$\begin{aligned}|E(f(T)) - E(f(V))| &= |\phi(1,0) - \phi(0,1)| \\ &= |\phi(a_0, b_0) - \phi(a_m, b_m)| \\ &\leq \sum_{k=0}^{m-1} |\phi(a_k, b_k) - \phi(a_{k+1}, b_{k+1})|.\end{aligned}$$

Since sin and cos are Lipschitz functions, $|a_k - a_{k+1}|$ and $|b_k - b_{k+1}|$ are bounded by $\pi/2m$ for every k. Therefore by Lemma 6 and the above inequality,

$$|E(f(T)) - E(f(V))| \leq \frac{C\pi^2}{8m} + \frac{9\pi n CK^2 L}{4\sigma^3} + \frac{n\pi^2 CKL^{2/3}}{8m\sigma^2}.$$

But m is arbitrary. So we can now send $m \to \infty$ and get the required bound. □

Example: Number of head runs

Let S_n be the number of head runs in n tosses of a p-coin. We have seen that S_n can be expressed as

$$S_n = \sum_{i=1}^{n} 1_{A_i},$$

where A_1 is the event that toss 1 turns up heads, and for $i \geq 2$, A_i is the event that toss i is heads and toss $i-1$ is tails. We will now prove a central limit theorem for S_n. To put this problem in the framework of Theorem 7, fix n and let $X_i = 1_{A_i}$. Take

$$N_i = \{1 \leq j \leq n : |j - i| \leq 1\}.$$

Then clearly, $i \in N_i$ and X_i and $(X_j)_{j \notin N_i}$ are independent. Next, let

$$M_i = \{1 \leq j \leq n : |j - i| \leq 2\}.$$

Again, it is clear that $M_i \supseteq N_i$, and $(X_j)_{j\in N_i}$ and $(X_j)_{j\notin M_i}$ are independent. Since the size of M_i is bounded above by 5 for each i, we can take $K = 5$. Next, observe that $E|X_i - E(X_i)|^3 \leq 1$, which means that we can take $L = 1$. Finally, an easy computation shows that

$$\begin{aligned} Var(S_n) &= \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Var(X_i) + 2 \sum_{i=2}^{n-1} Cov(X_i, X_{i+1}) + 2Cov(X_1, X_2) \\ &= np(1-p) - 2(n-2)p^2(1-p)^2 - 2p^2(1-p). \end{aligned}$$

If p is not 0 or 1 (which we may assume, to avoid trivialities), it can be verified that

$$p(1-p) > 2p^2(1-p)^2.$$

Therefore the above formula shows that $Var(S_n)$ behaves like a positive constant times n when n is large. Therefore by Theorem 7, we get that for any $f \in \mathcal{C}_b^\infty$,

$$\lim_{n\to\infty} E(f(T_n)) = E(f(Z)),$$

where $Z \sim N(0,1)$ and

$$T_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}}.$$

So, by Lemma 1, $T_n \to Z$ in distribution.

More about variance and covariance

The Cauchy–Schwarz inequality

Let X and Y be any two random variables. The following very useful inequality is called the **Cauchy–Schwarz inequality**:

$$|E(XY)| \le \sqrt{E(X^2)E(Y^2)}.$$

To prove this, note that by the AM-GM inequality,

$$uv \le \frac{u^2 + v^2}{2} \tag{33}$$

for any $u, v \ge 0$. Therefore, for any two nonnegative random variables U and V with $E(U^2) = E(V^2) = 1$, we have

$$E(UV) \le \frac{E(U^2) + E(V^2)}{2} = 1. \tag{34}$$

Now take any X and Y, and let

$$U = \frac{|X|}{\sqrt{E(X^2)}}, \quad V = \frac{|Y|}{\sqrt{E(Y^2)}}.$$

Then U and V are nonnegative random variables with $E(U^2) = E(V^2) = 1$. So we can apply (34). But that can be rewritten as

$$E|XY| \le \sqrt{E(X^2)E(Y^2)}.$$

Finally, the Cauchy–Schwarz inequality is obtained using $|E(XY)| \le E|XY|$.

A slightly different version of the Cauchy–Schwarz inequality, for the covariance of two random variables X and Y, is obtained by replacing X with $X - E(X)$ and Y with $Y - E(Y)$. This gives

$$\begin{aligned} |Cov(X,Y)| &= |E[(X - E(X))(Y - E(Y))] \\ &\le \sqrt{E(X - E(X))^2 E(Y - E(Y))^2} \\ &= \sqrt{Var(X)Var(Y)}. \end{aligned}$$

Correlation

The **correlation** between two random variables X and Y is defined as

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

By the Cauchy–Schwarz inequality the correlation between any two random variables is always a number between -1 and 1. It is not difficult to show that the correlation is 1 if and only if one random variable is an increasing linear function of the other, and -1 if and only if one random variable is a decreasing linear function of the other. The main step in proving this is that the difference between the two sides of the AM-GM inequality (33) equals $(u-v)^2/2$, which is zero if and only if $u = v$. The remaining details are left to the reader.

Bivariate normal distribution

The distribution of a pair of jointly normal random variables (X, Y) is called a **bivariate normal distribution**. Recall that a normal distribution is fully specified by its means and covariances. When the dimension is two, we have means, two variances, and one covariance (since covariance is symmetric). But the covariance can be expressed as correlation times the product of the standard deviations. Thus, a bivariate normal distribution is characterized by five parameters — two means μ_1 and μ_2, two variances σ_1^2 and σ_2^2, and one correlation, usually denoted by ρ. We write $(X,Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. The covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Suppose that $|\rho| < 1$. Then Σ is invertible and

$$\Sigma^{-1} = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}.$$

Also, $\det\Sigma = (1-\rho^2)\sigma_1^2\sigma_2^2$. Therefore, the p.d.f. of (X,Y) at a point (x,y) is

$$f_{X,Y}(x,y) = \frac{\exp\left(-\frac{\sigma_2^2(x-\mu_1)^2+\sigma_1^2(y-\mu_2)^2-2\rho\sigma_1\sigma_2(x-\mu_1)(y-\mu_2)}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}.$$

An important and useful result that follows from the above formula is the conditional distribution of Y given $X = x$. Since $X \sim N(\mu_1, \sigma_1^2)$, its p.d.f. at x is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right).$$

Therefore,

$$\begin{aligned}
f_{Y|X=x}(y) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
&= \frac{\exp\left(-\frac{\sigma_2^2\rho^2(x-\mu_1)^2+\sigma_1^2(y-\mu_2)^2-2\rho\sigma_1\sigma_2(x-\mu_1)(y-\mu_2)}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\right)}{\sqrt{2\pi(1-\rho^2)}\sigma_2} \\
&= \frac{\exp\left(-\frac{(\sigma_2\rho(x-\mu_1)-\sigma_1(y-\mu_2))^2}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\right)}{\sqrt{2\pi(1-\rho^2)}\sigma_2} \\
&= \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_2}\exp\left(-\frac{(y-\mu_2-\frac{\rho\sigma_2}{\sigma_1}(x-\mu_1))^2}{2\sigma_2^2(1-\rho^2)}\right).
\end{aligned}$$

This shows that given $X = x$, the conditional distribution of Y is $N(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2))$. Note that the conditional mean is a linear function of x and the conditional variance is a constant. Note that the unconditional variance of Y is σ_2^2. Thus, the information that $X = x$ reduces the variance of Y by a factor of $1-\rho^2$. If X and Y are strongly correlated (meaning that ρ is close to 1), the variance is reduced significantly. If, on the other hand, ρ is close to zero, there is not much reduction.

The Efron–Stein inequality

We have seen that upper bounds on variances are crucial for proving laws of large numbers. However, we have learnt only one method for calculating or bounding a variance — express the random variable as a sum of relatively simple random variables, and then express the variance as the sum of covariances. In many complicated problems, this is not possible (we will see an example soon). Fortunately, there is a simple upper bound, known as the **Efron–Stein inequality**, that is powerful enough to give useful upper bounds in a wide array of very complex problems — for which there are essentially no other ways of getting variance upper bounds[81].

Theorem 8 (Efron–Stein inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables (or vectors), and let $Y = f(X_1, \ldots, X_n)$ be a function of these variables such that $E(Y^2) < \infty$. Let $X_1', \ldots, X_n'$ be independent random variables (or vectors), independent of the X_i's, such that for each i, X_i' has the same distribution as X_i. Then*

$$Var(Y) \leq \frac{1}{2}\sum_{i=1}^{n} E[(Y - f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n))^2].$$

[81] There is a whole area of probability and analysis dedicated to understanding fluctuations of complicated random variables, known as **concentration of measure** or **concentration inequalities**. But as far as the order of fluctuations is concerned, there are still many problems where the Efron–Stein inequality gives the optimal or the (nearly) best available result.

Proof. For each i, let

$$X^{(i)} = (X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n),$$

and

$$X^{[i]} = (X_1', \dots, X_i', X_{i+1}, \dots, X_n).$$

We also define $X^{[0]} = (X_1, \dots, X_n)$. Then note that

$$\begin{aligned} Var(Y) &= E(Y^2) - (E(Y))^2 \\ &= E(Y^2) - E(Yf(X^{[n]})) \\ &= E[Y(f(X^{[0]}) - f(X^{[n]}))]. \end{aligned}$$

We can write the above as a telescoping sum:

$$Var(Y) = \sum_{i=1}^n E[Y(f(X^{[i-1]}) - f(X^{[i]}))].$$

Take any i. The random variable $Y(f(X^{[i-1]}) - f(X^{[i]}))$ is a function of the variables $X_1, \dots, X_n, X_1', \dots, X_n'$. Let us write it as

$$g(X_1, \dots, X_n, X_1', \dots, X_n'),$$

where $g : \mathbb{R}^{2n} \to \mathbb{R}$ is the function

$$\begin{aligned} &g(x_1, \dots, x_n, y_1, \dots, y_n) \\ &= f(x_1, \dots, x_n)(f(y_1, \dots, y_{i-1}, x_i, \dots, x_n) - f(y_1, \dots, y_i, x_{i+1}, \dots, x_n)). \end{aligned}$$

Now if we interchange X_i and X_i' in $g(X_1, \dots, X_n, X_1', \dots, X_n')$, the distribution of the resulting random variable remains unchanged. In particular, its expected value should remain the same as before. The random variable that comes out as a result of this interchange is

$$f(X^{(i)})(f(X^{[i]}) - f(X^{[i-1]})).$$

So, we get

$$E[Y(f(X^{[i-1]}) - f(X^{[i]}))] = E[f(X^{(i)})(f(X^{[i]}) - f(X^{[i-1]}))].$$

When two quantities are equal, their average is also the same quantity. Averaging the two sides of the above display gives

$$\frac{1}{2} E[(Y - f(X^{(i)}))(f(X^{[i-1]}) - f(X^{[i]}))].$$

Thus, we get

$$Var(Y) = \frac{1}{2} \sum_{i=1}^n E[(Y - f(X^{(i)}))(f(X^{[i-1]}) - f(X^{[i]}))].$$

Applying the Cauchy–Schwarz inequality to each term on the right gives

$$Var(Y) \leq \frac{1}{2}\sum_{i=1}^{n} \sqrt{E(Y - f(X^{(i)}))^2 E(f(X^{[i-1]}) - f(X^{[i]}))^2}.$$

Now, the expression $f(X^{[i-1]}) - f(X^{[i]})$ does not involve $X_1, \ldots, X_{i-1}$. So if we replace $X'_1, \ldots, X'_{i-1}$ in this expression by $X_1, \ldots, X_{i-1}$, its distribution should remain unchanged. This implies that

$$E(Y - f(X^{(i)}))^2 = E(f(X^{[i-1]}) - f(X^{[i]}))^2.$$

Plugging this into the previous bound completes the proof. □

Example: The traveling salesman problem

Let $X_1, \ldots, X_n$ be i.i.d. uniform distributed points from the unit square $[0, 1]^2$, where $n \geq 2$. In the **traveling salesman problem**, we seek a path through these points that starts and ends at the same vertex, and visits every vertex exactly once, such that the total length is minimized subject to these constraints. Let T_n be the length of this minimizing path. The goal of this section is to prove the following law of large numbers for T_n.

Theorem 9. *As $n \to \infty$, $T_n/E(T_n) \to 1$ in probability.*

In other words, when n is large, T_n is very likely to be close to its expected value, in the sense that the ratio of the two quantities is very likely to be close to 1. Actually, we will show a bit more. We will show that[82] $E(T_n) \geq C_1\sqrt{n}$ for some constant C_1 that does not depend on n, and we will show that $Var(T_n) \leq C_2$ where C_2 is another constant that does not depend on n. From these two results, we get $Var(T_n/E(T_n)) \to 0$ as $n \to \infty$. Since $E(T_n/E(T_n)) = 1$ for each n, this proves Theorem 9.

Fix n. For each i, let D_i be the distance of X_i to its nearest neighbor. among the other $n - 1$ points. The random variables $D_1, \ldots, D_n$ are not independent, but are identically distributed due to the symmetry of the situation.

Lemma 7. *For each i,*

$$E(D_i^2) \leq \frac{72}{\pi n} \quad \textit{and} \quad E(D_i) \geq \frac{1}{2\sqrt{2\pi n}}.$$

Proof. Since $D_1, \ldots, D_n$ are identically distributed, it suffices to prove the claim for $i = 1$. For $x \in \mathbb{R}^2$ and $t > 0$, let $B(x, t)$ denote the

[82] In fact, it is known that $E(T_n)/\sqrt{n}$ converges to a nonzero limit as $n \to \infty$. But we will not be able to prove that here.

Euclidean ball with center x and radius t. For $x \in [0,1]^2$ and $0 \leq t \leq 1/2$, it is not hard to see[83] that the area of the region $B(x,t) \cap [0,1]^2$ is bounded below by $\pi t^2/4$. On the other hand, if $1/2 \leq t \leq \sqrt{2}$, then $t/3 \leq 1/2$, which implies that the area of $B(x,t) \cap [0,1]^2$ is bounded below by the area of $B(x,t/3) \cap [0,1]^2$, which we know is bounded below by $\pi t^2/36$. For convenience, let us denote the constant $\pi/36$ by c.

Now take any $x \in [0,1]^2$. Since $X_1,\ldots,X_n$ are independent, the vectors $X_2,\ldots,X_n$, given $X_1 = x$, are still i.i.d. and uniformly distributed on $[0,1]^2$. Thus, for any $t \in [0,\sqrt{2}]$,

$$\begin{aligned} P(D_1 \geq t | X_1 = x) &= P(\cap_{i=2}^n \{\|X_i - x\| \geq t\}) \\ &= (P(\|X_2 - x\| \geq t))^{n-1} \\ &= (1 - \text{area}(B(x,t) \cap [0,1]^2))^{n-1} \\ &\leq (1 - ct^2)^{n-1}. \end{aligned}$$

Therefore, by the law of total probability for continuous random variables,

$$P(D_1 \geq t) = \int_{[0,1]^2} P(D_1 \geq t | X = x) dx \leq (1 - ct^2)^{n-1}.$$

If $t \geq \sqrt{2}$, the probability is zero since no two points in the unit cube can be at a distance greater than $\sqrt{2}$ from each other. Thus, by the tail integral formula for expectation,

$$\begin{aligned} E(D_1^2) &= \int_0^\infty 2t P(D_1 \geq t) dt \\ &\leq \int_0^\infty 2t(1 - ct^2)^{n-1} dt. \end{aligned}$$

Using the inequality $1 - x \leq e^{-x}$ that holds[84] for $x \geq 0$, we get

$$E(D_1^2) \leq \int_0^\infty 2t e^{-c(n-1)t^2} dt = \frac{1}{c(n-1)} \leq \frac{2}{cn},$$

where the last inequality holds because $n \geq 2$. This proves the desired upper bound.

On the other hand, by the inequality $(1-x)^m \geq 1 - mx$ that holds[85] for all $x \geq 0$ and positive integers m,

$$\begin{aligned} P(D_1 \geq t) &\geq (1 - \text{area}(B(x,t)))^{n-1} \\ &= (1 - \pi t^2)^{n-1} \geq 1 - \pi(n-1)t^2. \end{aligned}$$

By Markov's inequality, this gives

$$E(D_1) \geq t P(D_1 \geq t) \geq t(1 - \pi(n-1)t^2).$$

Taking $t = (2\pi(n-1))^{-1/2}$, we get

$$E(D_1) \geq \frac{1}{2\sqrt{2\pi(n-1)}} \geq \frac{1}{2\sqrt{2\pi n}},$$

[83] There are at least two sides of the square that are at distance $\geq 1/2$ from x, which means that at least one quarter of $B(x,t)$ must be contained in $[0,1]^2$.

[84] The two functions are equal at $x = 0$ and the derivative of $1 - x$ is dominated by the derivative of e^{-x} for all $x \geq 0$. When $x \geq 1$, the inequality is trivial.

[85] Again, check using derivatives.

which proves the required lower bound. □

The next lemma provides the first half of the argument for the proof of Theorem 9.

Lemma 8. *For any n,*

$$E(T_n) \geq \frac{\sqrt{n}}{2\sqrt{2\pi}}.$$

Proof. Let us fix a specific direction for traversing the optimal path. Then for X_i, there is a 'next point' on the path. Let us call it N_i. (Note that N_i is one of the other X_j's, but we do not care which one.) Clearly, $\|X_i - N_i\| \geq D_i$, and

$$T_n = \sum_{i=1}^{n} \|X_i - N_i\|.$$

Therefore by Lemma 7,

$$E(T_n) \geq \sum_{i=1}^{n} E(D_i) \geq \frac{n}{2\sqrt{2\pi n}},$$

which completes the proof. □

The second half of the argument for Theorem 9 is provided by the next lemma, which is proved using the Efron–Stein inequality.

Lemma 9. *For any n,*

$$Var(T_n) \leq \frac{144}{\pi}.$$

Proof. Fix n and take any $1 \leq i \leq n$. Let T_n' be the length of the new optimal path if we replace X_i by a new random vector X_i', which is also uniformly distributed on $[0,1]^2$ and is independent of $X_1, \ldots, X_n$. With the goal of applying the Efron–Stein inequality, we want to get an upper bound for $E(T_n - T_n')^2$. We do it in two steps. Let R_n be the length of the optimal path if we erase X_i and do not replace it with a new point. Then by the inequality[86] $(x+y)^2 \leq 2x^2 + 2y^2$, we get

$$E(T_n - T_n')^2 \leq 2E(T_n - R_n)^2 + 2E(R_n - T_n')^2.$$

But by the symmetry between X_i and X_i', the two terms on the right must be equal. Therefore

$$E(T_n - T_n')^2 \leq 4E(T_n - R_n)^2. \tag{35}$$

As before, let N_i be the point that comes after X_i in the optimal path, and now also let P_i be the point that comes before X_i. When we delete X_i, we can always find a path through the remaining $n-1$

[86] Expand $(x+y)^2$ and apply AM-GM to the cross-term.

points that is shorter than the old path, by declaring that N_i comes after P_i, keeping all else the same. Thus, $R_n \leq T_n$.

On the other hand, consider the optimal path through the remaining $n-1$ points, which we will henceforth call the 'second path'. Let Y_i be the nearest neighbor of X_i among the other $n-1$ points. Let Z_i be the point that comes after Y_i in the second path. Let us create a third path by redirecting the second path from Y_i to X_i and then to Z_i. This third path is a path through all n points, and it exceeds the length of the second path by

$$\|X_i - Y_i\| + \|Z_i - X_i\| - \|Z_i - Y_i\|.$$

This, by the triangle inequality, is bounded above by

$$2\|X_i - Y_i\| + \|Z_i - Y_i\| - \|Z_i - Y_i\| = 2\|X_i - Y_i\| = D_i.$$

Thus, $T_n \leq R_n + D_i$. Combining this with our previous observation that $R_n \leq T_n$, we get $|T_n - R_n| \leq D_i$. Therefore by Lemma 7 and (35),

$$E(T_n - T_n')^2 \leq 4E(D_i^2) \leq \frac{288}{\pi n}.$$

Summing over $i = 1, \ldots, n$, dividing by 2, and applying the Efron–Stein inequality completes the proof. □

As discussed before, Lemmas 8 and 9 jointly complete the proof of Theorem 9.

Incidentally, it is an **open problem** to prove a central limit theorem for T_n. It is believed (folklore) that $T_n - E(T_n)$ should converge in distribution to $N(0, \sigma^2)$ for some $\sigma^2 > 0$ as $n \to \infty$. The key obstacle to proving this is that we do not know how to show that the optimal path does not change much if a small number of points (or even one point) are shifted to new locations. It is conjectured that the optimal path has this kind of stability.