MATH 461

Probability Theory

Semester: Spring 2026

Professor: Neel Maitra

Contents

1 Combinatorial Analysis

Probability is a field that mathematically models uncertainty. One important skill is to know the number of ways that things can occur – the mathemaical theory behind this is combinatorial analysis.

1.1 The basic principles of counting

Definition 1.1 (Basic Principle of Counting)**.** If two experiments are performed where experiment 1 has m outcomes, and for each outcome of experiment 1, there are n outcomes for experiment 2, then

$$\text{Total number of outcomes} = m \times n \tag{1.1}$$

Proof. We start by enumerating all the possible outcomes for both experiments

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

where (i,j) represents the i^{th} outcome for experiment 1 and the j^{th} outcome for experiment 2 ■

Example 0

How many different two letter combinations are there?

Solution

There are 26 possible choices for the first letter and 26 possible choices for the second letter. Thus

$$26 \times 26 = 676$$

We can extend the basic principle of counting to the generalised basic principle of counting

Definition 1.2 (Generalised Basic Principle of Counting)**.** If r experiments are performed where the first experiment has n_1 outcomes, the second has n_2 outcomes,..., the r^{th} has n_r possible outcomes. Then,

$$\text{Total number of outcomes} = \prod_{i=1}^{r} n_i = n_1 \times \cdots n_r \tag{1.2}$$

Example 1

How many different seven-letter licence plates are there if the first three places are occupied by letters and the last four are occupied by numbers. No repetitions are allowed.

Solution

There are 26 choices for the first letter. Since no repetitions are allowed, there are 25 choices for the second letter and so on. There are 10 choices for the first number, 9 choices for the second number and so on. Thus,

$$\text{Total outcomes} = 26 \times 25 \times 24 \times 10 \times 9 \times 8 \times 7$$

Example 2

Consider a set A of n elements, $A = \{1, 2, \cdots, n\}$. How many different subsets of A exist?

Solution

There are n experiments corresponding to each element. For a given element, there are 2 outcomes: it is in the subset or not. Thus,

$$\text{Number of subsets} = 2^n$$

1.2 Permutations

Permutations refer to the ordering of objects

Definition 1.3 (Permutations)**.** Given a set of n objects, a permutation is a given ordering of the set of objects.

The number of permutations of n objects is

$$\text{Number of permutations} = n!$$

Proof. There are initially n objects for the first position in our odering, $n-1$ objects for the second,..., 1 object for the n^{th} position. Thus, by the generalised principle of counting,

$$\text{Number of orderings} = n \times n - 1 \times \cdots \times 1 = n! \tag{1.3}$$

■

Example 3

There are 10 books to be arranged on a bookshelf: 4 math books, 3 chemistry books, and 3 history books. If the math books have to be arranged together, how many arrangements of the books are there?

Solution

There are 4! ways to arrange the math books amongst themselves. If we now treat the math books as one unit, there are 7! ways to arrange the remaining books. Thus,

$$\text{Number of arrangements} = 7! \times 4!$$

If we have n objects, of which n_1 objects are alike, n_2 objects are like,...,n_r objects are like. The number of permutations is

$$\text{Number of permutations} = \frac{n!}{\prod_{i=1}^{r} n_i} \tag{1.4}$$

Example 4

Find the number of ways to arrange the letters in the word "PEPPER"

Solution

PEPPER has 6 letters so there are 6! arrangements. However, P repeats 3 times and E repeats 2 times. Thus, we have overcounted and so,

$$\text{Number of arrangements} = \frac{6!}{3! \times 2!}$$

1.3 Combinations

We are often interested in understanding how many groups of r objects can be selected from n objects, this is the idea of a combination

Definition 1.4 (Combination)**.** A combination refers to the number of unordered selections of r objects that can be made from a set of n objects. This number is given by

$$\text{Number of selections} = \binom{n}{r} = \frac{n!}{(n-r)!r!}, \qquad 0 \leq r \leq n \tag{1.5}$$

Proof. The number of choices to pick the first object from the group is n. The number of choices for the second object is $n-1$. The number of choices for the r^{th} object is $n-r+1$. Thus, the number of ordered selections is given by the generalised basic principle of counting,

$$n \times (n-1) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!}$$

Let X be the number of groups of r objects that can be formed from our n objects. Thus, within these groups there are $r!$ ways to permute objects. Thus, the number of different ordered selections that can be made is

$$X \cdot r! \tag{1.6}$$

Equating the two,

$$X \cdot r! = \frac{n!}{(n-r)!} \tag{1.7}$$

$$\implies X = \frac{n!}{(n-r)!r!} \tag{1.8}$$

■

$\binom{n}{r}$ is also known as a binomial coefficient since it is the coefficient of the terms in the binomial theorem, whose combinatorial proof is given at the end of this subsection.

By convention, $0! = 1$, thus

$$\binom{n}{0} = \binom{n}{n} = 1 \tag{1.9}$$

We also define

$$\binom{n}{k} = 0, \quad k < 0 \quad \text{or} \quad k > n \tag{1.10}$$

Example 5

Find the number of three letter groups that can be made from the alphabets.

Solution

There are 26 alphabets, of which we need to select 3. Thus, the total number of selections is

$$\text{Number of selections} = \binom{26}{3}$$

Example 6

From a group of 5 women and 7 men, find the number of different committees of 2 women and 3 men that can be made if 2 men are feuding and refuse to be in a committee together.

Solution

The total number of committees that can be made without any restrictions is $\binom{5}{2} \times \binom{7}{3}$.

The number of committees that can be made where the feuding men are in the committee are $\binom{5}{2} \times \binom{5}{1}$ (the 2 feuding men must be included in the committee, so we have to pick 1 men out of the remaining 5). Thus, the number of committees where the 2 feuding men are not together are

$$\text{Number of committees} = \binom{5}{2} \times \binom{7}{3} - \binom{5}{2} \times \binom{5}{1}$$

We can also think of combinations as the number of subsets of size r that can be made of a set of size n. Thus, when $r = n$, $\binom{n}{n} = 1$ because only one subset of size n exists in a set of size n: the set itself.

A useful identity involving binomial coefficients follows,

Identity 1.1.

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

Proof.

$$\begin{aligned}
\binom{n-1}{r-1} + \binom{n-1}{r} &= \frac{(n-1)!}{(n-r)!(r-1)!} + \frac{(n-1)!}{(n-r-1)!r!} \\
&= \frac{r(n-1)! + (n-1)!(n-r)}{(n-r)!r!} \\
&= \frac{(n-1)!(n-r+r)}{(n-r)!r!} \\
&= \frac{n!}{(n-r)!r!} \\
&= \binom{n}{r}
\end{aligned}$$

■

Theorem 1.1 (Binomial Theorem)**.** For $x, y \in \mathbb{R}$ and $n \in \mathbb{Z}^+$,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k \tag{1.11}$$

Proof. The most straightforward proof is via mathematical induction. A combinatorial approach is as follows,

$$(x+y)^n = (x+y)(x+y)\cdots(x+y)$$

To get a certain $x^{n-k}y^k$, we need to choose k such $(x+y)$ to form y^k. Out of the n available $(x+y)$, we need to pick k. Thus, there are $\binom{n}{k}$ ways to choose $(x+y)$. For example, to form y^n we need to choose all n available $(x+y)$ so there is $\binom{n}{n} = 1$ way to choose the $(x+y)$ terms. We sum over all of these to get the binomial theorem. ■

1.4 Multinomial Coefficients

Consider a set of n distinct items to be divided into r distinct groups where the first group has n_1 items, the second has n_2, ...m the r^{th} group has n_r items $(n_1 + \ldots n_r = n)$. Then, by the generalised basic principle of counting, the number of different divisions is

$$\binom{n}{n_1} \times \binom{n-n_1}{n_2} \cdots \times \binom{n-n_1-\cdots-n_{r-1}}{n_r} \tag{1.12}$$

We define this to be the multinomial coefficient

$$\binom{n}{n_1, n_2, \cdots, n_r} = \frac{n!}{\prod_{i=1}^{r} n_i!} = \frac{n!}{n_1! n_2! \cdots n_r!} \tag{1.13}$$

These are known are multinomial coefficients as they come from the multinomial theorem

Theorem 1.2 (Multinomial Theorem)**.** If $x_1, \cdots, x_r \in \mathbb{R}$ and $n \in \mathbb{Z}^+$, then

$$(x_1 + \cdots x_r)^n = \sum_{(n_1, \cdots n_r)} \binom{n}{n_1, \cdots, n_r} x_1^{n_1} x_2^{n_2} \dots x_r^{n_r} \tag{1.14}$$

where $(n_1, \dots n_r)$ represents the enumerations over all possible $n_1, \dots n_r$ such that $\sum_{i=1}^{r} n_i = n$

Proof. We apply a similar proof to the combinatorial proof for the binomial theorem. There must be $\binom{n}{n_1, \dots, n_r}$ ways to choose $(x_1 + \cdots + x_r)$ such that we get a given power of $x_1, \dots, x_r$ ■

Example 7

Compute $(x + y + z)^3$

$$(x+y+z)^3 = \binom{3}{3,0,0}x^3 + \binom{3}{0,3,0}y^3 + \binom{3}{0,0,3}z^3 + \binom{3}{2,1,1}x^2yz \tag{1.15}$$

$$+ \binom{3}{1,2,1}xy^2z + \binom{3}{1,1,2}xyz^2 + \binom{3}{2,1,0}x^2y + \binom{3}{1,2,0}xy^2z \tag{1.16}$$

$$+ \binom{3}{1,0,2}xz^2 + \binom{3}{0,1,2}yz^2 + \binom{3}{0,2,1}y^2z + \binom{3}{2,0,1}x^2z \tag{1.17}$$

$$+ \binom{3}{1,1,1}xyz \tag{1.18}$$

2 Axioms of Probability

2.1 Sample Space and Events

We need to first define what a random experiment is.

Definition 2.1 (Random Experiment)**.** A random experiment is an experiment whose outcome is not deterministic.

For example, when we toss a coin, we do not always get either a heads or a tails. While, the physics of the coin can be used to predict the outcome (knowing its mass, air resistance etc.), the outcome is too hard for us to predict with the information we have. Consider a random experiment whose outcome is not predictable, but whose set of all possible outcomes is known (such as tossing a coin or throwing a die).

Definition 2.2 (Sample Space)**.** The sample space S of an experiment is the set of all possible outcomes of the experiment

If our experiment is flipping a coin, then the sample space is

$$S_{coin} = \{H, T\} \tag{2.1}$$

If our experiment is rolling a 6-sided dice, then the sample space is

$$S_{dice} = \{1, 2, 3, 4, 5, 6\} \tag{2.2}$$

If our experiment is tossing two coins in succession, then the sample space is

$$S_{coins} = \{(H, H), (H, T), (T, H), (T, T)\} \tag{2.3}$$

If our experiment is tossing rolling two 6-sided die, then the sample space is

$$S_{die} = S_{die} \times S_{die} = \{(a, b)|a, b \in S_{die}\} \tag{2.4}$$

Definition 2.3 (Event)**.** An event E is a set of possible outcomes of an experiment.It is a subset of the sample space.

$$E \subset S \tag{2.5}$$

If an outcome of an experiment is contained in E, we say event E has occurred.

If our event is rolling an even number on a 6-sided dice, then

$$E = \{2, 4, 6\} \tag{2.6}$$

If we toss two coins and our event is getting the same face value on both tosses, then

$$E = \{(H, H), (T, T)\} \tag{2.7}$$

Definition 2.4 (Union)**.** For two events $E, F \subset S$, the union is a new event that contains all the outcomes either in E or F or both. It is denoted by,

$$E \cup F \tag{2.8}$$

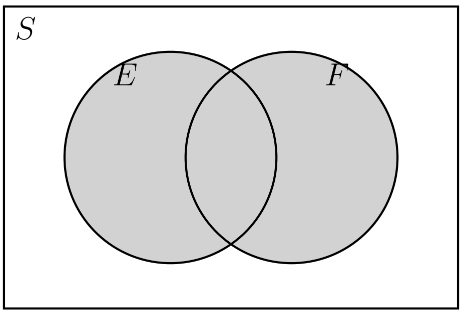


Figure 1: $E \cup F$

If we toss two coins and E is the event the first coin lands heads, $E = \{(H, H), (H, T)\}$ and F is the event the second coin lands heads, $F = \{(H, H), (T, H)\}$, then

$$E \cup F = \{(H, H), (H, T), (T, H)\} \tag{2.9}$$

is the event that at least one of the tossed coins is a head.

Definition 2.5 (Intersection)**.** For two events $E, F \subset S$, the intersection is a new event that contains all the outcomes that are in both E and F. It is denoted by,

$$E \cap F \tag{2.10}$$

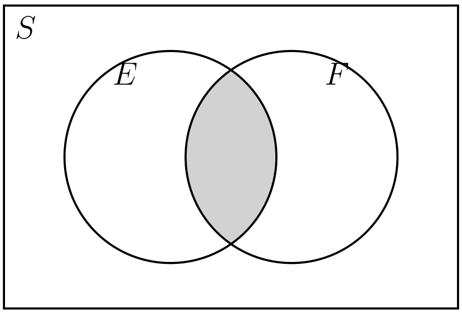


Figure 2: $E \cap F$

If we roll a die and $E = \{1, 3, 5\}$ where the roll is an odd number and $F = \{3, 6\}$ where the roll is a multiple of 3 then,

$$E \cap F = \{3\} \tag{2.11}$$

which is the event the roll is an odd multiple of 3.

Definition 2.6 (Null event)**.** The null event $\varnothing$ is the event where no outcome occurs.

If $E = \{1, 3, 5\}$ (the roll is an odd number) and $F = \{2, 4, 6\}$ (the roll is an even number), then

$$E \cap F = \varnothing \tag{2.12}$$

because it is not possible to roll an even and odd number at the same time.

Definition 2.7 (Mutually Exclusive)**.** Two events $E, F \subset S$ are mutually exclusive if

$$E \cap F = \varnothing \tag{2.13}$$

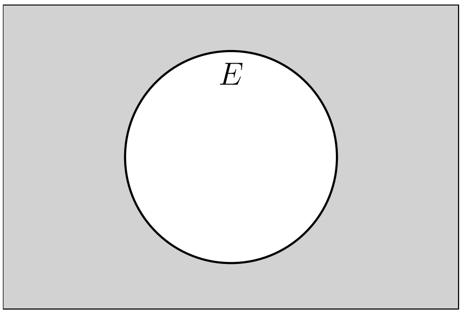


Figure 3: E^c

Definition 2.8 (Complement)**.** The complement of an event $E \subset S$ is the set of outcomes that are not in E, it is denoted by

$$E^c \tag{2.14}$$

It holds that,

$$(E^c)^c = E \tag{2.15}$$

Since an experiment must result in some outcome, we have,

$$S^c = \varnothing \tag{2.16}$$

If we have multiple events $E_1, E_2, \cdots \subset S$, then their union is that event that consists of all outcomes in E_n for at least one $n = 1, 2, \cdots$

$$\bigcup_{i=1}^{\infty} E_i = \{x \in S | \exists i \geq 1 \text{ s.t } x \in E_i\} \tag{2.17}$$

Their intersection is the event that consists of the outcomes that are in all E_n for $n = 1, 2, \cdots$

$$\bigcap_{i=1}^{n} E_i = \{x \in S | \forall i \geq 1, \ x \in E_i\} \tag{2.18}$$

If all the outcomes of E are in F, we say E is contained in F ie $E \subset F$.

Thus, we say two events are equal if

$$E \subset F \quad \text{and} \quad F \subset E \tag{2.19}$$

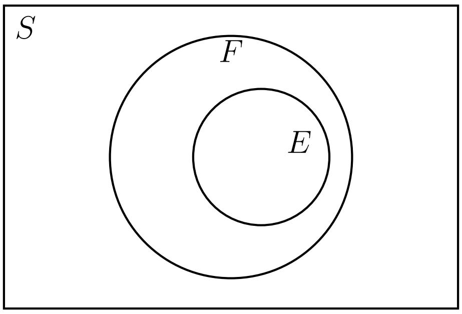


Figure 4: $E \subset F$

Events follow the following rules,

1. Commutative Laws

$$E \cup F = F \cup E \tag{2.20}$$

$$E \cap F = F \cap E \tag{2.21}$$

2. Associative Laws

$$(E \cup F) \cup G = E \cup (F \cup G) \tag{2.22}$$

$$(E \cap F) \cap G = E \cap (F \cap G) \tag{2.23}$$

3. Distributive Laws

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G) \tag{2.24}$$

$$(E \cap F) \cup G = (E \cap G) \cap (F \cap F) \tag{2.25}$$

The laws can be verified via Venn diagrams. We also have DeMorgan's laws

Theorem 2.1 (DeMorgan's Laws)**.** For events $E_1, \cdots, E_n \subset S$,

$$\left(\bigcup_{i=1}^{n} E_i\right)^c = \bigcap_{i=1}^{n} E_i^c \tag{2.26}$$

$$\left(\bigcap_{i=1}^{n} E_i\right)^c = \bigcup_{i=1}^{n} E_i^c \tag{2.27}$$

Proof. Let $x \in (\bigcup_{i=1}^{n} E_i)^c$, then $x \notin (\bigcup_{i=1}^{n} E_i)$, which means x is not in any of the events E_1 for $i = 1, \cdots, n$. Thus, x is contained in E_i^c for all $i = 1, \cdots, n$. For the other way, assume $x \in \bigcap_{i=1}^{n} E_i^c$, then x is contained in E_i^c for all $i = 1, \cdots, n$, which means $x \notin \bigcup_{i=1}^{n} E_i$ which means $x \in (\bigcup_{i=1}^{n} E_i)^c$.

To prove the second law, we use the first law,

$$\bigcap_{i=1}^{n} (E_i^c)^c = \bigcup_{i=1}^{n} E_i^c$$

$$\implies \bigcap_{i=1}^{n} E_i = \left(\bigcup_{i=1}^{n} E_i^c\right)^c$$

Taking complements gives,

$$\left(\bigcap_{i=1}^{n} E_i\right)^c = \bigcup_{i=1}^{n} E_i^c$$

■

For two events, DeMorgan's laws says that

$$(E \cup F)^c = E^c \cap F^c \tag{2.28}$$

$$(E \cap F)^c = E^c \cup F^c \tag{2.29}$$

2.2 Axioms of Probability

Definition 2.9 (Probability of an event)**.** For an event $E \subset S$, the probability of E

occuring is the chance that E occurs. It is denoted by,

$$\mathbb{P}(E) \tag{2.30}$$

We used to intuitively define probability via relative frequency. Perform an experiment n times. If $n(E)$ is the number of times E occurs then,

$$\mathbb{P}(E) = \lim_{n \to \infty} \frac{n(E)}{n} \tag{2.31}$$

That is, the probability is the the relative frequency of E as we perform our experiment infinity many times. However, this assumes that $\frac{n(E)}{n}$ converges, and that it converges to the same value each time we repeat our experiment infinitely many times.

The modern axiomatic approach instead is to start with more evident assumptions and show that we the relative frequency definition is naturally satisfied. In essence, for each event $E \subset S$, we assign a number $\mathbb{P}(E)$ which is the probability of E occuring. This number satisfies a set of axioms which agrees with our intuition of the relative frequency definition.

Definition 2.10 (Axioms of Probability)**.** Consider an experiment whose sample space is S. For each event $E \subset S$, we assign a number $\mathbb{P}(E)$ that satisfies the following axioms

1. Axiom 1
$$0 \leq \mathbb{P}(E) \leq 1$$

2. Axiom 2
$$\mathbb{P}(S) = 1$$

3. Axiom 3: For mutually exclusive events $E_1, E_2 \cdots (E_i \cap E_j = \varnothing$ when $i \neq j)$
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

$\mathbb{P}(E)$ is the probability of event E

Axiom 1 puts a bound on the probability of an event. Axiom 2 says that if we perform a random experiment with sample S, the probability that the outcome is in S is 1. Axiom 3 says that for mutually exclusive events, the probability that at least one event occurs is the sum of their respective probabilities.

The consequences of the axioms of probability are as follows,

Theorem 2.2 (Probability of a null event)**.**

$$\mathbb{P}(\varnothing) = 0$$

Proof. Define $E_1 = S$ and $E_i = \varnothing$ for $i > 1$. Since $S^c = \varnothing$, S and $\varnothing$ are mutually exclusive. Thus,

$$\mathbb{P}(S) = \sum_{i=1}^{n} \mathbb{P}(E_i) = P(E_1) + \sum_{i=2}^{n} \mathbb{P}(E_i)$$

Since $E_1 = S$ and $E_i = \varnothing$ for $i > 1$,

$$\mathbb{P}(S) = \mathbb{P}(S) + \sum_{i=2}^{\infty} \mathbb{P}(\varnothing)$$

Thus,

$$\mathbb{P}(\varnothing) = 0 \tag{2.32}$$

■

Theorem 2.3 (Probability of finite mutually exclusive events)**.** Let $E_1, \cdots, E_n$ be mutually exclusive events. Then,

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

Proof. Let $E_{n+1} =, \cdots = \varnothing$. Then,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \mathbb{P}(E_1 \cup E_2 \cup \cdots E_n \cup \varnothing \cdots) \tag{2.33}$$

$$= \mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) \tag{2.34}$$

$$= \sum_{i=1}^{n} \mathbb{P}(E_i) + \sum_{i=n+1}^{\infty} \mathbb{P}(\varnothing) \tag{2.35}$$

$$= \sum_{i=1}^{n} \mathbb{P}(E_i) \tag{2.36}$$

where the last equality comes from theorem 2.2. ■

For two mutually exclusive events A and B we have,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \tag{2.37}$$

Theorem 2.4 (Probability of a complement)**.**

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

Proof. We know that E and E^c are mutually exclusive events with $E \cup E^c = S$. Then by axiom 3,

$$\mathbb{P}(S) = \mathbb{P}(E) + \mathbb{P}(E^c)$$
$$\implies \mathbb{P}(E^c) = \mathbb{P}(S) - \mathbb{P}(E)$$
$$\implies \mathbb{P}(E^c) = 1 - \mathbb{P}(E)$$

■

Theorem 2.5 (Monotonicity Principle)**.** For two events $E, F \subset S$, if $E \subset F$ then

$$\mathbb{P}(E) \leq \mathbb{P}(F)$$

Proof. We see that E and $F \cap E^c$ are mutually exclusive,

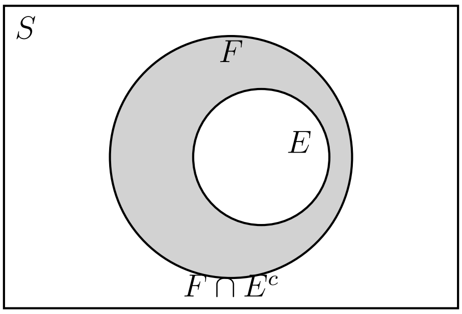


Figure 5: $F \cap E^c$

Also,

$$\begin{aligned} E \cup (F \cap E^c) &= (E \cup F) \cap (E \cup E^c) \\ &= F \cap S \\ &= F \end{aligned}$$

Thus,

$$\begin{aligned} &\mathbb{P}(F) = \mathbb{P}(E) + \mathbb{P}(F \cap E^c) \text{ (By theorem 2.3)} \\ \implies &\mathbb{P}(F) \geq \mathbb{P}(E) \end{aligned}$$

■

Theorem 2.6. For any two events $E, F \subset S$,

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$$

Proof. Note that we can decompose $E \cup F$ as two mutual events,

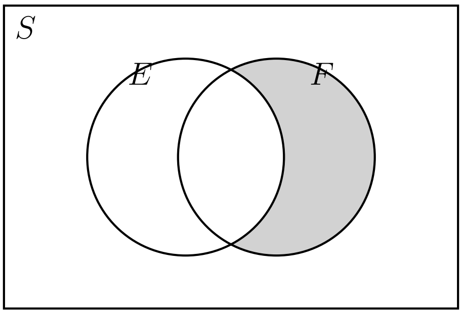


Figure 6: $E^c \cap F$

$$E \cup F = E \cup (E^c \cap F)$$

Then by axiom 3,

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F \cap E^c)$$

Since $F = (E \cap F) \cup (E^c \cap F)$,

$$\mathbb{P}(F) = \mathbb{P}(E \cap F) + \mathbb{P}(E^c \cap F)$$

Thus,

$$\mathbb{P}(E^c \cap F) = \mathbb{P}(F) - \mathbb{P}(E \cap F)$$

Substitute this into the equation for $\mathbb{P}(E \cup F)$ to get,

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F) \tag{2.38}$$

■

This can be generalised to n events using the inclusion-exclusion principle,

Theorem 2.7 (Inclusion-Exclusion principle)**.** For $E_1, \cdots, E_n \subset S$,

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{1 \leq i < j \leq n} \mathbb{P}(E_i \cap E_j) + \sum_{1 \leq i < j < k \leq n} \mathbb{P}(E_i \cap E_j \cap E_k) - \cdots + (-1)^{n-1} \mathbb{P}(E_1 \cap \cdots \cap E_n)$$

Proof. We use mathematical induction where $n \geq 1$. For our base case $n = 1$, we have trivially $\mathbb{P}(E) = \mathbb{P}(E)$. We proved the proposition for $n = 2$ above in theorem 2.6. Then, assume our proposition is true for $n = k$ where $k \in \mathbb{Z}^+$ ie

$$P\left(\bigcup_{i=1}^{k} E_i\right) = \sum_{i=1}^{k} \mathbb{P}(E_i) - \sum_{1\leq i<j\leq k} \mathbb{P}(E_i \cap E_j) + \sum_{1\leq i<j<p\leq k} \mathbb{P}(E_i \cap E_j \cap E_p) - \cdots + (-1)^{k-1}\mathbb{P}(E_1 \cap \cdots \cap E_k)$$

Define F to be the proposition for $n = k$. Then for $n = k + 1$ we have,

$$\begin{aligned} P(F \cup E_{k+1}) &= \mathbb{P}(F) + \mathbb{P}(E_{k+1}) - \mathbb{P}(F \cap E_{k+1}) \\ &= \sum_{i=1}^{k+1} \mathbb{P}(E_i) - \sum_{1\leq i<j\leq k+1} \mathbb{P}(E_i \cap E_j) + \sum_{1\leq i,j,p\leq k+1} \mathbb{P}(E_i \cap E_j \cap E_p) - \cdots + (-1)^k \mathbb{P}(E_1 \cap \cdots \end{aligned}$$

■

For 3 events we have,

$$\mathbb{P}(E \cup F \cup G) = \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(E \cap F) - \mathbb{P}(E \cap G) - \mathbb{P}(F \cap G) + \mathbb{P}(E \cap F \cap G) \tag{2.39}$$

2.3 Uniform Probability Space

Definition 2.11 (Uniform Probability Space)**.** A uniform probability space is a sample space where all the outcomes of an experiment are equally likely. For a sample space of n outcomes,

$$S = \{A_1, A_2, \cdots, A_n\}$$

,

$$\mathbb{P}(A_i) = \frac{1}{n}, \forall i$$

Examples of uniform probability spaces include tossing a fair coin and rolling a fair dice. We can say that for any event E,

$$\mathbb{P}(E) = \frac{\text{Number of outcomes in E}}{\text{Number of outcomes in S}} \tag{2.40}$$

We can find the number of outcomes using combinatorics.

Example 8

A committee of 5 people is to be selected from a group of of 6 men and 9 women. If the selection is made randomly, find the probability that the committee is made of 3 men and 2 women.

Solution

There are $\binom{15}{5}$ ways to make a 5 person committee. There are $\binom{6}{3}\binom{9}{2}$ ways to make a committee of 3 men and 2 women. Thus,

$$\frac{\binom{6}{3}\binom{9}{2}}{\binom{15}{5}} = \frac{240}{1001}$$

3 Conditional Probability

Often times, we adjust the probability of events occurring based on information we have.

Example 9

Suppose we have two events: E and F when we roll two fair six-sided die. E is the event we roll a 5 once and F is the event our sum is 2. If E did not occur, then $\mathbb{P}(F) = \frac{1}{36}$. If E did occur, then $\mathbb{P}(F) = 0$ since it is impossible to get a sum of 2 if we roll a 5.

In the example above, we had information that we rolled a 5. Thus, we updated our probability of F occurring from $\frac{1}{36}$ to 0. This is known as conditional probability.

Definition 3.1 (Conditional Probability)**.** If we have two events $E, F \subset S$. Then, the probability that event F occurs, given that event E has already occurred is denoted by,

$$\mathbb{P}(F|E) \tag{3.1}$$

If $\mathbb{P}(F) > 0$, then

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)} \tag{3.2}$$

The idea behind conditional probability is that if event F occurs after event E occurs, then the occurrence must be in both E and F. Further, since E has already occurred, our sample space reduces to just the outcomes in E. If all the outcomes are likely, we have,

$$\begin{aligned}\mathbb{P}(F|E) &= \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)} \\ &= \frac{\frac{|F \cap E|}{|S|}}{\frac{|E|}{|S|}} \\ \mathbb{P}(F|E) &= \frac{|F \cap E|}{|E|} \end{aligned} \tag{3.3}$$

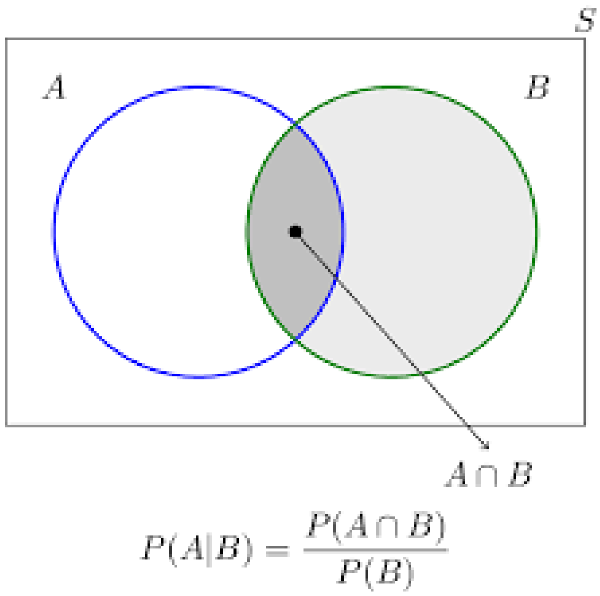


Figure 7: Conditional Probability

Example 10

We see that the sample space reduces to B. A coin is flipped twice. Assuming that all outcomes are equally likely $S = \{HH, TH, HT, TT\}$, what is the probability that both flips on a head given that the first flip is heads.

Solution

Let E = first flip is heads and F = both flips land on a head. Then, $\mathbb{P}(F|E) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)}$ We know $\mathbb{P}(E) = 0.5$ and $F \cap E = \{HH\}$ so $\mathbb{P}(F \cap E) = 0.25$. Then, $\mathbb{P}(F|E) = 0.5$

From the definition of conditional probability we have

$$\mathbb{P}(F \cap E) = \mathbb{P}(E|F)\mathbb{P}(F) \tag{3.4}$$

We can use conditional probability to give us the general form of the intersection of many events via the multiplication rule.

Lemma 3.1 (Multiplication Rule)**.** For events $E_1 \cdots E_n \subset S$ then,

$$\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right) = \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1 \cap E_2)\cdots\mathbb{P}(E_n|E_1 \cap \cdots \cap E_{n-1}) \tag{3.5}$$

Proof. Applying the definition of conditional probability we have,

$$\mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1 \cap E_2)\cdots\mathbb{P}(E_n|E_1 \cap \cdots \cap E_{n-1}) \tag{3.6}$$

$$= \mathbb{P}(E_1)\frac{\mathbb{P}(E_2 \cap E_1)}{\mathbb{P}(E_1)}\frac{\mathbb{P}(E_1 \cap E_2 \cap E_3}{\mathbb{P}(E_1 \cap E_2)}\cdots\frac{\mathbb{P}(E_1 \cap \cdots \cap E_{n-1})}{\mathbb{P}(E_1 \cap \cdots \cap E_{n-2})}\frac{\mathbb{P}(E_1 \cap \cdots \cap E_n)}{\mathbb{P}(E_1 \cap \cdots \cap E_{n-1})} \tag{3.7}$$

$$= \mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right) \tag{3.8}$$

■

Example 11

Consider a box with 8 red balls and 4 white balls. Suppose two balls are picked without replacement. Find the probability both balls are red.

Solution

Let B_1 = the first ball picked is red and B_2 = the second ball picked is red. We have

$$\mathbb{P}(B_1 \cap B_2) = \mathbb{P}(B_1)\mathbb{P}(B_2|B_1) \tag{3.9}$$

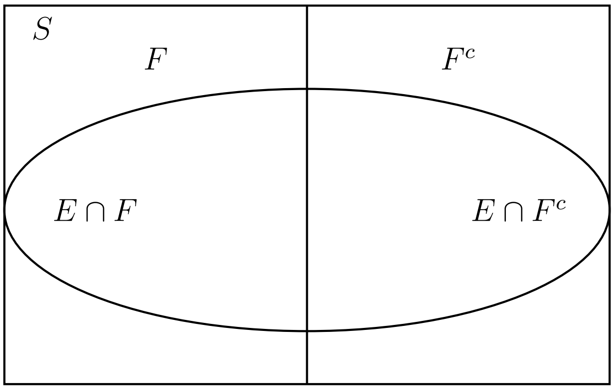
$$= \frac{8}{12}\cdot\frac{7}{11} \tag{3.10}$$

$$= \frac{56}{132} \tag{3.11}$$

Consider an event F that partitions our sample space, and an event E that is made up by intersections along the partition.

$$\mathbb{P}(E) = \mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c) \tag{3.12}$$

$$= \mathbb{P}(F)\mathbb{P}(E|F) + \mathbb{P}(F^c)\mathbb{P}(E|F^c) \tag{3.13}$$



$$P(E) = P(E \cap F) + P(E \cap F^c)$$

Figure 8: Partition of the sample space by F and decomposition of E into $E \cap F$ and $E \cap F^c$.

We can generalise this for n events via the law of total probability,

Lemma 3.2 (Law of Total Probability)**.** Suppose $F_1, \dots, F_n \subset S$ are mutually exclusive events such that $\bigcup_{i=1}^{n} F_i = S$. Then,

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(F_i)\mathbb{P}(E|F_i) \tag{3.14}$$

Proof. Use induction on n. ■

Example 12

Three machines make parts at a factory. We know the following about the manufacturing process:

1. Machine 1 makes 60% of the parts
2. Machine 2 makes 30% of the parts
3. Machine 3 makes 10% of the parts
4. Of the parts machine 1 makes, 7% are defective
5. Of the parts machine 2 makes, 15% are defective
6. Of the parts machine 3 makes, 30% are defective

A part is randomly selected. What is the probability it is defective

Solution

Let D be the event a randomly selected part is defective. Then, by the law of total probability

$$\begin{aligned}\mathbb{P}(D) &= \mathbb{P}(M_1)\mathbb{P}(D|M_1) + \mathbb{P}(M_2)\mathbb{P}(D|M_2) + \mathbb{P}(M_3)\mathbb{P}(D|M_3) && (3.15)\\ &= 0.1 \cdot 0.3 + 0.3 \cdot 0.15 + 0.10 \cdot 0.30 && (3.16)\\ &= 0.117 && (3.17)\end{aligned}$$

Example 13

There are four urns. The first urn contains 4 blue balls and 6 red balls. The second urn contains 1 blue ball and 6 red balls. The third urn contains 3 blue balls and 5 red balls. The fourth urn contains 5 blue balls and 4 red balls. An urn is randomly selected and a ball is picked from the urn. What is the probability the ball is blue?

Solution

Let B be the event a blue ball is picked and U_i be the event urn i is picked where $i = 1, 2, 3, 4$. Then,

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(U_1)\mathbb{P}(B|U_1) + \mathbb{P}(U_2)\mathbb{P}(B|U_2) + \mathbb{P}(U_3)\mathbb{P}(B|U_3) + \mathbb{P}(U_4)\mathbb{P}(B|U_4) && (3.18)\\ &= \frac{1}{4}\left(\frac{4}{10} + \frac{1}{7} + \frac{3}{8} + \frac{5}{9}\right) && (3.19)\\ &\approx 0.37 && (3.20)\end{aligned}$$

The law of total probability is important because it lends itself to Bayes' theorem.

3.1 Bayes' Theorem

We create a new test for cancer. With this test, we test 1000 people for cancer. We have two groups: 500 people actually have cancer while another 500 people form a control group without cancer. These are the test results,

1. 500 People with cancer: 450 people test positive, 50 people test negative (false negative)

2. 500 people without cancer: 100 people test positive (false positive) and 400 people test negative

Let B be the event a person has cancer and T be the event they test positive. From the table, we can compute the probability that they test positive given they have cancer $\mathbb{P}(T|B) = \frac{450}{500}$. However, it is often better for the creators of the test to know what the probability of actually having cancer is, given that you test positive for it. Namely, it is better to know $\mathbb{P}(B|T)$. Thus, we now need to be able to relate $\mathbb{P}(B|T)$ to $\mathbb{P}(T|B)$. This relation comes from Bayes' Theorem

Theorem 3.3 (Bayes' Theorem)**.** For two events $A, B \subset S$

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \tag{3.21}$$

Proof. We have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{3.22}$$

Thus,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) \tag{3.23}$$

Also

$$\mathbb{P}(B \cap A) = \mathbb{P}(B|A)\mathbb{P}(A) \tag{3.24}$$

Since $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$, equating the two gives

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \tag{3.25}$$

■

$\mathbb{P}(B)$ is known as the prior probability. Given that we have no information about the test, $\mathbb{P}(B)$ gives the probability of a person having a cancer. $\mathbb{P}(A|B)$ is known as the update probability. $\mathbb{P}(A|B)$ updates the probability based on knowledge of the test. $\mathbb{P}(A)$ is a normalisation constant and $\mathbb{P}(B|A)$ is known as the posterior probability. We can generalise Bayes' theorem to n events using the law of total probability.

Theorem 3.4 (General Bayes' Theorem)**.** Let $A, B_1, \ldots, B_n \subset S$. Then,

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A|B_i)} \tag{3.26}$$

Proof. Finish ■

For the two event case, this simplifies to

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(B)\mathbb{P}(A|B) + \mathbb{P}(B^c)\mathbb{P}(A|B^c)} \tag{3.27}$$

Example 14

We have a test that is 99% accurate in detecting cancer, however the cancer is very rare so only 0.001% of the population have this cancer. What is the probability of having the cancer given a positive test.

Solution

Let A be the event an individual tests positive for cancer and B be the event they have cancer. We have

$$\mathbb{P}(A) = 0.00001 \tag{3.28}$$

$$\mathbb{P}(A|B) = 0.99 \tag{3.29}$$

We want $\mathbb{P}(B|A)$. Using Bayes' theorem,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(B)\mathbb{P}(A|B) + \mathbb{P}(B^c)\mathbb{P}(A|B^c)} \tag{3.30}$$

Substituting,

3.2 Independence

Two experiments are independent if the outcome of one experiment does not affect the outcome of another experiment. For example, if you toss a coin twice. Whether you get heads or tails on the second toss does not depend on whether you got a heads or tails on the first toss. We can formally define independence.

Theorem 3.5. For $E, F \subset S$, E and F are independent if

$$\mathbb{P}(E|F) = \mathbb{P}(E) \tag{3.31}$$

$$\mathbb{P}(F|E) = \mathbb{P}(F) \tag{3.32}$$

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F) \tag{3.33}$$

Example 15

Suppose you toss two die. Let E_1 be the event that the sum of the values is a 6, let E_2 be the event that sum of the values is a 7, and let F be the event that the first dice is a 4. Are E_1 and F independent, and are E_2 and F independent.

Solution

$\mathbb{E}_1 \cap F$ is the event that the first die is a 4 and the sum of the values is a 6. Thus, the second die must be a 2 and so $\mathbb{P}(E_1 \cap F) = \frac{1}{36}$. Now the outcomes of E_1 and F are

$$E_1 = \{(1,5),(2,4),(3,3),(4,2),(5,1)\} \tag{3.34}$$

$$F = \{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\} \tag{3.35}$$

Then,

$$\mathbb{P}(E_1) = \frac{5}{36} \mathbb{P}(F) = \frac{6}{36} = \frac{1}{6} \tag{3.36}$$

Since $\mathbb{P}(E_1)\mathbb{P}(F) \neq \mathbb{P}(E_1 \cap F)$, the two events are not independent. Now consider E_2. $E_2 \cap F$ is the event that the sum of the die is a 7 and the first die is a 4. Then the only valid outcome is (4, 3).

$$\mathbb{P}(E_2 \cap F) = \frac{1}{36} \tag{3.37}$$

Considering just the outcomes in E_2 we have,

$$E_2 = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \tag{3.38}$$

Then,

$$\mathbb{P}(E_2) = \frac{6}{36} = \frac{1}{6} \tag{3.39}$$

Since $\mathbb{P}(E_2)\mathbb{P}(F) = \mathbb{P}(E_2 \cap F)$, these two events are independent.

Defining the independence of 3 events is more involved.

Definition 3.2 (Independence of 3 events)**.** For $E, F, G \subset S$, E, F, G are independent if

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F) \tag{3.40}$$

$$\mathbb{P}(E \cap G) = \mathbb{P}(E)\mathbb{P}(G) \tag{3.41}$$

$$\mathbb{P}(F \cap G) = \mathbb{P}(F)\mathbb{P}(G) \tag{3.42}$$

$$\mathbb{P}(E \cap F \cap G) = \mathbb{P}(E)\mathbb{P}(F)\mathbb{P}(G) \tag{3.43}$$

If the first three hold, we say the events are pairwise independent. However, pairwise independence does not imply independence since we need the last statement to hold. We can generalise the statement for n events.

Definition 3.3 (Independence of n events)**.** For n evenets $E_1, \ldots E_n \subset S$, these events are independent if for every finite subset of Ω of $\{E_1, \ldots E_n\}$, we have

$$\mathbb{P}\left(\bigcap_{i\in\Omega} E_i\right) = \prod_{i\in\Omega} \mathbb{P}(E_i) \tag{3.44}$$

Thus four events E, F, G, H are independent only if all of the following are satisfied

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F) \tag{3.45}$$

$$\mathbb{P}(E \cap G) = \mathbb{P}(E)\mathbb{P}(G) \tag{3.46}$$

$$\mathbb{P}(E \cap H) = \mathbb{P}(E)\mathbb{P}(H) \tag{3.47}$$

$$\mathbb{P}(F \cap G) = \mathbb{P}(F)\mathbb{P}(G) \tag{3.48}$$

$$\mathbb{P}(F \cap H) = \mathbb{P}(F)\mathbb{P}(H) \tag{3.49}$$

$$\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}(H) \tag{3.50}$$

$$\mathbb{P}(E \cap F \cap G) = \mathbb{P}(E)\mathbb{P}(F)\mathbb{P}(G) \tag{3.51}$$

$$\mathbb{P}(E \cap F \cap H) = \mathbb{P}(E)\mathbb{P}(F)\mathbb{P}(H) \tag{3.52}$$

$$\mathbb{P}(E \cap G \cap H) = \mathbb{P}(E)\mathbb{P}(G)\mathbb{P}(H) \tag{3.53}$$

$$\mathbb{P}(F \cap G \cap H) = \mathbb{P}(F)\mathbb{P}(G)\mathbb{P}(H) \tag{3.54}$$

$$\mathbb{P}(E \cap F \cap G \cap H) = \mathbb{P}(E)\mathbb{P}(F)\mathbb{P}(G)\mathbb{P}(H) \tag{3.55}$$

We also have a notion of conditional independence.

Definition 3.4 (Conditional Independence)**.** Events F and G are conditionally independent given an event H if

$$\mathbb{P}(F \cap G|H) = \mathbb{P}(F|H)\mathbb{P}(G|H) = \mathbb{P}(F|G \cap H)\mathbb{P}(G|F \cap H) \tag{3.56}$$

We can generalise this for n events,

Definition 3.5 (Conditional Independence of n events)**.** A collection of events $E_1 \dots E_n$ are independent given an event F if for any finite subset Ω of $\{E_1 \dots E_n\}$

$$\mathbb{P}\left(\left(\bigcap_{i \in \Omega} E_i\right) |F\right) = \prod_{i \in \Omega} \mathbb{P}(E_i|F) \tag{3.57}$$

4 Discrete Random Variables

4.1 Introduction to Random Variables

Whenever we perform an experiment, we often care about quantifying some outcome as opposed to the actual outcome itself. For example, if we roll two die, we may care more about the sum of the face of the die as opposed to the actual values on each dice. Alternatively, if we toss a coin, we may care about the number of heads we get as opposed to the actual sequence of heads and tails. For this, we need random variables.

Definition 4.1 (Random Variables)**.** A random variable X is a real-valued function from the sample space to the real numbers, $X : S \to \mathbb{R}$. That is, a random variable associates to each outcome in the sample space S a real number

We tend to characterise a random variable by the values that it can take, and the probability of these values. If the random variable takes on the value k with probability p, we denote

$$\mathbb{P}(X = k) = p \tag{4.1}$$

Example 16

If we toss 3 fair coins, we can define our random variable to be the number of heads we get. Then, X can take values $0, 1, 2, 3$. The outcomes of our experiment for each random value is,

$$X = 0 : \{TTT\}$$
$$X = 1 : \{HTT, THT, TTH\}$$
$$X = 2 : \{HHT, HTH, THH\}$$
$$X = 3 : \{HHH\}$$

Then,

$$\mathbb{P}(X=0)=\frac{1}{8}$$
$$\mathbb{P}(X=1)=\frac{3}{8}$$
$$\mathbb{P}(X=2)=\frac{3}{8}$$
$$\mathbb{P}(X=3)=\frac{1}{8}$$

4.2 Probability Mass Functions and Support

Definition 4.2 (Discrete Random Variable)**.** A random variable X is discrete if it can take at most countably many values.

For discrete random variables, we define an associated function called the probability mass function.

Definition 4.3 (Probability Mass Function)**.** For a discrete random variable X, the associated probability mass function $f_X : \mathbb{R} \to [0,1]$ is defined by,

$$f_X(x) = \mathbb{P}(X = x) \;\; \forall x \in \mathbb{R} \tag{4.2}$$

The probability mass function maps each value $x \in \mathbb{R}$ to the probability of the random variable taking on that value. Thus, we say it is the probability distribution of a discrete random variable; it provides the values the random variable can take and its associated probabilities. If X cannot take on the value x then $f_X(x) = 0$. A function f is a valid probability mass function of a discrete random variable X if To this end, we define the support of a random variable.

Definition 4.4 (Support)**.** The support of a random variable X is a subset of $\mathbb{R}$

which contains the set of values the random variable can take,

$$\mathrm{supp}(X) = \{x \in \mathbb{R} \mid f_X(x) > 0\} \tag{4.3}$$

Since the sum of all the probabilities must add to 1

$$\sum_{x \in \mathrm{supp}(X)} f_X(x) = 1 \tag{4.4}$$

In fact, a given function is a valid probability mass function of a discrete random variable X if and only if

- $f_X(x) \geq 0 \forall x \in \mathrm{supp}(X)$
- $\sum_{x \in \mathrm{supp}(X)} f_X(k) = 1$

When we graph the probability mass function, we tend to get histograms. Suppose we roll two die and compute their sum. The associated probability mass function is

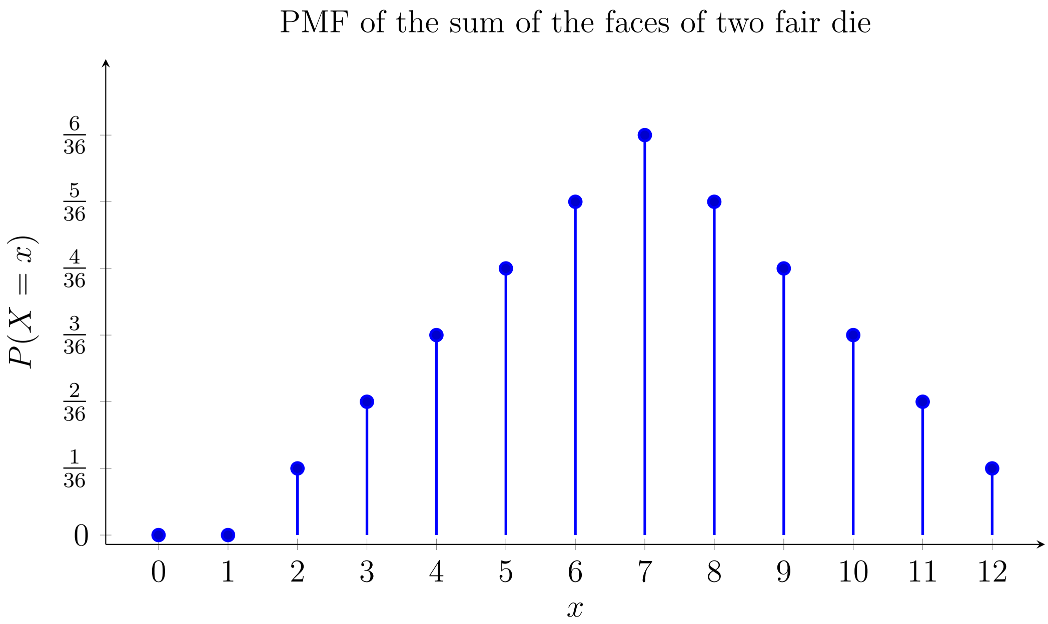


Figure 9: Probability mass function of the sum of two fair six-sided dice.

4.2.1 Expectation

Definition 4.5 (Expectation of a Discrete Random Variable)**.** If X is a discrete random variable with associated probability mass function f_X, then the expectation of X is,

$$\mathbb{E}(X) = \sum_{x \in \text{supp}(X)} x f_X(x) = \sum_{x \in \text{supp}(X)} x \mathbb{P}(X = x) \tag{4.5}$$

The expectation is the weighted average of all possible outcomes. With the frequency interpretation of probably, it can be thought of as the long run average value that X takes on when we repeat our experiment n times.

Proof. Let $X^{(i)}$ be the value of X in the i^{th} repetion. Then,

$$\begin{aligned}
\mathbb{E}(X) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X^{(i)} && (4.6)\\
&= \lim_{n\to\infty} \sum_{x \in \text{supp}(X)} x \frac{\text{frequency of } x \text{ in } n \text{ repeated experiments}}{n} && (4.7)\\
&= \sum_{x \in \text{supp}(X)} x \lim_{n\to\infty} \frac{\text{frequency of } x \text{ in } n \text{ repeated experiments}}{n} && (4.8)\\
&= \sum_{x \in \text{supp}(X)} x \mathbb{P}(X = x) && (4.9)\\
&= \sum_{x \in \text{supp}(X)} x f_X(x) && (4.10)
\end{aligned}$$

■

Since expectation is the long run average value of the random variable X, it is sometimes called the mean of the random variable.

Example 17

Find the expected value of rolling a dice.

Solution

Let X be the random variable that denotes the outcome of rolling the die. Then X can take on values $1, 2, 3, 4, 5, 6$. Each of these values have probability greater than 0. Thus, $\text{supp}(X) = \{1, 2, 3, 4, 5, 6\}$ Thus,

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x \in \text{supp}(X)} x f_X(x) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5\end{aligned}$$

Thus, if we roll a die infinitely many times, the average value we get is 3.5.

Example 18

We say $\mathbb{I}_{\mathbb{A}}$ is the indicator variable for the event A if,

$$\mathbb{I}_{\mathbb{A}} = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur} \end{cases} \tag{4.11}$$

Find the expected value of $\mathbb{I}_{\mathbb{A}}$

Solution

A not occuring is the same as A^c occuring. Thus,

$$\mathbb{E}(\mathbb{I}_{\mathbb{A}}) = 1 \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}(A^c) = \mathbb{P}(A) \tag{4.12}$$

Thus, the expected value of the indicator variable is the probability of event A occurring.

The expectation can also be thought of as the center of mass of the probability distribution. If we had a system where the masses at a point x_i are equal to the probability of x_i, the point at which the system would balance would be the expected value of X.

Remark The probability concept of expectation is analogous to the physical concept of the *center of gravity* of a distribution of mass. Consider a discrete random variable X having probability mass function $p(x_i), i \geq 1$. If we now imagine a weightless rod in which weights with mass $p(x_i), i \geq 1$, are located at the points $x_i, i \geq 1$ (see Figure 4.4), then the point at which the rod would be in balance is known as the center of gravity. For those readers acquainted with elementary statics, it is now a simple matter to show that this point is at $E[X]$.† ■

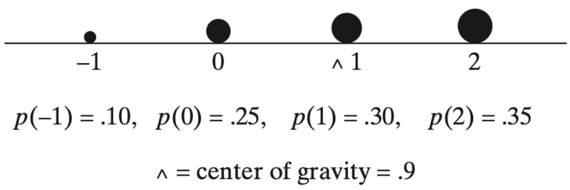


Figure 10: Expectation as Center of Gravity

4.2.2 Variance

Consider the following random variables and their associated probability mass functions: $X = 0$ with probability 1, $Y = \pm 1$ with probability $\frac{1}{2}$ and $Z = \pm 100$ with probability $\frac{1}{2}$ All of these have the same expectation or measure of center: $\mathbb{E}(X) = \mathbb{E}(Y) = \mathbb{E}(Z) = 0$.

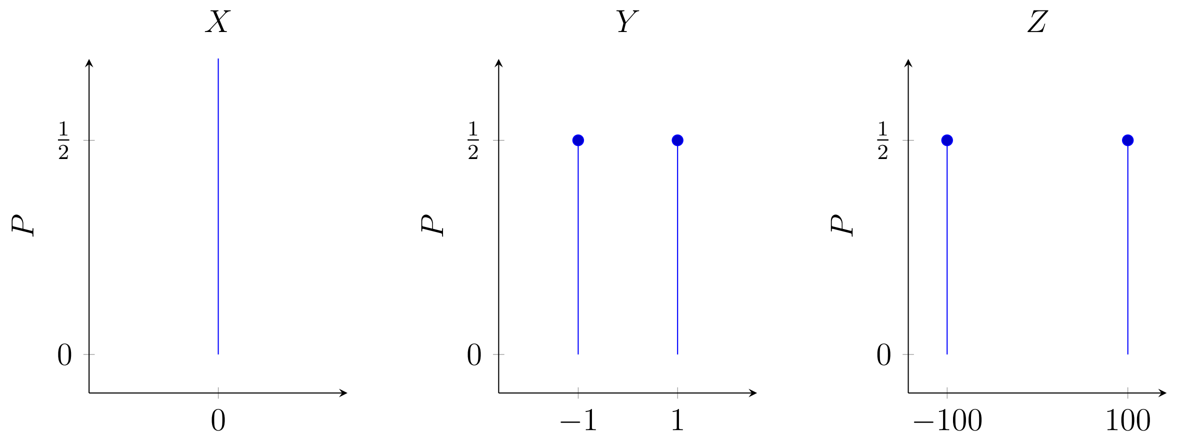


Figure 11: Probability mass functions of X, Y, and Z.

However, the probability distributions themselves are spread out very differently. We quantify the measure of spread using variance.

Definition 4.6 (Variance of a Discrete Random Variable)**.** If X is a discrete random variable, then the variance of X is given by,

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \tag{4.13}$$

We define the expectation of the function of the random variable as,

$$\mathbb{E}(g(X)) = \sum_{x \in \mathrm{supp}(X)} g(x) f_X(x) \tag{4.14}$$

Thus,

$$\mathrm{Var}(X) = \sum_{x \in \mathrm{supp}(X)} x^2 f_X(x) - \left(\sum_{x \in \mathrm{supp}(X)} x f_X(x) \right)^2 \tag{4.15}$$

A higher variance means the distribution is spread out more from the mean. This means that our experiment is less deterministic. Some properties from the definition of expectation and variance follows,

Theorem 4.1 (Linearity of Expectation)**.** For two constants a and b,

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b \tag{4.16}$$

Proof.

$$\begin{aligned}
\mathbb{E}(aX + b) &= \sum_{x \in \mathrm{supp}(X)} (ax + b) f_X(x) \\
&= a \sum_{x \in \mathrm{supp}(X)} x f_X(x) + b \sum_{x \in \mathrm{supp}(X)} f_X(x) \\
&= a\mathbb{E}(X) + b(1) \\
&= a\mathbb{E}(X) + b
\end{aligned}$$

■

This says that translating your probability distribution by some constant b will translate your expectation by b as well. Since expectation is a linear operator, the underlying vector space is

$$\chi = \{X \mid X \text{ is a random variable}\} \tag{4.17}$$

Then $\mathbb{E} : \chi \to \mathbb{R}$ is a linear operator.

Theorem 4.2. For two constants a and b,

$$\text{Var}(aX + b) = a^2\text{Var}(X) \tag{4.18}$$

Proof.

$$\begin{aligned}
\text{Var}(aX + b) &= \mathbb{E}((aX + b)^2) - (\mathbb{E}(aX + b))^2 \\
&= \mathbb{E}(a^2X^2 + 2abX + b^2) - (a\mathbb{E}(X) + b)^2 \\
&= a^2\mathbb{E}(X^2) + 2ab\mathbb{E}(X) + b^2) - (a^2(\mathbb{E}(X))^2 + 2ab\mathbb{E}(X) + b^2) \\
&= a^2\mathbb{E}(X^2) - a^2(\mathbb{E}(X))^2 \\
&= a^2(\mathbb{E}(X^2) - (\mathbb{E}(X)^2)) \\
&= a^2\text{Var}(X)
\end{aligned}$$

■

This says that translating your probability distribution by some constant b does not affect its spread. A crucial fact involving the linearity of expectation is that for random variables $X_1, X_2, \cdot, X_n$

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{N} \mathbb{E}(X_i) \tag{4.19}$$

This holds true regardless of the dependence between each X_i

Example 19

Suppose you run n trials where trial i succeeds independently with probability p_i. What is the expected number of successes?

Solution

Let X_i be defined as

$$X_i = \mathbb{I}_{\text{trial } i \text{ succeeds}} \tag{4.20}$$

Then,

$$\mathbb{E}(X_i) = p_i \tag{4.21}$$

The total number of successes is then given by

$$S = \sum_{i=1}^{n} X_i \tag{4.22}$$

Thus,

$$\mathbb{E}(S) = \sum_{i=1}^{n} p_i \tag{4.23}$$

4.3 Bernoulli Random Variable

We start by defining a Bernoulli experiment as follows

Definition 4.7 (Bernoulli Experiment)**.** A Bernoulli Experiment is a single random experiment with two outcomes: a success with probability p and a failure with probability $1 - p$.

For example, a Bernoulli experiment could be tossing a dice and getting a 6. Then, a success would be getting a 6 and a failure is not getting a 6. For a Bernoulli experiment, we can define a Bernoulli random variable

Definition 4.8 (Bernoulli Random Variable)**.** A Bernoulli random variable X is a random variable where $\text{supp}(X) = \{0, 1\}$. In other words, there are two outcomes. 1 refers to a "success" and 0 refers to a failure. The probability of a success is p and the probability of a failure is $1 - p$.

$$\mathbb{P}(X = 1) = p \tag{4.24}$$

$$\mathbb{P}(X = 0) = 1 - p \tag{4.25}$$

A Bernoulli random variable is denoted as

$$X \sim \text{Bernoulli(p)} \tag{4.26}$$

Lemma 4.3. If $X \sim \text{Bernoulli}(p)$ then,

$$\mathbb{E}(X) = p \tag{4.27}$$

$$\text{Var}(X) = p(1 - p) \tag{4.28}$$

Proof.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x \in \text{supp}(X)} x f_X(x) \\
&= 0 \cdot f_X(0) + 1 \cdot f_X(1) \\
&= \mathbb{P}(X = 1) \\
&= p \\
\text{Var}(X) &= \mathbb{E}(X^2) - (E(X))^2 \\
&= \sum_{x \in \text{supp}(X)} x^2 f_X(x) - \left(\sum_{x \in \text{supp}(X)} x f_X(x) \right)^2 \\
&= 0^2(1 - p) + 1^2 p - p^2 \\
&= p - p^2 \\
&= p(1 - p)
\end{aligned}$$

■

4.4 Binomial Random Variable

The Bernoulli trial and random variable is the foundation for a binomial experiment and random variable.

Definition 4.9 (Binomial Experiment)**.** Given an independent bernoulli experiment, a binomial experiment repeats the bernoulli experiment n times and counts the number of successes.

From a binomial experiment, we define a binomial random variable

Definition 4.10 (Binomial Random Variable)**.** A Binomial Random Variable X is a random variable that characterises a binomial experiment. It is modelled by two parameters: n (the number of times we repeat the Bernoulli trial) and p (the probability of a success). It is denoted as

$$X \sim \text{Bin}(n, p) \tag{4.29}$$

The associated probability mass function is used to determine the number of successes. It says that given $X \sim \text{Bin}(n, p)$, the probability of k successes is

$$f_X(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.30}$$

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

$$\text{supp}(X) = \{0, 1, \ldots, n\} \tag{4.31}$$

We know that the probability mass function defined above is valid since

$$\sum_{k \in \text{supp}(X)} f_X(k) = \sum_{x=0}^{n} \binom{n}{x} p^k (1-p)^{n-k} \tag{4.32}$$

$$= (p + 1 - p)^n \tag{4.33}$$

$$= 1 \tag{4.34}$$

Example 20

Suppose we flip 10 coins. What is the probability we get 5 heads?

Example 21

This is a binomial experiment with $X \sim \text{Bin}\left(10, \frac{1}{2}\right)$. Then

$$\mathbb{P}(X = 5) = \binom{10}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^5 \tag{4.35}$$

Lemma 4.4. If $X \sim \text{Bin}(n, p)$ then,

$$\mathbb{E}(X) = np \tag{4.36}$$

$$\text{Var}(x) = np(1-p) \tag{4.37}$$

Proof. We can write out our binomial random variable as the sum of n Bernoulli random variables.

$$X = \sum_{i=1}^{n} X_i \tag{4.38}$$

where X_i is the Bernoulli random variable corresponding to the i^{th} trial. Then,

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \\
&= E(X_1 + \cdots + X_n) \text{ (By the linearity of expectation)} \\
&= \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) \\
&= \sum_{i=1}^{n} \mathbb{E}(X_i)
\end{aligned}$$

We know if X Bernoulli(p) then $\mathbb{E}(X) = p$. Thus,

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{i=1}^{n} p \\
&= np
\end{aligned}$$

Further,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \mathrm{Var}(X_1 + \cdots + X_n) \\
&= \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) \\
&= \sum_{i=1}^{n} \mathrm{Var}(X_i)
\end{aligned}$$

We know if X Bernoulli(p) then $\mathrm{Var}(X) = p(1-p)$. Thus,

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{i=1}^{n} p(1-p) \\
&= np(1-p)
\end{aligned}$$

■

In fact,

$$X \sim \mathrm{Bernoulli}(p) \implies X \sim \mathrm{Bin}(1, p) \tag{4.39}$$

Thus,

$$\mathbb{E}(X) = p \tag{4.40}$$

$$\mathrm{Var}(x) = p(1-p) \tag{4.41}$$

4.5 Poisson Random Variable

Definition 4.11 (Poisson Random Variable)**.** A random variable X is said to be a Poisson random variable with parameter $\lambda > 0$ if its distribution, or probability mass function is given by

$$\mathbb{P}(X = k) = f_X(k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{4.42}$$

where

$$\text{supp}(X) = \mathbb{Z}_{\geq 0} = \{0, 1, \cdots\} \tag{4.43}$$

The Poisson probability mass function is a valid mass function since $f_X(k) \geq 0 \forall k \in \{0, 1, \cdots\}$. Further,

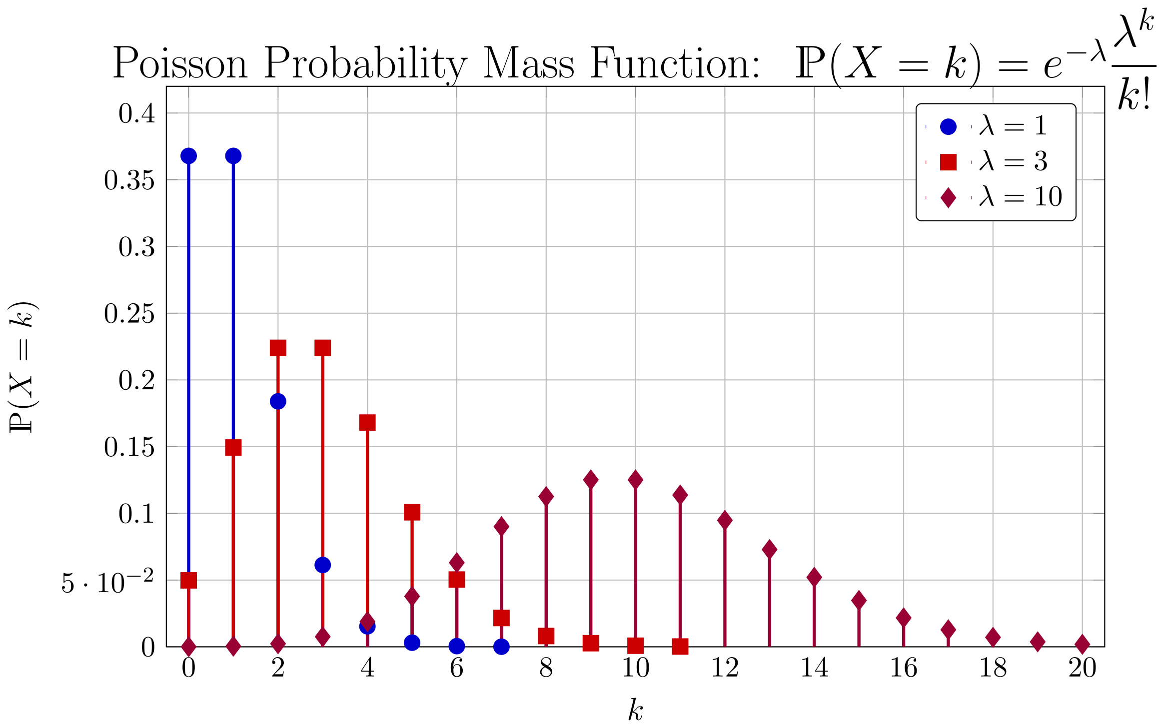
$$\sum_{k \in \text{supp}(X)} f_X(x) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \tag{4.44}$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \tag{4.45}$$

$$= e^{-\lambda} e^{\lambda} \text{ (Via the Taylor Series of } e^x) \tag{4.46}$$

$$= e^0 \tag{4.47}$$

$$= 1 \tag{4.48}$$



Poisson distributions are used to model "rare" events that occur independently over some window (time, space, volume etc). The notion of rareness comes from the relation between the Poisson and Binomial distribution. Suppose $Y \sim \text{Bin}(n, p)$ where n is large and p is small. If we define $\lambda = np$ and it is moderate, then Y is approximately a Poisson

Distribution given by

$$Y \approx \text{Poi}(np) \tag{4.49}$$

The idea is that if you are counting a large number of independent trials, each of which has a tiny probability of success, then the total number of successes is well-modelled by a Poisson distribution. In essence, if n is large and p is small, then

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{e^{-np}(np)^k}{k!} \tag{4.50}$$

We can formally prove this.

Theorem 4.5 (Rare Event Limit of Poisson Distribution)**.** Let Y be a discrete random variable such that $Y \sim \text{Bin}(n, p)$. If n is large and p is small, and $\lambda = np$ is moderate. Then,

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{e^{-np}(np)^k}{k!} \tag{4.51}$$

Proof. Define $\lambda = np \implies p = \frac{\lambda}{n}$. Thus,

$$\begin{aligned}
\mathbb{P}(Y = k) &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} && (4.52)\\
&= \frac{n!}{(n-k)!k!} && (4.53)\\
&= \frac{\lambda^k}{n^k} \frac{1}{k!} \prod_{i=0}^{k-1} (n-i) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} && (4.54)\\
&= \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \frac{\lambda^k}{k!} \prod_{i=0}^{k-1} \frac{n-i}{n} && (4.55)
\end{aligned}$$

Now the product term is given by

$$\begin{aligned}
\prod_{i=0}^{k-1} \frac{n-i}{n} &= \frac{n}{n} \cdot \left(\frac{n-1}{n}\right) \cdot \left(\frac{n-2}{n}\right) \cdots \left(\frac{n-k+1}{n}\right) && (4.56)\\
&= 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) && (4.57)
\end{aligned}$$

We have k to be fixed. We need to take each term in our equation for $\mathbb{P}(Y = k)$ to infinity to get an approximation for large n,

$$\lim_{n\to\infty} 1 \cdot \left(1-\frac{1}{n}\right) \cdot \left(1-\frac{2}{n}\right) \cdots \left(1-\frac{k-1}{n}\right) = 1 \tag{4.58}$$

Further,

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda} \tag{4.59}$$

And,

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^k = 1^{-k} = 1 \tag{4.60}$$

Then,

$$P(Y=k) \approx \frac{e^{-\lambda}\lambda^k}{k!} \tag{4.61}$$

■

Lemma 4.6. If $X \sim \text{Poi}(\lambda)$, then

$$\mathbb{E}(X) = \lambda \tag{4.62}$$

$$\text{Var}(X) = \lambda \tag{4.63}$$

Proof.

$$E(X) = \sum_{x\in\text{supp}(X)} x f_X(x) \tag{4.64}$$

$$= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \tag{4.65}$$

Since the $x=0$ term will make the product 0, we re-index with $x=1$,

$$E(x) = \sum_{x=1}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \tag{4.66}$$

$$= \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} \tag{4.67}$$

$$= \lambda e^{-\lambda}\sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \tag{4.68}$$

Let $n = x - 1$. Then,

$$\mathbb{E}(X) = \lambda e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \tag{4.69}$$

$$= \lambda e^{-\lambda} e^{\lambda} \tag{4.70}$$

$$= \lambda \tag{4.71}$$

Now,

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 \tag{4.72}$$

$$= \sum_{x=0}^{\infty} \left(x^2 \frac{e^{-\lambda}\lambda^x}{x!} \right) - \lambda^2 \tag{4.73}$$

$$= \sum_{x=0}^{\infty} \left(\frac{x\lambda^x e^{-\lambda}}{(x-1)!} \right) - \lambda^2 \tag{4.74}$$

Neglecting the $x = 0$ term again,

$$\sum_{x=1}^{\infty} \left(\frac{x\lambda^x e^{-\lambda}}{(x-1)!} \right) - \lambda^2 \tag{4.75}$$

Let $n = x - 1 \implies x = n + 1$. Then,

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{n=0}^{\infty} \left(\frac{(n+1)\lambda^{n+1} e^{-\lambda}}{n!} \right) - \lambda^2 && (4.76)\\
&= \lambda e^{-\lambda} \sum_{n=0}^{\infty} \frac{(n+1)\lambda^n}{n!} - \lambda^2 && (4.77)\\
&= \lambda e^{-\lambda} \left(\sum_{n=0}^{\infty} \frac{n\lambda^n}{n!} + \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \right) - \lambda^2 && (4.78)\\
&= \lambda e^{-\lambda} \left(\sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!} + e^{\lambda} \right) - \lambda^2 && (4.79)\\
&= \lambda e^{-\lambda} \left(\lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} + e^{\lambda} \right) - \lambda^2 && (4.80)\\
&= \lambda e^{-\lambda} \left(\lambda e^{\lambda} + e^{\lambda} \right) - \lambda^2 && (4.81)\\
&= \lambda^2 + \lambda - \lambda^2 && (4.82)\\
&= \lambda && (4.83)
\end{aligned}$$

■

Thus, the mean and the variance of the Poisson distribution is simply the Poisson parameter. This also makes sense. Recall, that the Poisson distribution is an approximation of the Binomial distribution for large n, small p and moderate $\lambda = np$ Recall that if $X \sim \mathrm{Bin}(n, p)$, then $\mathbb{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$. With our approximations then,

$$\mathbb{E}(X) = np = \lambda \tag{4.84}$$

$$\mathrm{Var}(X) = np(1-p) \approx \lambda \tag{4.85}$$

Where the last equality comes from the fact that if p is really small, $1 - p \approx p$

Example 22

The monthly worldwide average number of airplane crashes of commercial airlines is 3.5. What is the probability of at least two crashes next month. What is the probability at least two such accidents occur in the next two months.

Solution

Since airplane crashes are rare events spread over many flights. We have $X \sim \text{Poi}(3.5)$. Thus,

$$\mathbb{P}(X \geq 2) = 1 - P(X = 0) - P(X = 1) \tag{4.86}$$
$$= 1 - \frac{e^{-3.5}3.5^0}{0!} - \frac{3.5e^{-3.5}}{1!} \tag{4.87}$$
$$= 1 - 4.5e^{-3.5} \tag{4.88}$$
$$\approx 0.864 \tag{4.89}$$

For two months, we have $X \sim \text{Poi}(7)$

4.5.1 The Poisson Paradigm

The Binomial to Poisson approximation required each trial to be independent. However, the Poisson Paradigm states that this approximation is valid for even weakly dependent trials. To define weak dependence, we need to define the degree of an event.

Definition 4.12 (Degree of an Event)**.** Given events $A_1, A_2, \cdots, A_n$, the degree of an event $\deg(A_i)$ is defined as the number of events A_j $(j \neq i)$ such that A_i and A_j are dependence.

Then,

Definition 4.13 (Weakly Dependent Events)**.** Events $A_1, A_2, \cdots, A_n$ are weakly dependent if there exists a constant L such that

$$\deg(A_i) \leq L \forall i \in \{1, 2, \cdots, n\} \tag{4.90}$$

Thus, no matter how many events we have, each individual event depends on at most L others. Then,

Definition 4.14 (Poisson Paradigm)**.** Consider n events $A_1, A_2, \cdots, A_n$ with $\mathbb{P}(A_i) = p_i$. If all p_i are small and the events are independent or weakly dependent, then the number of the events that occur can be approximated as a Poisson distribution with $\lambda = \sum_{i=1}^{n} p_i$

Example 23

A total of $2n$ people, consisting of n married couples are randomly seated (all possible orderings are equally likely) at a round table. Let C_i denote the event that the members of couple i are seated together for $i \in \{1, \cdots, n\}$. Find

1. a). $\mathbb{P}(C_i)$
2. b). $\mathbb{P}(C_j|C_i)$ for $j \neq i$
3. c). Approximate the probability for n large that no married couples are seated next to each other

Solution

1. a). Let us fix one member of couple i. Then, there are $2n-1$ seats left, with 2 seats being next to the fixed member. Thus,

$$\mathbb{P}(C_i) = \frac{2}{2n-1} \tag{4.91}$$

b). If one couple is already seated next to each other

Example 24

Customers enter a store on a given day with the number of customers described by $X \sim \text{Poi}(\lambda)$. Each customer independently buys something with probability p. Let Y be the number of customers who buy something and Z be the number of customers who don't buy something. We want to find the distributions of Y and Z.

Solution

Let us assume n customers enter a store. Then, $Y|X = n \sim \text{Bin}(n, p)$.. Thus,

$$\mathbb{P}(Y = k) = \sum_{n=k}^{\infty} \mathbb{P}(Y = k|X = n) \cdot \mathbb{P}(X = n) \tag{4.92}$$

$$= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{e^{-\lambda}\lambda^n}{n!} \tag{4.93}$$

$$= \sum_{n=k}^{\infty} \frac{e^{-\lambda}\lambda^n}{k!(n-k)!} p^k (1-p)^{n-k} \tag{4.94}$$

$$= \frac{(\lambda p)^k e^{-\lambda}}{k!} \sum_{n=k}^{\infty} \frac{[(1-p)\lambda]^{n-k}}{(n-k)!} \tag{4.95}$$

$$= \frac{(\lambda p)^k e^{-\lambda}}{k!} \sum_{m=0}^{\infty} \frac{[(1-p)\lambda]^{m}}{m!} \tag{4.96}$$

$$= \frac{(\lambda p)^k e^{-\lambda}}{k!} e^{(1-p)\lambda} \tag{4.97}$$

$$= \frac{e^{-\lambda p}(\lambda p)^k}{k!} \tag{4.98}$$

Thus,

$$Y \sim \text{Poi}(\lambda p) \tag{4.99}$$

By an identical argument where we now have $1-p$ since customers don't buy any item,

$$Z \sim \text{Poi}(\lambda(1-p)) \tag{4.100}$$

The above is known as Poisson splitting. A $\text{Poi}(\lambda)$ stream where each event is independently classified into one of two types with probability p and $1-p$, produces two independent Poisson streams with rate λp and $\lambda(1-p)$. The independence of Y and Z is especially surprising — after all, $Y + Z = X$ so knowing Y completely determines Z given X. But marginally (without conditioning on X), they turn out to be independent. This is a special property of the Poisson that no other distribution shares.

4.6 Geometric Random Variable

Given an infinite sequence of non-negative numbers x_n such that the sum converges to some non-negative value $L < \infty$

$$\sum_{n=0}^{\infty} x_n = L \tag{4.101}$$

We can construct a random variable X by setting

$$\mathbb{P}(X = n) = \frac{x_n}{L} \tag{4.102}$$

This is a valid PMF since

$$\sum_{n=0}^{\infty} \mathbb{P}(X = n) = \sum_{n=0}^{\infty} \frac{x_n}{L} \tag{4.103}$$

$$= \frac{L}{L} \tag{4.104}$$

$$= 1 \tag{4.105}$$

If we set $x_n = p(1-p)^{n-1}$ where $0 < p < 1$ and $n = 1, 2, \cdots$, then,

$$\sum_{n=1}^{\infty} x_n = p + p(1-p) + p(1-p)^2 + \cdots \tag{4.106}$$

$$= \frac{p}{1-(1-p)} \tag{4.107}$$

$$= \frac{p}{p} \tag{4.108}$$

$$= 1 \tag{4.109}$$

Thus, in this case, $L = 1$ and $f_X(n) = p(1-p)^{n-1}$. This is known as the Geometric distribution.

Definition 4.15 (Geometric Random Variable)**.** A random variable X is a Geometric random variable with parameter p if

$$\mathbb{P}(X = n) = p(1-p)^{n-1} \tag{4.110}$$

where $\text{supp}(X) = \{1, 2, \cdots\}$

$X \sim \text{G}(p)$ has the interpretation of us performing a sequence of independent Bernoulli trials with probability p and $\mathbb{P}(X = n)$ is the probability we get our first success on the n^{th} trial.

4.6.1 Memorylesness of the Geometric Variable

Note that

$$\mathbb{P}(X > k) = \sum_{k+1}^{\infty} p(1-p)^{n-1} \tag{4.111}$$

$$= p(1-p)^k + p(1-p)^{k+1} + \cdots \tag{4.112}$$

$$= p(1-p)^k(1 + p + p^2 + \dots) \tag{4.113}$$

$$= p(1-p)^k \sum_{n=0}^{\infty} p^n \tag{4.114}$$

$$= (1-p)^k \tag{4.115}$$

This is the probability that the first k trials are all failures. Now for any $k, n > 0$

$$\mathbb{P}(X = k + n | X > k) = \frac{\mathbb{P}(X = k + n)}{\mathbb{P}(X > k)} \tag{4.116}$$

$$= \frac{p(1-p)^{k+n-1}}{(1-p)^k} \tag{4.117}$$

$$= p(1-p)^{n-1} \tag{4.118}$$

$$= p(1-p)^n \tag{4.119}$$

$$= \mathbb{P}(X = n) \tag{4.120}$$

$\mathbb{P}(X = k + n | X > k)$ is the probability our first success is on the $k + n$ trial given that we failed the first k trials. Thus, given that the first k trials were failures, the remaining waiting time for the first success also has the same distribution $\text{G}(p)$ as if we were starting to fresh. This makes sense, if we were to flip a coin and we get tails the first 9 times. The probability we get a head on the 10th time does not depend on us getting a tails the previous 9 times.

Lemma 4.7. If $X \sim \mathrm{G}(p)$ then,

$$\mathbb{E}(X) = \frac{1}{p} \tag{4.121}$$

$$\mathrm{Var}(X) = \frac{1-p}{p^2} \tag{4.122}$$

Proof.

$$\mathbb{E}(X) = \sum_{x \in \mathrm{supp}(X)} x f_X(x) \tag{4.123}$$

$$= \sum_{n=1}^{\infty} np(1-p)^{n-1} \tag{4.124}$$

$$= p \sum_{n=1}^{\infty} n(1-p)^{n-1} \tag{4.125}$$

We need to recognise that $nx^{n-1} = \frac{d}{dx}x^n$ Then,

$$\sum_{n=1}^{\infty} nx^{n-1} = \sum_{n=1}^{\infty} \frac{d}{dx}x^n \tag{4.126}$$

We can swap the sum and the derivative because power series converge uniformly inside their radius of convergence. So,

$$\sum_{n=1}^{\infty} nx^{n-1} = \frac{d}{dx}\left(\sum_{n=1}^{\infty} x^n\right) \tag{4.127}$$

$$= \frac{d}{dx}\frac{x}{1-x} \tag{4.128}$$

$$= \frac{1}{(1-x)^2} \tag{4.129}$$

Now let $x = 1-p \implies p = 1-x$. Then,

$$\mathbb{E}(X) = p \cdot \frac{1}{p^2} \tag{4.130}$$

$$= \frac{1}{p} \tag{4.131}$$

Now,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 && (4.132)\\
&= \sum_{n=1}^{\infty} n^2 p(1-p)^{n-1} - \frac{1}{p^2} && (4.133)\\
&= p\sum_{n=1}^{\infty} n^2 (1-p)^{n-1} - \frac{1}{p^2} && (4.134)\\
& && (4.135)
\end{aligned}$$

We can write $n^2 = n(n-1) + n$ Then,

$$\begin{aligned}
\mathrm{Var}(X) &= p\sum_{n=1}^{\infty} n^2 x^{n-1} - \frac{1}{p^2} && (4.136)\\
&= p\sum_{n=1}^{\infty} n(n-1)x^{n-2} + p\sum_{n=1}^{\infty} nx^{n-1} - \frac{1}{p^2} && (4.137)\\
&= px\sum_{n=2}^{\infty} n(n-1)x^{n-2} + p\sum_{n=1}^{\infty} nx^{n-1} - \frac{1}{p^2} && (4.138)\\
&= px\sum_{n=2}^{\infty} \frac{d^2}{dx^2}x^n + p\sum_{n=1}^{\infty} nx^{n-1} - \frac{1}{p^2} && (4.139)\\
&= px\frac{d^2}{dx^2}\sum_{n=2}^{\infty} x^n + p\sum_{n=1}^{\infty} nx^{n-1} - \frac{1}{p^2} && (4.140)\\
&= px\frac{d^2}{dx^2}\left(\frac{x^2}{1-x}\right) + p\sum_{n=1}^{\infty} nx^{n-1} - \frac{1}{p^2} && (4.141)\\
&= px\frac{2}{(1-x)^3} + \frac{1}{p} - \frac{1}{p^2} && (4.142)\\
&= p(1-p)\frac{2(1-p)}{p^3} + \frac{1}{p} - \frac{1}{p^2} && (4.143)\\
&= \frac{2p(1-p)}{p^3} + \frac{1}{p} - \frac{1}{p^2} && (4.144)\\
&= \frac{2(1-p)}{p^2} + \frac{p}{p^2} - \frac{1}{p^2} && (4.145)\\
&= \frac{2-2p+p-1}{p^2} && (4.146)\\
&= \frac{1-p}{p^2} && (4.147)
\end{aligned}$$

■

Example 25

Two independent machines A and B fail each day with probabilities p and q respectively. Let $T_A \sim \mathrm{G}(p)$ be the first day machine A fails and $T_b \sim \mathrm{G}(q)$ be the first day machine B fails. The machines are independent. Find $\mathbb{P}(\min(T_A, T_B) > k)$ and $\mathbb{P}(T_A < T_B)$.

Solution

$\mathbb{P}(\min(T_A, T_B) > k)$ is the probability that neither machine has failed in the first k days. Since T_A and T_B are independent,

$$\mathbb{P}(\min(T_A, T_B) > k) = \mathbb{P}(T_A > k \cap T_B > k) \tag{4.148}$$
$$= \mathbb{P}(T_A > k)\mathbb{P}(T_B > k) \tag{4.149}$$
$$= (1-p)^k(1-q)^k \tag{4.150}$$
$$= [(1-p)(1-q)]^k \tag{4.151}$$

Note that this is the failure rate of a geometric distribution with $1 - p' = (1 - p)(1-q)$. Thus, $1-(1-p)(1-q) = p + q - pq$. Thus, the minimum of the waiting times of two independent Geometric waiting times is itself Geometric, with a success probability equal to the probability that at least one event fails. Now, we want to find $\mathbb{P}(T_A < T_B)$. Let $T_A = k$ We then require $T_B > k$.

$$\mathbb{P}(T_A < T_B) = \sum_{k=1}^{\infty} \mathbb{P}(T = k)\mathbb{P}(T_A > k) \tag{4.152}$$
$$= \sum_{k=1}^{\infty} p(1-p)^{k-1}(1-q)^k \tag{4.153}$$
$$= p(1-q)\sum_{k=1}^{\infty}((1-p)(1-q))^{k-1} \tag{4.154}$$
$$= \frac{p(1-q)}{1-(1-p)(1-q)} \tag{4.155}$$
$$= \frac{p(1-q)}{p+q-pq} \tag{4.156}$$

4.7 Negative Binomial Random Variable

Definition 4.16. X is a negative binomial random variable with parameter $0 < p < 1$ and $r \geq 1$ if its probability mass function is given by

$$\mathbb{P}(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r} \tag{4.157}$$

where

$$\text{supp}(X) = \{r, r+1, r+2 \cdots\} \tag{4.158}$$

We can interpret the negative binomial distribution as a generalisation of the Geometric distribution. While the Geometric distribution measured the probability of getting our first success on the n^{th} trial, the negative binomial distribution measures the probability of getting the r^{th} success on the n^{th} trial. To understand the probability distribution, out of the $n-1$ trials, exactly $r-1$ must be successes. The number of ways to arrange this is $\binom{n-1}{r-1}$ with each configuration having probability $p^{r-1}(1-p)^{n-1-(r-1)} = p^{r-1}(1-p)^{n-r}$. Further, trial n must be a success. Thus, multiplying everything gives the PMF above. When $r = 1$, the negative binomial distribution collapses to the Geometric distribution.

Lemma 4.8. If $X \sim \text{NegBin}(p, r)$ then,

$$\mathbb{E}(X) = \frac{r}{p} \tag{4.159}$$

$$\text{Var}(X) = \frac{r(1-p)}{p^2} \tag{4.160}$$

Proof. ∎

The equations for expectation and variance make sense. These are the expectation and variance of the geometric distribution multiplied by r. waiting for the r^{th} success is like waiting r independent "first successes" chained together. Each "leg" of the wait contributes $\frac{1}{p}$ to the expected time and $\frac{1-p}{p^2}$ to the variance. Since the legs are independent, both the means and variances add.

Example 26

The Banach Match Problem. A pipe-smoking mathematician always carries two matchboxes, one in each pocket. Each time he needs a match, he reaches into a random pocket (left or right, each with probability $\frac{1}{2}$. Both boxes initially contain N matches. At some point, he reaches into a pocket and finds that box empty. What is the probability that the other box has exactly k matches remaining?

Solution

Since one of the boxes is empty. He must have drawn from it N times. To find that it is empty, he must have drawn from that box $N+1$ times. Further, he has drawn from the box $N-k$ times. Thus, the total number of draws is $2N-k+1$. The last draw (draw number $2N-k+1$) is the one that discovered the empty box. Thus, the remaining $2N-k$ draws picked out N matches. There are $\binom{2N-k}{N}$ ways for this to happen. Each sequence has probability $\left(\frac{1}{2}\right)^{2N-k+1}$. Since either the left or the right pocket could be empty, we multiply by a factor of 2 to get

$$\mathbb{P}(\text{other box has } k \text{ matches}) = 2\binom{2N-k}{N}\left(\frac{1}{2}\right)^{2N-k+1} \tag{4.161}$$

$$= \binom{2N-k}{N}\left(\frac{1}{2}\right)^{2N-k} \tag{4.162}$$

We can reparamaterise the negative binomial distribution in terms of failures. If $l = n - r$ is the number of failures before the r^{th} success then $l \in \{0, 1, \cdots\}$ and

$$\mathbb{P}(X = r + l) = \binom{l+r-1}{l} p^r (1-p)^l \tag{4.163}$$

This way of writing the distribution is often more convenient for algebraic manipulations. The name of the distribution comes from the negative binomial theorem which generalises the ordinary binomial theorem to negative exponents. For $|x| < 1$,

$$(1-x)^{-r} = \sum_{l=0}^{\infty} \binom{l+r-1}{l} x^l \tag{4.164}$$

The expansion produces the same coefficients in the probability mass function. We can apply this to show that the obtained probability mass function is valid

$$\sum_{l=0}^{\infty} p^r (1-p)^l = p^r \sum_{l=0}^{\infty} \binom{l+r-1}{l} (1-p)^l \tag{4.165}$$

$$= p^r (1-(1-p))^{-r} \tag{4.166}$$

$$= p^r \cdot p^{-r} \tag{4.167}$$

$$= 1 \tag{4.168}$$

4.8 Multinomial Random Variable

The Binomial distribution has two outcomes per trial. The multinomial distribution generalises this to k outcomes $O_1, O_2, \cdots, O_k$ per trial where $P(O_i) = p_i$ such that

$$\sum_{i=1}^{k} p_i = 1 \tag{4.169}$$

For example, $k = 6$ when we roll a die. The probability of observing exactly n_i occurrences of outcome O_i where $(n_1 + n_2 + \cdots + n_k = n)$ is

$$\mathbb{P}(X_1 = n_1, X_2 = n_2, \cdots, X_k = n_k) = \binom{n}{n_1, n_2, \cdots, n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \tag{4.170}$$

This is a valid probability mass function by the multinomial theorem

$$\sum_{n_1+\cdots+n_k=n} \binom{n}{n_1, n_2, \cdots, n_k} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} = (p_1 + \cdots p_k)^n = 1 \tag{4.171}$$

The result is a random vector

$$X : S \to \mathbb{R}^k = (X_1, \cdots, X_k) \tag{4.172}$$

where each X_i is the number of times outcome O_i occurred

Definition 4.17 (Multinomial Random Variable)**.** X is a multinomial random variable

4.9 Hypergeometric Distribution

The Binomial distribution models sampling with replacement since each draw is independent of each other. The hypergeometric distribution models sampling with replacement, where earlier trials affect later ones.

Definition 4.18 (Hypergeometric Random Variable)**.** A random variable X is a Hypergeometric random variable with parameter N, m, and n if

$$\mathbb{P}(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{4.173}$$

where $\operatorname{supp}(X) = \{0, 1, \cdots, n\}$. N is typically interpreted as the "total population", m is the number of "special items", and n is the number of items picked.

For example, if we have an urn with N balls where m are white balls and $N - m$ are red balls. Let us assume we draw n balls without replacement. The number of white balls we get in our draw is distributed as $X \sim \text{HyperGeo}(N, m, n)$. We need $N > \max(m, n)$. Continuing with our example, there are a total of $\binom{N}{n}$ ways to choose n balls from N balls. The total number of ways to choose k white balls from m white balls is $\binom{m}{k}$. The ways to then choose $n - k$ red balls from $N - m$ red balls is $\binom{N-m}{n-k}$. Thus, the probability of a favourable outcome is given by the PMF. This is a valid PMF via Vandermonde's identity which states that

$$\sum_{k=0}^{n} \binom{m}{k}\binom{N-m}{n-k} = \binom{N}{n} \tag{4.174}$$

The proof of Vandermonde's identity is combinatorial. If we want to choose k objects from m objects, there are $\binom{m}{k}$ ways to do this. The remaining $n - k$ objects must be chosen from the remaining $N - m$ objects. The sum over all possible values of k must give you the number of ways to choose n objects from N objects.

4.10 Other Random Variables

Other common distributions described by discrete random variables include the Benford distribution and Zipf distribution. Benford's distribution says that

$$\mathbb{P}(X = n) = \log_b \left(1 + \frac{1}{n}\right) \tag{4.175}$$

where b is the base of the number system (typically 10). his describes the distribution of the leading digit of "naturally occurring" numbers — things like population counts, financial figures, physical constants. The surprising prediction is that smaller digits are far more likely to be the leading digit. For base 10:

$$\mathbb{P}(X = 1) = \log(2) \approx 0.301 \tag{4.176}$$

Thus, o digit 1 appears as the leading digit about 30% of the time, not the 11% you'd expect if digits were uniform. Zipf's distribution says that

$$\mathbb{P}(X = n) = cn^{-\alpha} \tag{4.177}$$

where $n = 1, 2, \ldots$. where c is a normalizing constant (chosen so the pmf sums to 1) and $\alpha > 1$ This models phenomena where frequency drops off as a power of rank. Classic examples include word frequencies in natural language (the most common word appears roughly twice as often as the second most common), city sizes, and website traffic.

4.11 Cumulative Distribution Functions

Definition 4.19 (Cumulative Distribution Function)**.** The cumulative distribution function of a random variable X is a function F such that

$$F(a) = \mathbb{P}(X \leq a), a \in (-\infty, \infty) \tag{4.178}$$

The CDF is a single function that encodes everything about the distribution of X. For discrete random variables, F is a staircase function: it jumps at each value X can take, and is flat in between. For discrete random variables,

$$F(a) = \sum_{x \leq a} \mathbb{P}(X = x) \tag{4.179}$$

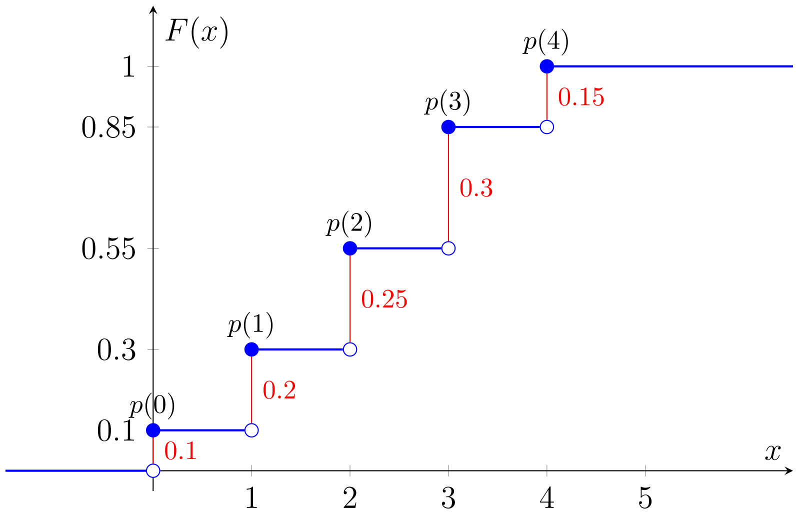


Figure 12: CDF of a discrete random variable X with support $\{0, 1, 2, 3, 4\}$.

Cumulative distribution functions must satisfy five key properties:

1. Bounded: $0 \leq F(a) \leq 1 \forall a$. Probabilities cannot be outside $[0, 1]$
2. Non-decreasing: If $a \leq b$ then $F(a) \leq F(b)$. As you slide b to the right, you include more values, so the cumulative probability can only go up (or stay the same). It can never decrease.
3. $\lim_{a \to \infty} F(a) = 1$. Eventually you capture all the probability mass
4. $\lim_{a \to -\infty} F(a) = 0$. Far enough to the left, nothing has accumulated yet.
5. F is right-continuous. If $a_n \downarrow a$ then $\lim_{n \to \infty} F(a_n) = F(a)$

A function F is right-continuous at a point a if

$$\lim_{x \downarrow a} F(x) = F(a) \tag{4.180}$$

In words: as you approach a from the right (from values slightly larger than a, the function values converge to $F(a)$. There's no gap between where the function is heading and where it actually lands. The CDF is right-continuous as it comes directly from the definition. $F(a) = \mathbb{P}(X \leq a)$. Thus if you have a decreasing sequence $a_n \downarrow a$ i.e. $a_1 > a_2 > \cdots$ where $a_n \to a$. Then, the following events are nested

$$\cdots \{X \leq a_3\} \subseteq \{X \leq a_2\} \subseteq \{X \leq a_1\} \tag{4.181}$$

Each event is contained in the previous one, because as a_n decreases, fewer values of X satisfy the inequality.

The CDF is important as every probability question about X can be answered from F

$$\mathbb{P}(a < X \leq b) = F(b) - F(a) \tag{4.182}$$

$$\mathbb{P}(X > b) = 1 - F(b) \tag{4.183}$$

$$\mathbb{P}(X < b) = F(b^-) = \lim_{x \uparrow b} F(x) \tag{4.184}$$

$$\mathbb{P}(X = b) = F(b) - F(b^-) \tag{4.185}$$

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X = a) + P(a < X \leq b) \tag{4.186}$$

$$= F(a) - F(a^-) + F(b) - F(a) \tag{4.187}$$

$$= F(b) - F(a^-) \tag{4.188}$$

The last formula is particularly important: the probability that X equals a specific value is the size of the jump in F at that point. If F is continuous at b then $\mathbb{P}(X = b) = 0$. The distinction between $\leq$ and $<$ matters for discrete distributions. $F(b) = \mathbb{P}(X \leq b)$ includes the point b while $F(b^-) = \mathbb{P}(X < b)$ does not. The difference is exactly $P(X = b)$.

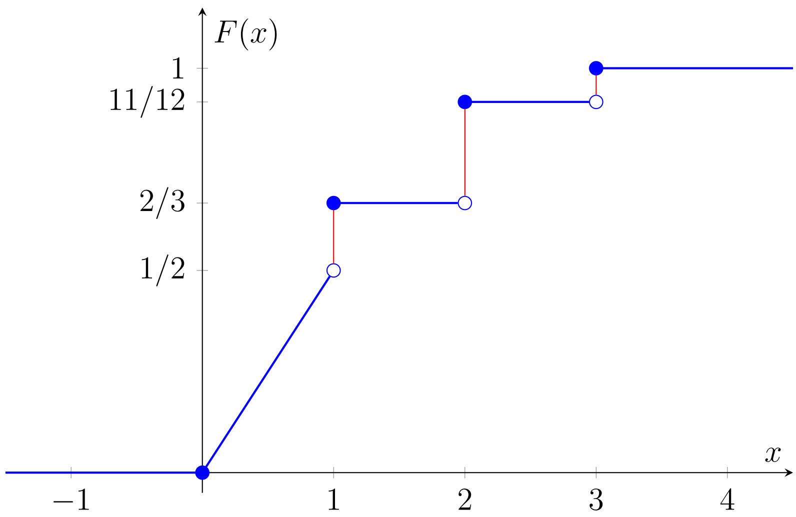


Figure 13: CDF $F(x)$ with a continuous piece on $[0, 1)$ and jumps at $x = 1, 2, 3$.

Example 27

Consider a random variable X described by CDF

$$F(x) = \begin{cases} 0, & x < 0, \\ x/2, & 0 \leqslant x < 1, \\ 2/3, & 1 \leqslant x < 2, \\ 11/12, & 2 \leqslant x < 3, \\ 1, & 3 \leqslant x. \end{cases} \tag{4.189}$$

Find $\mathbb{P}(X < 3), \mathbb{P}(X = 1), \mathbb{P}\left(X > \frac{1}{2}\right)$ and $\mathbb{P}(2 < X \leq 4)$

Solution

A graph of F is shown above. Now,

$$\mathbb{P}(X < 3) = F(3^-) \tag{4.190}$$
$$= \frac{11}{12} \tag{4.191}$$

$$\mathbb{P}(X = 1) = F(1) - F(1^-) \tag{4.192}$$
$$= \frac{2}{3} - \frac{1}{2} \tag{4.193}$$
$$= \frac{1}{6} \tag{4.194}$$

$$\mathbb{P}\left(X > \frac{1}{2}\right) = 1 - F\left(\frac{1}{2}\right) \tag{4.195}$$
$$= 1 - \frac{1}{4} \tag{4.196}$$
$$= \frac{3}{4} \tag{4.197}$$

$$\mathbb{P}(2 < X \leq 4) = F(4) - F(2) \tag{4.198}$$
$$= 1 - \frac{11}{12} \tag{4.199}$$
$$= \frac{1}{12} \tag{4.200}$$

5 Continuous Random Variables

5.1 Introduction to Continuous Random Variables

All the random variables and distributions we've seen so far have been discrete. Namely, their support is at most countable (a countable or countably infinite set of values). Further, they are completely characterised by their PMF $\mathbb{P}(X = k)$. We now discuss random variables whose support are uncountable. Their support can be intervals like $(-\infty, \infty), (0, 1)$ or $(0, \infty)$. Real-world examples include the time until the next earthquake, the height of a randomly selected person, or the exact temperature at noon.

For discrete random variables, we could assign probabilities to each value in its support $\mathbb{P}(X = k)$. We cannot do the same for continuous random variables. Suppose X is uniformly distributed on $[0, 1]$. By symmetry, every point should have the same probability, say $\mathbb{P}(X = x) = c \forall x \in [0, 1]$. If $c > 0$ then summing over uncountably many points would give infinite total probability which is invalid. If $c = 0$ then every point has probability zero, which is fine — but it means $\mathbb{P}(X = x) = 0$ which tells us nothing useful. To resolve this, we assign probabilities to intervals instead.

Definition 5.1 (Continuous Random Variables)**.** X is a continuous random variable if there is a non-negative function $f : \mathbb{R} \to [0, \infty)$ on $(-\infty, \infty)$ such that for any reasonable set B

$$\mathbb{P}(X \in B) = \int_B f(x)dx \tag{5.1}$$

for any set B of $(-\infty, \infty)$. The function f is called the probability density function of X

We define a reasonable set as a set that is Borel measurable (do not worry about what this means!). If we let $B = (-\infty, \infty)$ we have the normalisation condition

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

This is the continuous analogue of the following equation for discrete random variables.

$$\sum_{x \in \text{supp}(X)} f_X(x)dx = 1 \tag{5.3}$$

If we let $B = [a, b]$, then we get the probability of an interval to be

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx \tag{5.4}$$

The probability of a single point is

$$\mathbb{P}(X = a) = \int_{\{a\}} f(x)dx = 0 \tag{5.5}$$

This fact implies that for continuous random variables, unlike discrete random variables, it doesn't matter whether we use strict or non-strict inequalities. Thus,

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b) \tag{5.6}$$

This is because the boundary point contributes to zero probability. Thus, the cumulative distribution function of a continuous random variable is defined as

$$F(a) = \mathbb{P}(X \leq a) = \mathbb{P}(X < a) = \int_{-\infty}^{a} f(x)dx \tag{5.7}$$

The relationship between the probability density function and the cumulative distribution function is given by the fundamental theorem of calculus,

$$F'(a) = \frac{dF}{dx}|_{x=a} = f(a) \tag{5.8}$$

This is the continuous analogue of the discrete relationship

$$\mathbb{P}(X = k) = F(k) - F(k^-) \tag{5.9}$$

The density $f(a)$ is not a probability, in fact it can be greater than 1. Instead, for small ϵ

$$\mathbb{P}\left(a - \frac{\epsilon}{2} < X < a + \frac{\epsilon}{2}\right) = \int_{a-\frac{\epsilon}{2}}^{a+\frac{\epsilon}{2}} f(x)dx \approx \epsilon f(a) \tag{5.10}$$

Thus, $f(a)$ measures the probability per unit length near a

Example 28

Suppose X is a continuous random variable whose probability density function is given by $f(x) = C(4x - 2x^2)$ for $0 < x < 2$. Find C and $\mathbb{P}(X > 1)$.

Solution

$$\int_0^2 C(4x - 2x^2)dx = \frac{8C}{3} = 1 \tag{5.11}$$

$$\implies C = \frac{3}{8} \tag{5.12}$$

$$\mathbb{P}(X > 1) = \int_1^2 \frac{3}{8}(4x - 2x^2)dx = \frac{1}{2} \tag{5.13}$$

Example 29

If X is a continuous random variable with PDF f_X and CDF F_X. Find the probability density of $Y = X^2$

Solution

We first find the CDF and differentiate to find the PDF. Thus,

$$G_Y(y) = \mathbb{P}(Y \leq y) \tag{5.14}$$
$$= \mathbb{P}(X^2 \leq y) \tag{5.15}$$
$$= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \tag{5.16}$$
$$= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \tag{5.17}$$

Thus,

$$g_Y(y) = G'_Y(y) \tag{5.18}$$
$$= \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}} \tag{5.19}$$

Definition 5.2. The expectation of a continuous random variable X is defined as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{5.20}$$

This is precisely the continuous version of the discrete analogue of expectation where the sum is now replaced by an integral.

Theorem 5.1 (The Law of the Unconscious Statistician)**.** The Law of the Unconscious Statistician says that if X is a continuous random variable with probability density f and g is a function, then

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx \tag{5.21}$$

This is useful because we do not have to find the density of $g(X)$. From this law, we get the variance of a continuous random variable X to be

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 && (5.22)\\ &= \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} x f(x)dx\right)^2 && (5.23)\end{aligned}$$

The same results for expectation and variance for discrete RV's hold true for continuous RV's

$$\begin{aligned}\mathbb{E}(aX+b) &= a\mathbb{E}(X)+b && (5.24)\\ \mathrm{Var}(aX+b) &= a^2\mathrm{Var}(X) && (5.25)\end{aligned}$$

5.2 Uniform Random Variables

Definition 5.3 (Uniform Random Variable)**.** A continuous RV X is a Uniform Random Variable if it has a PDF

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases} \tag{5.26}$$

We denote a uniform random variable as $X \sim U(a,b)$

Uniform random variables follow a key property called shift invariance. Namely,

$$\mathbb{P}(x_1 < X < x_2) = \mathbb{P}(x_1 + h < X < x_2 + h) = \frac{x_2 - x_1}{b-a} \tag{5.27}$$

as long as both intervals lie within (a,b). Namely, the probability of an interval depends only on the length of the interval and not the actual interval itself.

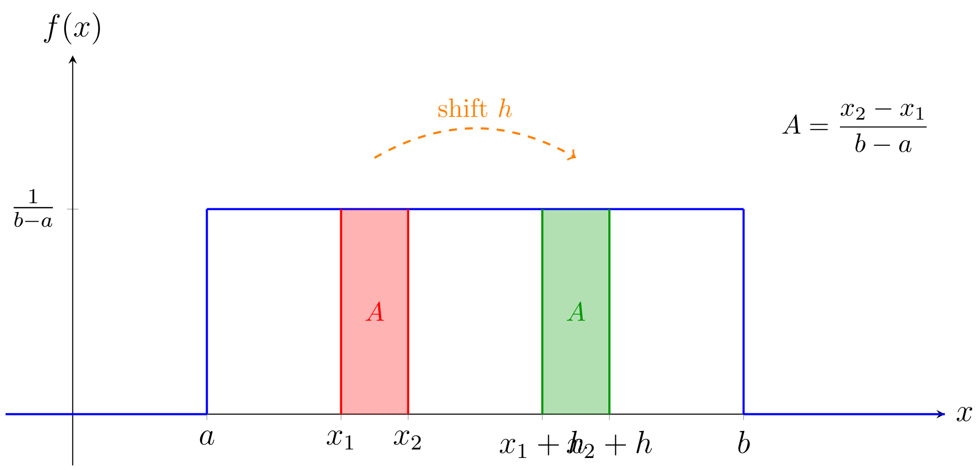


Figure 14: Shift invariance of the uniform distribution: both shaded regions have equal area $P(x_1 < X < x_2) = P(x_1 + h < X < x_2 + h) = \frac{x_2 - x_1}{b-a}$.

Lemma 5.2. For a uniform random variable $X \sim U(a, b)$

$$\mathbb{E}(X) = \frac{a+b}{2} \tag{5.28}$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12} \tag{5.29}$$

Remark. The mean is exactly the midpoint of the uniform interval. The variance depends only on the length of the interval.

Proof.

$$\mathbb{E}(X) = \int_a^b \frac{x}{b-a} dx \tag{5.30}$$

$$= \frac{1}{b-a}\left(\frac{b^2 - a^2}{2}\right) \tag{5.31}$$

$$= \frac{b+a}{2} \tag{5.32}$$

$$\mathrm{Var}(X) = \int_a^b \frac{x^2}{b-a} dx - \left(\frac{b+a}{2}\right)^2 \tag{5.33}$$

$$= \frac{1}{b-a}\left(\frac{b^3 - a^3}{3}\right) - \frac{a^2 + b^2 + 2ab}{4} \tag{5.34}$$

$$= \frac{(b-a)^2}{12} \tag{5.35}$$

■

Example 30

A bus travels between two cities A and B that are 100 miles apart. If the bus has a breakdown, the distance from the breakdown to city A is distributed as $X \sim U(0, 100)$. Three service stations are available. Configuration 1 (S_1)says that service stations are available at 0, 50, and 100 miles from city A. Configuration 2 (S_2) says that service stations are available at 25, 50, 75 miles from city A. Which configuration is more efficient?

Solution

Let $g_i(u)$ be the distance to the nearest service station for configuration i. Then,

$$\mathbb{E}(g_1(u)) = \tag{5.36}$$

Example 31

A stick of length 1 is split at a random point $U \sim U(0, 1)$. We're given a fixed point $p \in (0, 1)$. What is the expected value of the length of the piece that contains p?

Solution

The split creates two pieces. Left piece has length U from 0 to U. The right piece has length $1 - U$ for U to 1. There are two cases depending on where the split U happens relative to p. If $U < p$, then p is in the $1 - U$ regime. If $U \geq p$ then p is in the U regime. Thus, the length of the piece containing p is

$$g(U) = \begin{cases} 1 - U, & U < p \\ U, & U \geq p \end{cases} \tag{5.37}$$

Thus,

$$\mathbb{E}(g(U)) = \int_0^p 1 - U dU + \int_p^1 U dU = p - p^2 + \frac{1}{2} \tag{5.38}$$

5.3 Normal Random Variables

Definition 5.4 (Normal Random Variables)**.** A continuous random variable X is a normal random variable with parameters μ and σ^2 if its density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } x \in (-\infty, \infty) \tag{5.39}$$

We denote it as $X \sim N(\mu, \sigma^2)$

A normal distribution looks like a bell curve whose centre is at $x = \mu$ and whose width is σ. Normal distributions are important as many quantities follow a normal distribution

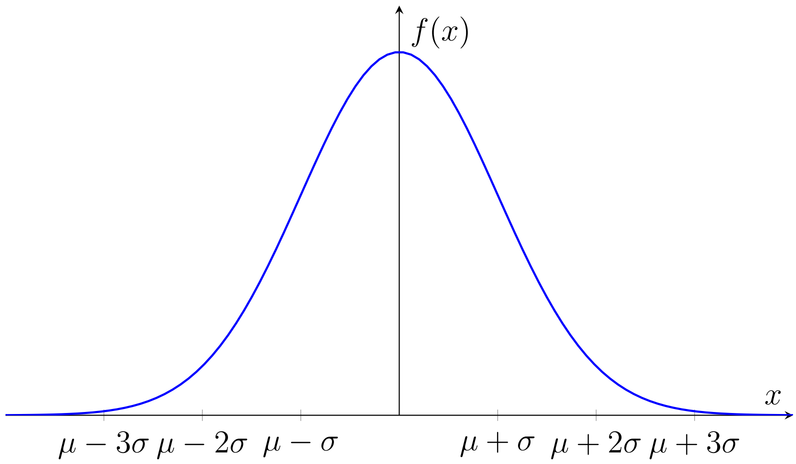


Figure 15: Normal distribution $\mathcal{N}(\mu, \sigma^2)$.

like heights, measurement errors, test scores etc. The Poisson distribution was related to the number of occurences of at most weakly dependent events, each with a small chance of occurence. The normal distribution is related to the sum of weakly dependent variables, each with a moderate chance of occurence.

The prefactor $\frac{1}{\sqrt{2\pi\sigma^2}}$ is used to normalise so that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1 \tag{5.40}$$

The standard normal distribution is the distribution $N \sim (0, 1)$ where $\mu = 0$ and $\sigma^2 = 1$.

It shifts the distribution to center it at 0 and scales it to have unit variance. This means all computations involving any $N(\mu, \sigma^2)$ can be reduced to computations involving $N(0, 1)$. This is important since the density of the standard normal distribution φ is significantly easier to work with.

$$\varphi(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{5.41}$$

If we have a random variable $X \sim N(\mu, \sigma^2)$, we can reduce it to its standard normal via

$$Z = \frac{X - \mu}{\sigma} \tag{5.42}$$

Then,

$$Z \sim N(0, 1) \tag{5.43}$$

Shifting to the standard normal is useful because we can make use of the standard integral result

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi} \tag{5.44}$$

Normal distributions satisfy a special property wherein they are closed under linear transformations. Namely, if $X \sim N(\mu, \sigma^2)$ then

$$aX + b \sim N(a\mu + b, a^2\sigma^2) \tag{5.45}$$

Thus, applying a linear transformation to a normal distribution results in another normal distribution. Very few probability distributions follow this result. In fact, the standardising $Z = \frac{X-\mu}{\sigma}$ is just a linear transformation with $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$

Lemma 5.3. If $X \sim N(\mu, \sigma^2)$ then,

$$\mathbb{E}(X) = \mu \tag{5.46}$$

$$\text{Var}(X) = \sigma^2 \tag{5.47}$$

Proof.

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{5.48}$$

Let $z = \frac{x-\mu}{\sigma} \implies x = \mu + \sigma z$. Then,

$$\begin{aligned}
\mathbb{E}(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z) e^{-\frac{z^2}{2}} dz && (5.49)\\
&= \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz + \frac{\sigma}{\sqrt{\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz && (5.50)\\
&= \frac{\mu}{\sqrt{2\pi}} \cdot \sqrt{2\pi} + 0 && (5.51)\\
&= \mu + 0 && (5.52)\\
&= 0 && (5.53)
\end{aligned}$$

where the second integral is 0 since we have an odd function integrated over a symmetric interval. Now,

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} x^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx - \mu^2 \tag{5.54}$$

Let $z = \frac{x-\mu}{\sigma} \implies x^2 = \mu^2 + 2\mu\sigma z + \sigma^2 z^2$. Thus,

$$\begin{aligned}
\mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{\infty} \mu^2 e^{-\frac{z^2}{2}} dz + 2\mu\sigma \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz + \sigma^2 \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz \right) - \mu^2 && (5.55)\\
&= \frac{1}{\sqrt{2\pi}} \left(\mu^2 \sqrt{2\pi} + 0 + \sigma^2 \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz \right) - \mu^2 && (5.56)
\end{aligned}$$

We use integration by parts on the integral inside the parentheses. Let $u = z \implies du = 1$ and $dv = ze^{-\frac{z^2}{2}} \implies v = -e^{-\frac{z^2}{2}}$. Then,

$$\begin{aligned}
\int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz &= \left(-ze^{-\frac{z^2}{2}} \right)_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dx && (5.57)\\
&= 0 + \sqrt{2\pi} && (5.58)
\end{aligned}$$

since the boundary term vanishes as $ze^{-\frac{z^2}{2}} \to 0$ as $z \to \pm\infty$ Thus,

$$\mathrm{Var}(X) = \frac{1}{\sqrt{2\pi}} \left(\mu^2\sqrt{2\pi} + \sigma^2\sqrt{2\pi}\right) - \mu^2 \tag{5.59}$$

$$= \mu^2 + \sigma^2 - \mu^2 \tag{5.60}$$

$$= \sigma^2 \tag{5.61}$$

■

For $X \sim N(\mu, \sigma^2)$

1. About 68.2% of the probability lies within 1 standard deviation ($\mu \pm \sigma$)
2. About 95.4% of the probability lies within 2 standard deviations ($\mu \pm 2\sigma$)
3. About 99.7% of the probability lies within 3 standard deviations ($\mu \pm 3\sigma$)

Thus, anything beyond 3 standard deviations of the mean happens less than 0.3% of the time The cumulative distribution function of the standard normal distribution is defined

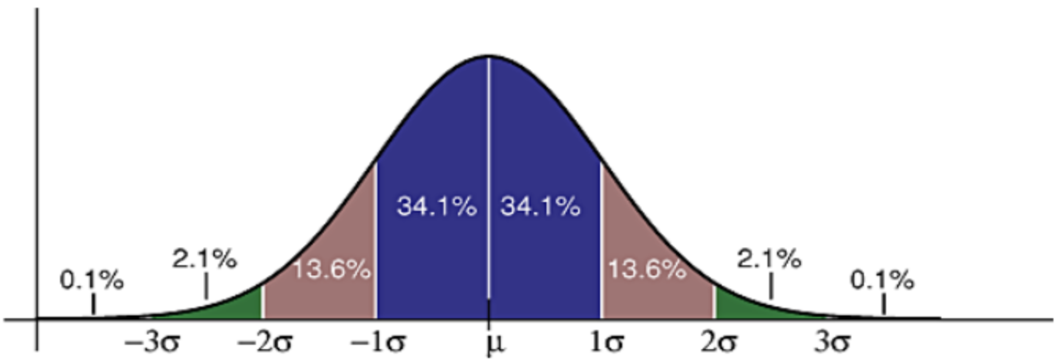


Figure 16: Normal distribution and probabilities

as

$$\Phi(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy \tag{5.62}$$

This has no closed form expression. Thus, we typically use a table of values

According to the table,

$$\mathbb{P}(Z \leq 1.14) = \Phi(1.14) = 0.8729 \tag{5.63}$$

For $\mathbb{P}(X > a)$ we have

$$\mathbb{P}(X > a) = 1 - \mathbb{P}(X \leq a) = 1 - \Phi(a) \tag{5.64}$$

For negative values we use symmetry,

$$\Phi(-a) = 1 - \Phi(a) \tag{5.65}$$

Then,

$$\mathbb{P}(a < Z < b) = \Phi(b) - \Phi(a) \tag{5.66}$$

To compute probabilities using the standard normal, we typically normalise first. Thus, if $X \sim N(\mu, \sigma^2)$,

$$\mathbb{P}(X \leq a) = \mathbb{P}\left(Z \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{5.67}$$

An important lemma comes from the standard normal density $\varphi(x)$

Lemma 5.4 (Stein's Lemma)**.** For any differentiable function g,

$$\mathbb{E}(g'(X)) = \mathbb{E}(Xg(X)) \tag{5.68}$$

Proof. By law of the unconscious statistician and using the standard normal density,

$$\mathbb{E}(g'(X)) = \int_{-\infty}^{\infty} g'(x)\varphi(x)dx \tag{5.69}$$

Use integration by parts where $u = g(x) \implies du = g'(x)$ and the product rule $\varphi' =$

$-x\varphi(x)$

$$\int_{-\infty}^{\infty} g'(x)\varphi(x)dx = [g(x)\varphi(x)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} g(x)\varphi'(x)dx \tag{5.70}$$

$$= 0 + \int_{-\infty}^{\infty} xg(x)\varphi(x)dx \tag{5.71}$$

$$= \mathbb{E}[Xg(X)] \tag{5.72}$$

■

Thus, if any random variable X satisfies $\mathbb{E}(g'(X)) = \mathbb{E}(Xg(X))$ then X must the standard normal distribution $X \sim N(0,1)$. No other distribution satisfies this identity.

Example 32

Suppose $X \sim N(3,9)$. Find $\mathbb{P}(2 < X < 5)$, $\mathbb{P}(X > 0)$ and $\mathbb{P}(|X-3| > 6)$

Solution

$$\begin{aligned}
\mathbb{P}(2 < X < 5) &= \mathbb{P}\left(\frac{2-3}{3} < Z < \frac{5-3}{3}\right) && (5.73)\\
&= \mathbb{P}\left(-\frac{1}{3} < Z < \frac{2}{3}\right) && (5.74)\\
&= \Phi(0.667) - (1 - \Phi(0.333)) && (5.75)\\
&= \Phi(0.667) + \Phi(0.333) - 1 && (5.76)\\
&= 0.7486 + 0.6293 - 1 && (5.77)
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(X > 0) &= \mathbb{P}\left(Z > \frac{0-3}{3}\right) && (5.78)\\
&= \mathbb{P}(Z > -1) && (5.79)\\
&= 1 - \mathbb{P}(Z < -1) && (5.80)\\
&= 1 - (1 - \Phi(1)) && (5.81)\\
&= 1 - (1 - 0.8413) && (5.82)\\
&= 0.84143 && (5.83)
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(|X-3| > 6) &= \mathbb{P}(X - 3 < -6) + \mathbb{P}(X - 3 > 6) && (5.84)\\
&= \mathbb{P}(X < -3) + \mathbb{P}(X > 9) && (5.85)\\
&= \mathbb{P}\left(Z < -2\right) + \mathbb{P}(Z > 2) && (5.86)\\
&= 1 - \Phi(2) + 1 - \Phi(2) && (5.87)\\
&= 2(1 - \Phi(2)) && (5.88)\\
&= 2(1 - 0.9772) && (5.89)\\
&= 0.0456 && (5.90)
\end{aligned}$$

It turns out we can approximate the binomial distribution to the standard normal distribution. Let $X \sim \text{Bin}(n, p)$. We know that

$$\mathbb{E}(X) = np \tag{5.91}$$

$$\text{Var}(X) = np(1-p) \tag{5.92}$$

The standardised Binomial is given by

$$S_n = \frac{S_n - np}{\sqrt{np(1-p)}} \tag{5.93}$$

For large n, the standardised Binomial is approximately a standardised normal distribution,

$$S_n = \frac{X - np}{\sqrt{np(1-p)}} \approx N(0,1) \tag{5.94}$$

This is typically a good approximation when $np(1-p) \geq 10$. We can contrast this with the Poisson distribution. The Poisson distribution works well when n is large, p is small and $\lambda = np$ is moderate. The Normal approximation works well when n is large. p doesn't necessarily have to be small. However, if we naively try to compute $\mathbb{P}(X = k)$ we have

$$\mathbb{P}(S = k) = \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} = \frac{k - np}{\sqrt{np(1-p)}}\right) \approx \mathbb{P}\left(Z = \frac{k-np}{\sqrt{np(1-p)}}\right) \tag{5.95}$$

However, this is 0 since the Normal distribution is continuous and so the probability at a point is not defined. However X is discrete and so $\mathbb{P}(X = k)$ is defined. Thus, we add a continuity correction. Since X is always an integer, the event $X = k$ is the same as $k - \frac{1}{2} < X < k + \frac{1}{2}$. Thus,

$$\mathbb{P}(X = k) = \mathbb{P}\left(k - \frac{1}{2} < X < k + \frac{1}{2}\right) \approx \mathbb{P}\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}} < Z < \frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \tag{5.96}$$

This produces a small but non-zero probability, which is a much better approximation.

Example 33

Each item produced by a manufacturer is, independently, of acceptable quality with probability 0.95. Approximate the probability that at most 10 of the next 150 items produced are unacceptable.

Solution

Let X be the number of unacceptable items. Thus, $X \sim \text{Bin}(150, 0.05)$. Then,

$$\mathbb{E}(X) = np = 7.5 \tag{5.97}$$

$$\text{Var}(X) = np(1-p) = 7.5(0.95) = 7.125 \tag{5.98}$$

$$\mathbb{P}(X \leq 10) = \mathbb{P}\left(Z \leq \frac{10 - 7.5}{\sqrt{7.125}}\right) \tag{5.99}$$

$$= \mathbb{P}(Z \leq 0.94) \tag{5.100}$$

$$= \Phi(0.94) \tag{5.101}$$

$$= 0.8264 \tag{5.102}$$

Example 34

The ideal size of a first year class at a particular college is 150 students. The college, knowing from past experience that, on the average, only 30 percent of those accepted for admission will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 first-year students attend this college.

Solution

Let X be the number of students that accept. Thus, $X \sim \text{Bin}(450, 0.3)$. Then,

$$\mathbb{E}(X) = np = 450(0.3) = 135 \tag{5.103}$$

$$\text{Var}(X) = np(1-p) = 135(0.7) = 94.5 \tag{5.104}$$

Thus,

$$\mathbb{P}(X > 150) = \mathbb{P}\left(Z > \frac{150-135}{\sqrt{94.5}}\right) = \mathbb{P}(Z > 1.54) = 1-\Phi(1.54) = 1-0.9382 = 0.0618 \tag{5.105}$$

If we use the continuity correction we have

$$\begin{aligned}
\mathbb{P}(X > 150) &= \mathbb{P}(X \geq 151) && (5.106)\\
&= \mathbb{P}(X > 150.5) && (5.107)\\
&= \mathbb{P}\left(Z > 1.59\right) && (5.108)\\
&= 1 - \Phi(1.59) && (5.109)\\
&= 1 - 0.9441 && (5.110)\\
&= 0.0559 && (5.111)
\end{aligned}$$

Thus, the continuity correction is slightly more accurate.

5.4 Exponential Random Variables

Definition 5.5 (Exponential Random Variable)**.** A continuous random variable X is an exponential random variable with parameter $\lambda > 0$ if it has density

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{5.112}$$

We denote this as $X \sim \text{Exp}(\lambda)$

The density starts at $f(0) = \lambda$ and decays exponentially where a larger λ means a faster decay. The density is more concentrated near zero, so the random variable tends to take smaller values.

This is a valid probability density function since

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\infty} \lambda e^{-\lambda x} dx \tag{5.113}$$

$$= -[e^{-\lambda x}]_0^\infty \tag{5.114}$$

$$= -[0 - 1] \tag{5.115}$$

$$= 1 \tag{5.116}$$

For $a \geq 0$, the CDF is given by

$$F(a) = \int_0^a \lambda e^{-\lambda x} dx \tag{5.117}$$

$$= -[e^{-\lambda x}]_0^a \tag{5.118}$$

$$= -[e^{-\lambda a} - 1] \tag{5.119}$$

$$= 1 - e^{-\lambda a} \tag{5.120}$$

The tail of the exponential distribution is given by

$$\mathbb{P}(X > a) = 1 - F(a) = e^{-\lambda a} \tag{5.121}$$

The exponential decay of the tail is a defining feature of the distribution.

Lemma 5.5. If $X \sim \text{Exp}(\lambda)$ then,

$$\mathbb{E}(X) = \frac{1}{\lambda} \tag{5.122}$$

$$\text{Var}(X) = \frac{1}{\lambda^2} \tag{5.123}$$

Proof.

$$\mathbb{E}(X) = \int_0^\infty \lambda x e^{-\lambda x} dx \tag{5.124}$$

Let $u = \lambda x \implies du = \lambda dx$. Then,

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty \frac{u}{\lambda} e^{-u} du &(5.125)\\ &= \frac{1}{\lambda}\int_0^\infty u e^{u} du &(5.126) \end{aligned}$$

Use integration by parts where $\alpha = u$ and $dv = e^{-u}$. Thus, $d\alpha = 1$ and $v = -e^{-u}$. Then,

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{\lambda}\left([-ue^{-u}]_0^\infty - \int_0^\infty e^{-u} du\right) &(5.127)\\ &= \frac{1}{\lambda}[0-(-1)] &(5.128)\\ &= \frac{1}{\lambda} &(5.129) \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{Var}(X) &= \int_0^\infty \lambda x^2 e^{-\lambda x} dx - \frac{1}{\lambda^2} &(5.130)\\ &= \left([-x^2 e^{-\lambda x}]_0^\infty + 2\int_0^\infty x e^{-\lambda x} dx\right) - \frac{1}{\lambda^2} &(5.131)\\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} &(5.132)\\ &= \frac{1}{\lambda^2} &(5.133) \end{aligned}$$

■

In practice, the exponential distribution is often the time until some event (like a phone call or earthquake occurs). The exponential distribution, like the geometric distribution, also exhibits memorylessness. Recall that

$$\mathbb{P}(X > t) = e^{-\lambda t} \tag{5.134}$$

Thus,

$$\mathbb{P}(X > t + s | X > t) = \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > t)} \tag{5.135}$$
$$= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} \tag{5.136}$$
$$= e^{-\lambda s} \tag{5.137}$$

Thus, conditionally on $X > t$, the remaining waiting time is a fresh exponential distribution. In other words, given that you've already waited t units with no event, the remaining waiting time has the same $Exp(\lambda)$ distribution as if you just started waiting. The past is irrelevant. The system has no memory of how long you've been waiting. The exponential distribution is the only continuous distribution with the memorylessness property.

Example 35

A car battery's lifetime (in miles) is distributed exponentially with a mean of 10,000 miles. What is the probability of completing a 5,000-mile trip?

Example 36

Let X be the car battery's lifetine in miles. Thus,

$$\mathbb{E}(X) = \frac{1}{\lambda} = 10000 \tag{5.138}$$
$$\implies \lambda = \frac{1}{10000} \tag{5.139}$$

Thus, $X \sim \text{Exp}\left(\frac{1}{10000}\right)$. Then,

$$\mathbb{P}(X > 5000) = e^{-\frac{5000}{10000}} \tag{5.140}$$
$$= \frac{1}{\sqrt{e}} \tag{5.141}$$
$$\approx 0.607 \tag{5.142}$$

If the car had already been drive 3000 miles, the probability that the car lasts another

5000 miles is still

$$e^{-0.5} \approx 0.607 \tag{5.143}$$

We can see that the exponential distribution is not a good model for a car's battery lifetime!

5.5 Gamma Random Variables

Definition 5.6 (Gamma Random Variable)**.** A continuous random variable X is a Gamma random variable with parameters α, λ if it has density

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & x >= 0 \\ 0, & x < 0 \end{cases} \tag{5.144}$$

where $\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx$ is the Gamma function. We denote this as $X \sim \Gamma(\alpha, \lambda)$

When $\alpha = 1$, we get the exponential distribution. Thus,

$$\Gamma(1, \lambda) = \text{Exp}(\lambda) \tag{5.145}$$

The Gamma distribution has a beautiful interpretation in the context of Poisson processes. If events arrive according to a Poisson process with rate λ (meaning the number of events in any time interval of length t is Poi() then

1. The time until the first event is distributed as $\Gamma(1, \lambda) = \text{Exp}(\lambda)$
2. The time until the n^{th} event is distributed as $\Gamma(n, \lambda)$

So $\Gamma(n, \lambda)$ is the waiting time for n events in a Poisson process. This is the continuous analogue of the Negative Binomial, which counts the number of trials until n successes in a discrete Bernoulli process.

Discrete Bernoulli trials	**Continuous** Poisson process
Geo(p): trials until 1st success	Exp(λ): time until 1st event
NegBin(n, p): trials until n-th success	$\Gamma(n, \lambda)$: time until n-th event

Table 1: Discrete and continuous analogues for counting successes/events.

5.6 Other Continuous Random Variables

Weibull Random Variable

Definition 5.7 (Weibull Random Variable)**.** A continuous random variable X is a Weibull random variable with parameters ν, $\alpha > 0, \beta > 0$ if it has cumulative distributive function

$$F(x) = \begin{cases} 1 - e^{-\left(\frac{x-\nu}{\alpha}\right)^{\beta}}, & x \geq \nu \\ 0, & x < \nu \end{cases} \tag{5.146}$$

Its density is given by differentiation

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\nu}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x-\nu}{\alpha}\right)^{\beta}} \tag{5.147}$$

The Weibull is used extensively in reliability engineering to model lifetimes of devices, components, and individuals. Its power comes from the shape parameter β

1. $\beta < 1$: The failure rate decreases over time. This models "infant mortality" — if a device survives its initial period, it becomes more reliable. Think of manufacturing defects that cause early failures.

2. $\beta = 1$: The failure rate is constant. The Weibull reduces to the Exponential (with $\nu = 0$ this gives $F(x) = 1 - e^{-x/\alpha}$ which is Exp $\left(\frac{1}{\alpha}/\right)$. This is the memoryless case

3. $\beta > 1$: The failure rate increases over time. This models ageing and wear — the longer a device has been running, the more likely it is to fail soon. This is the most realistic model for mechanical components.

The Exponential, with its constant failure rate, is a special case that sits right at the boundary between improving and deteriorating reliability.

Beta Random Variables

Definition 5.8 (Beta Random Variables)**.** A continuous random variable X is a Beta random variable with parameters a, b if it has density

$$f(x) = \begin{cases} \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{5.148}$$

where $B(a, b) = \int_0^1 x^{(a-1)}(1-x)^{b-1}dx$ is the Beta function (which serves as the normalisation constant). We denote it as $X \sim B(a, b)$

The Beta distribution models random phenomena that take values in a bounded interval — specifically $[0, 1]$. This makes it natural for modeling probabilities, proportions, percentages, and rates. For example: the fraction of defective items in a batch, the probability that a customer clicks an ad, or the proportion of time a server is busy. The parameters a and b control the shape, making the Beta distribution extremely versatile.

1. $a = b = 1$: $f(x) = 1$ which is $U(0, 1)$. The Uniform is a special case of the Beta! The density is flat — all values in $[0, 1]$ are equally likely.

2. $a = b > 1$: Symmetric, bell-shaped, centered at $\frac{1}{2}$. As $a = b$ grows, the distribution concentrates more tightly around $\frac{1}{2}$

3. $a > b$: Skewed towards 1. The distribution favours larger values.

4. $a < b$: Skewed toward 0. The distribution favours smaller values.

5. $a, b < 1$: U-shaped. The distribution concentrates near 0 and 1, with little mass in the middle.

5.7 Mixtures of Random Variables

Recall that any function F_X that satisfies

1. F_X is bounded
2. F_X is non-decreasing
3. $\lim_{x\to\infty} F_X = 1$
4. $\lim_{x\to-\infty} F_X = 0$
5. F_X is right-continuous

can define a valid random variable. Conversely, given any function with these four properties, we can construct a random variable X such that $\mathbb{P}(X \in (a, b)) = F_X(b) - F_X(a)$. For discrete random variables, F_X is a staircase function with jumps while for continuous random variables, F_X is differentiable and $\mathbb{P}(X \in (a, b)) = \int_a^b f_X(x)dx$ where $f_X = F_X'$. In general, a probability CDF can be a mixture of both discrete and continuous components. The CDF might have jumps (atoms) in some places and smooth continuous pieces elsewhere. For example, this is a valid CDF

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.3 + 0.7x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases} \tag{5.149}$$

We can verify $\mathbb{P}(X \le t) = 0$ for $t < 0$, $\mathbb{P}(X = 0) = F(0) - F(0^-) = 0.3 - 0 = 0.3$ and for $0 < a < b \le 1$

$$\mathbb{P}(X \in (a, b)) = \int_a^b 0.3 + 0.7x dx \tag{5.150}$$

The continuous part has density $f(x) = 0.7$ on $[0, 1]$. This is 0.7 times $U(0, 1)$. So X is a mixture: with probability 0.3 it equals exactly 0, and with probability 0.7 it's drawn from $U(0, 1)$. Another real-world example of a mixed CDF is waiting in a queue. When you arrive, with probability p the queue is empty and you are served immediately

$(X = 0)$ and with probability $1 - p$, the queue is busy and your wait time is distributed perhaps exponentially. Thus, the CDF has a jump of size of p at $x = 0$ and then increases smoothly for $x > 0$. This cannot be described by a pmf alone or a pdf alone, we you need the CDF framework to handle both components simultaneously.

TABLE 5.1: AREA $\Phi(x)$ UNDER THE STANDARD NORMAL CURVE TO THE LEFT OF X

X	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Figure 17: Standard normal table. The row gives the first decimal place. The column gives the second decimal place.

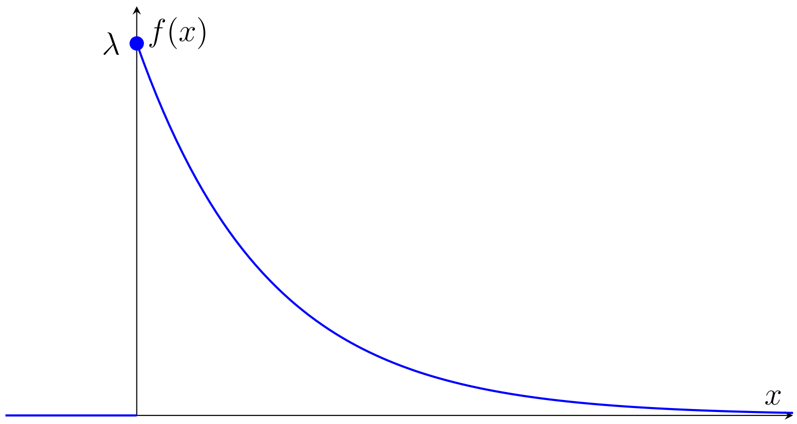


Figure 18: PDF of the exponential distribution Exp(λ).

Below: $\nu = 0$, $\beta = c$, $\alpha = 1$.

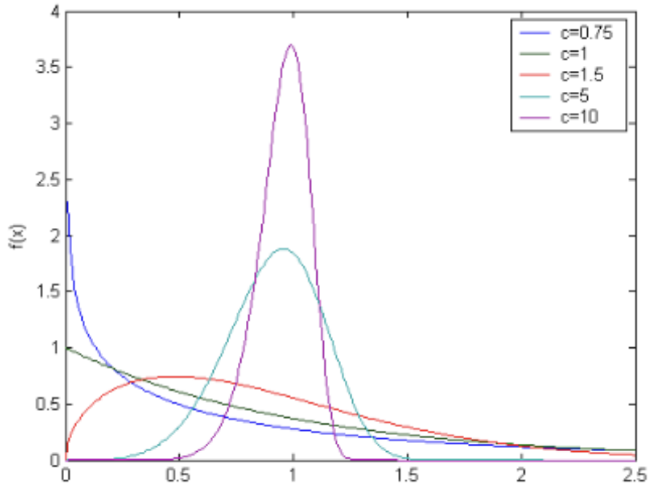


Figure 19: Weibull Distribution

6 Jointly Distributed Random Variables

6.1 Introduction to Jointly Distributed Random Variables

Everything we have covered so far has involved studying isolated random variables. However, we now begin to study collections of two or more random variables. In practice, we almost never encounter isolated random variables. Real phenomena involve multiple interacting quantities:

- X_1 could model the price of one stock, X_2 could model the price of another stock and we want to understand how they are correlated
- X_1 could model cholestorol, X_2 could model blood pressure, and we want to understand how they are correlated
- X_1 could model rainfall in Illinois, X_2 could model rainfall in Indiana, and we want to understand how weather is linked across regions

Most advanced topics in statistics (time series, multivariate analysis, regression, factor models) and probability (Markov chains, stochastic processes) fundamentally involve collections of random variables. Understanding joint distributions is the foundation for all of these.

It is important to understand that all random variables from a random experiment live on the same probability space. They are all functions $X_1, X_2, \ldots, X_n : S \to \mathbb{R}$ possibly dependent on each other. For example, suppose we toss a coin n times where the probability of getting a heads is p. Then, if X is the number of heads we get and Y is the number of tails we get, we have

$$X \sim \text{Bin}(n, p) \tag{6.1}$$

$$Y \sim \text{Bin}(n, 1-p) \tag{6.2}$$

It will always hold that $X + Y = n$. They are completeley dependent as Y is determined entirely by X.

The joint cumulative distribution function of two random variables X and Y is defined as

$$F(a, b) = \mathbb{P}(X \leq a, Y \leq b) = \mathbb{P}(X \leq a \cap Y \leq b), -\infty < a, b < \infty \tag{6.3}$$

The condition $X \leq a$ restricts you to everything to the left of the vertical line $x = a$. The condition $Y \leq b$ restricts you to everything below the horizontal line $y = b$ The intersection of these two regions is the lower-left rectangle $(-\infty, a] \times [-\infty, b]$ The joint

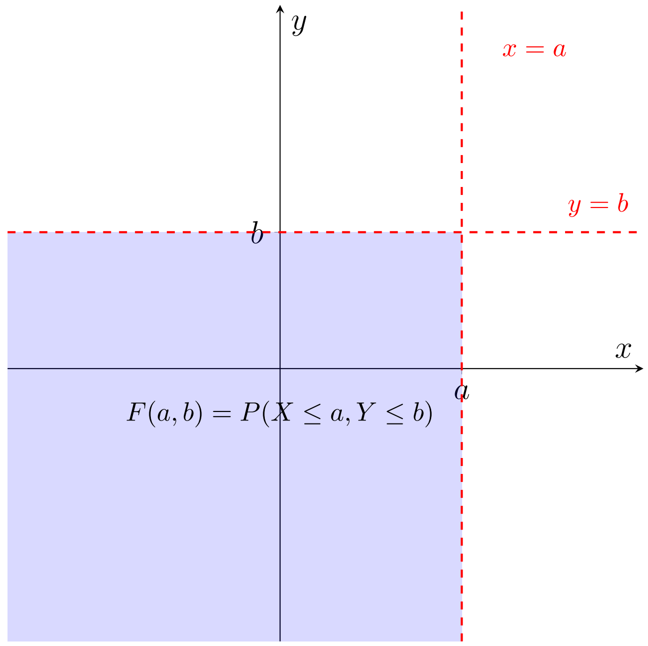


Figure 20: The joint CDF $F(a, b)$ is the probability mass in the shaded region $(-\infty, a] \times (-\infty, b]$.

CDF thus measures how much probability mass is in the lower-left quadrant determined by the point (a, b). This single function encodes all probabilistic information about X and Y together. All joint CDFs must satisfy the following property

1. $0 \leq F(x, y) \leq 1 \forall x, y$. This is immediate sincd $F(x, y)$ gives a probability

2. Boundary behaviour: $F(\infty, \infty) = 1$ and $F(x, -\infty) = F(-\infty, y) = F(-\infty, \infty) = 0$.

   . If you send either argument to $-\infty$ you're asking for the probability that the random variable is less than something impossibly small, which is zero. If both go to ∞ you're asking for the probability of the entire sample space, which is 1.

3. Monotonicity: F is non-decreasing in each argument separately. If you fix y and increase x, then F can only go up since you are increasing probability. The same is true vice-versa

These conditions are sufficient. If $\mathcal{F}$ is the class of functions that satisfies these properties then for any $F \in \mathcal{F}$, we can construct an $\mathbb{R}^2$-valued random variable whose joint CDF is exactly F. In other words, there's a one-to-one correspondence between joint CDFs and distributions of random points in the plane.

Example 37

Suppose you roll two die. Let X be the value of the first die and Y be the value of the second die. Find $\mathbb{P}(X \leq 3, Y \leq 2)$.

Example 38

Thus, we look for cases where the first die is either a 1,2, or 3 and the second die is either a 1 or 2. The cases are $(1, 1), (1, 2), (2, 1), (2, 2), (3, 1), (3, 2)$. Thus,

$$\mathbb{P}(X \leq 3, Y \leq 2) = \frac{1}{6} \tag{6.4}$$

The individual CDFs of X and Y are special cases of the joint CDF. To isolate X, we let $b \to \infty$ so the constraint on Y disappears. To isolate Y, we let $a \to \infty$ so the constraint on X disappears.

$$F_X(a) = \lim_{b\to\infty} F(a, b) = \mathbb{P}(X \leq a, Y < \infty) = \mathbb{P}(X \leq a) \tag{6.5}$$

$$F_Y(b) = \lim_{a\to\infty} F(a, b) = \mathbb{P}(X < \infty, Y \leq b) = \mathbb{P}(Y \leq b) \tag{6.6}$$

The complement via the joint CDF is given by the inclusion-exclusion principle

$$\mathbb{P}(X > a, Y > b) = 1 - \mathbb{P}(X \le a, Y \le b) \tag{6.7}$$
$$= 1 - \mathbb{P}(X \le a) - \mathbb{P}(Y \le b) + \mathbb{P}(X \le a, Y \le b) \tag{6.8}$$
$$= 1 - F_X(a) - F_Y(b) + F(a, b) \tag{6.9}$$

The probability that X, Y falls in a rectangle $(a_1, a_2] \times (b_1, b_2]$ is

$$\mathbb{P}(a_1 < X \le a_2, b_1 < Y \le b_2) = F(a_2, b_2) - F(a_1, b_2) - F(a_2, b_1) + F(a_1, b_1) \tag{6.10}$$

This is the 2D analogue of

$$\mathbb{P}(a < X \le b) = F(b) - F(a) \tag{6.11}$$

There are two main classes of joint distributions: jointly discrete and jointly continuous.

Jointly Discrete Random Variables

If we have two discrete random variables X and Y, they are characterised by their joint probability mass function

$$p(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x \cap Y = y) \tag{6.12}$$

The probability mass function of X, also known as the marginal pmf of X, is obtained by summing over all possible values of y and vice-versa

$$f_X(x) = \mathbb{P}(X = x) = \sum_y p(x, y) \tag{6.13}$$
$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x p(x, y) \tag{6.14}$$

This can be thought of as the discrete analogue of integrating out the other variable: to find the probability that$X = x$ regardless of what Y is, you add up $\mathbb{P}(X = x, Y = y)$

over all possible y values.

Jointly Continuous Random Variables

If we have two continuous random variables X and Y, they are jointly continuous if there is a non-negative function $f(x, y)$ such that for any region C in the plane

$$\mathbb{P}((X, Y) \in C) = \int\int_{(x,y)\in C} f(x, y)dxdy \tag{6.15}$$

This is the volume under the surface $z = f(x, y)$ above the region C. In the special case $f = 1$ (uniform density), this reduces to the area of C. If C is a cartesian product of two sets $C = A \times B = \{(x, y)|x \in A \cap y \in B\}$, the double integral factors

$$\mathbb{P}((X, Y) \in C) = \int_A dx \int_B dyf(x, y) \tag{6.16}$$

This does not mean $f(x, y) = f_X(x) \cdot f_Y(y)$. In general, the integral factors over the region, but the integrand might not. For the joint CDF

$$F(a, b) = \mathbb{P}(X \leq a, Y \leq b) \tag{6.17}$$

$$= \mathbb{P}((X, Y) \in (-\infty, a] \times (-\infty, b]) \tag{6.18}$$

$$= \int_{-\infty}^{a} dx \int_{-\infty}^{b} dyf(x, y) \tag{6.19}$$

the joint probability density is recovered by differentiating twice

$$f(a, b) = \frac{\partial^2}{\partial a \partial b} F(a, b) \tag{6.20}$$

Note that for small da, db, if $f(a, b)$ is continuous at (a, b) then

$$\mathbb{P}(a < X \leq a + da, b < Y \leq b + db) = \int_{a}^{a+da} dx \int_{b}^{b+db} dyf(x, y) \tag{6.21}$$

$$\approx f(a, b)dadb \tag{6.22}$$

Thus, $f(a, b)$ is a measure of how likely X, Y is near a, b.

Each indiviudal random variable is continuous. The probability density of X (also known as the marginal density) is obtained by integrating out Y and vice-versa

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \tag{6.23}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx \tag{6.24}$$

This is the continuous analogue of summing over the other variable. I.e., to find how much probability density is near a particular value of x regardless of y, you "collapse" the joint density along the y-axis by integrating. All the probability at different y gets accumulated into the marginal density at x.

Example 39

Two fair dice are rolled. Let X be the largest value obtained and Y be the sum of the two die. Find the joint probability mass function.

Solution

Each outcome (d_1, d_2) has a probability of $\frac{1}{36}$ happening where $X = \max(d_1, d_2)$ and $Y = d_1 + d_2$. For a specific (a, b) pair, we need to count how many outcomes (d_1, d_2) satisfy $\max(d_1, d_2) = a$ and $d_1 + d_2 = b$. If the largest die is a, then both die are between 1 and a and one is a. Thus, b must satisfy

$$a + 1 \leq b \leq 2a \tag{6.25}$$

For $\max(d_1, d_2) = a$ and $d_1 + d_2 = b$. If $b = 2a$, only (a, a) works so there is 1 outcome. If $a + 1 \leq b \leq 2a$, the other die must be $b - a$ which must satisfy $1 \leq b - a < a$. This gives two outcomes: $(a, b - a)$ and $(b - a, a)$. Thus,

$$p(a, b) = \begin{cases} \frac{1}{36}, & b = a \\ \frac{2}{36}, & a + 1 \leq b < 2a \end{cases} \tag{6.26}$$

for $a \in \{1, 2, 3, 4, 5, 6\}$

Example 40

The joint probability density of X and Y is given by

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y}, & 0 < x, y < \infty \\ 0, & \text{otherwise} \end{cases} \tag{6.27}$$

Find $\mathbb{P}(X < Y)$

Solution

$X < Y$ is everything above the line $y = x$. Thus,

$$\mathbb{P}(X < Y) = \int_0^\infty \int_x^\infty 2e^{-x}e^{-2y}dydx \tag{6.28}$$
$$= \int_0^\infty -e^{-3x}dx \tag{6.29}$$
$$= \frac{1}{3} \tag{6.30}$$

This makes sense since $2e^{-x}e^{-2y} = e^{-x} \cdot 2e^{-2y}$. Thus, $f(x, y)$ factors into $X \sim \text{Exp}(1)$ and $Y \sim \text{Exp}(2)$, Thus, Y decays faster so we would expect $\mathbb{P}(X < Y)$ to be smaller than $\frac{1}{2}$

Example 41

The joint density function of X and Y is given by

$$f(x, y) = \begin{cases} c(y^2 - x^2)e^{-y}, & -y \leq x \leq y, 0 < y < \infty \\ 0, & \text{otherwise} \end{cases} \tag{6.31}$$

Find c, the marginal densities of X and Y, $\mathbb{E}(X)$ and $\mathbb{P}(0 < X < 1, Y < 1)$.

Solution

The graph is given below. Now,

$$\int_0^\infty \int_{-y}^{y} c(y^2 - x^2)e^{-y} dx dy = 1 \tag{6.32}$$

$$\implies c\int_0^\infty \int_{-y}^{y} y^2e^{-y} - x^2e^{-y} dx dx = 1 \tag{6.33}$$

$$\implies c\int_0^\infty 2y^3e^{-y} - \frac{2y^3}{3}e^{-y} dy = 1 \tag{6.34}$$

$$\implies c\int_0^\infty \frac{4y^3}{3}e^{-y} dy = 1 \tag{6.35}$$

$$\implies 8c = 1 \tag{6.36}$$

$$\implies c = \frac{1}{8} \tag{6.37}$$

Solution

$$f(x,y) = \begin{cases} \frac{1}{8}(y^2 - x^2)e^{-y}, & -y \leq x \leq y, 0 < y < \infty \\ 0, & \text{otherwise} \end{cases} \tag{6.38}$$

$$f_X(x) = \int_{-\infty}^{\infty} \frac{1}{8}(y^2 - x^2)e^{-y} dy \tag{6.39}$$

$$= \int_{|x|}^{\infty} \frac{1}{8}(y^2 - x^2)e^{-y} dy \tag{6.40}$$

$$= \frac{1+|x|}{4} e^{-|x|} \tag{6.41}$$

$$f_Y(y) = \int_{-\infty}^{\infty} \frac{1}{8}(y^2 - x^2)e^{-y} dy \tag{6.42}$$

$$\tag{6.43}$$

$$= \int_{-y}^{y} \frac{1}{8}(y^2 - x^2)e^{-y} dy \tag{6.44}$$

$$= \frac{y^3}{6} e^{-y}, y > 0 \tag{6.45}$$

This is the Gamma density. Namely, $f_Y(y) = \frac{y^{4-1}e^{-y}}{\Gamma(4)} \implies Y \sim \Gamma(4,1)$.

Solution

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx \tag{6.46}$$

$$= \int_{-\infty}^{\infty} x \frac{1+|x|}{4} e^{-x} dx \tag{6.47}$$

This is an odd function so $\mathbb{E}(X) = 0$.

Solution

We integrate the joint density over the region $\{0 < x < 1,\ 0 < y < 1\}$, but we must also respect the support constraint $x \leq y$. Since $x > 0$, the constraint $-y \leq x$ is automatic. The binding constraint is $x \leq y$, so for a given $y \in (0, 1)$, x ranges from 0 to $\min(y, 1) = y$ (since $y < 1$) (draw it out to see):

$$\mathbb{P}(0 < X < 1, Y < 1) = \int_0^1 \int_0^y \frac{1}{8}(y^2 - x^2)e^{-y} dxdy \tag{6.48}$$

$$= \frac{16}{e} \tag{6.49}$$

The ideas above can be extended to more than two random variables. If we have n random variables $X_1, X_2, \ldots, X_n$, their joint cumulative distribution function is defined as

$$F(a_1, a_2, \ldots, a_n) = \mathbb{P}(X_1 \leq a_1, X_2 \leq a_2, \ldots X_n < a_n) \tag{6.50}$$

Further, $X_1, X_2, \ldots, X_n$ are jointly continuous if there is a non-negative function $f(x_1, x_2, \ldots, x_n)$ called the joint probability density function such that for any reasonable set C in n-dimensional space,

$$\mathbb{P}((X_1, X_2, \ldots, X_n) \in C) = \int \int \cdots \int_{(x_1, x_2, \ldots, x_n) \in C} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n \tag{6.51}$$

6.2 Independence of Joint Distributions

Definition 6.1 (Independence of Random Variables)**.** Two random variables X and Y are independent if, for any sets A and B,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \tag{6.52}$$

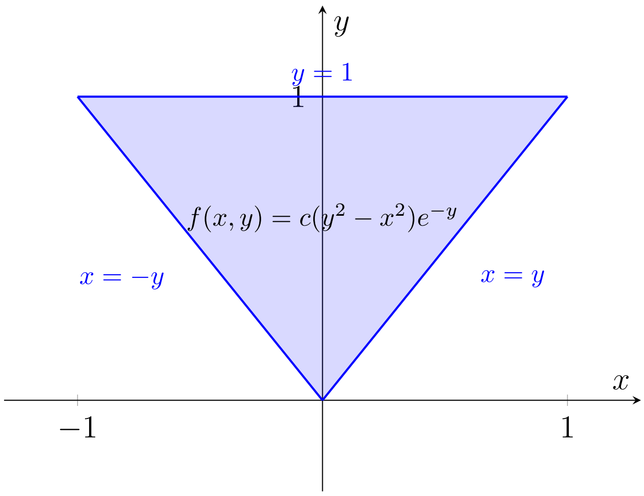


Figure 21: Support of $f(x, y) = c(y^2 - x^2)e^{-y}$ over the region $-y \leqslant x \leqslant y$, $0 < y < 1$.

In other words, the events $\{X \in A\}$ and $\{Y \in B\}$ must be independent events.

If two random variables are not independent, they are dependent. It can be shown that independence is equivalent to

Lemma 6.1. If two random variables X and Y are independent, then

$$F(a, b) = F_X(a)F_Y(b)\ \forall\ a, b \tag{6.53}$$

where F is the joint CDF of X and Y, F_X is the marginal CDF of X and F_Y is the marginal CDF of Y

Proof.

$$F(a, b) = \mathbb{P}(X \le a, Y \le b) \tag{6.54}$$

$$= \mathbb{P}(X \le a)\mathbb{P}(Y \le b) \tag{6.55}$$

$$= F_X(a)F_Y(b) \tag{6.56}$$

■

Lemma 6.2. If two discrete random variables X and Y are independent, then

$$p(x, y) = p_X(x)p_Y(y)\ \forall\ x, y \tag{6.57}$$

where $p(x, y)$ is the joint PMF of X and Y, p_X is the marginal PMF of X and p_Y is the marginal PMF of Y

Proof.

$$p(x, y) = \mathbb{P}(X = x, Y = y) \tag{6.58}$$

$$= \mathbb{P}(X = x)\mathbb{P}(Y = y) \tag{6.59}$$

$$= p_X(x)p_Y(y) \tag{6.60}$$

■

Lemma 6.3. If two jointly continuous random variables X and Y are independent, then

$$f(x, y) = f_X(x)f_Y(y)\,\forall\, x, y \tag{6.61}$$

where $f(x, y)$ is the joint PDF of X and Y, f_X is the marginal PDF of X and f_Y is the marginal PDF of Y.

Proof.

$$\mathbb{P}((X, Y) \in C) = \int dx \int_C dy f(x, y) \tag{6.62}$$

$$= \mathbb{P}(X \in C)\mathbb{P}(Y \in C) \tag{6.63}$$

$$= \int_C f_X(x, y)dx \int_C f_Y(x, y)dy \tag{6.64}$$

■

However, we don't have to compute the marginal PDF or PMF to check independence. It is enough to check the joint PMF or PDF factors into a function of x times a function of y alone. We can see this for the discrete case. Let

$$p(x, y) = h(x)g(y) \tag{6.65}$$

Then, the marginal PMF of X is

$$p_X(x) = \sum_y h(x)g(y) \tag{6.66}$$

$$= h(x)\sum_y g(y) \tag{6.67}$$

$$= c \cdot h(x) \tag{6.68}$$

for some constant c. Similarly, $p_Y(y) = d \cdot g(y)$ for some constant d. Now,

$$\sum_x \sum_y p(x, y) = 1 \tag{6.69}$$

Then,

$$\sum_x \sum_y p(x, y) = \sum_x \sum_y h(x)g(y) \tag{6.70}$$

$$= \sum_x h(x) \sum_y g(y) \tag{6.71}$$

$$= cd \tag{6.72}$$

$$= 1 \tag{6.73}$$

Thus,

$$p(x, y) = h(x)g(y) \tag{6.74}$$

$$= cdh(x)g(y) \tag{6.75}$$

$$= p_X(x)p_Y(y) \tag{6.76}$$

Similarly for the continuous case. Let

$$f(x, y) = h(x)g(y) \tag{6.77}$$

Then, the marginal PDF of X is

$$f_X(x) = \int_{-\infty}^{\infty} h(x)g(y)\,dy \tag{6.78}$$
$$= h(x)\int_{-\infty}^{\infty} g(y)\,dy \tag{6.79}$$
$$= c \cdot h(x) \tag{6.80}$$

for some constant c. Similarly, $f_Y(y) = d \cdot g(y)$ for some constant d. Now,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1 \tag{6.81}$$

Then,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x)g(y)\,dx\,dy \tag{6.82}$$
$$= \int_{-\infty}^{\infty} h(x)\,dx \int_{-\infty}^{\infty} g(y)\,dy \tag{6.83}$$
$$= cd \tag{6.84}$$
$$= 1 \tag{6.85}$$

Thus,

$$f(x,y) = h(x)g(y) \tag{6.86}$$
$$= cd\,h(x)g(y) \tag{6.87}$$
$$= f_X(x)f_Y(y) \tag{6.88}$$

However, we can only do this when the factoring works over the entire support. The region where $f(x,y) > 0$ must also be a "rectangle" (a product of intervals), not some oddly shaped region that couples x and y together.

Example 42

Let $f(x,y) = 6e^{-2x}e^{-3y}, 0 < x, y < \infty$. This factors into $h(x) = 6e^{-2x}$ and $g(y) = e^{-3y}$. The support is also the interval $(0,\infty) \times (0,\infty)$ which is a product of two intervals. Thus, X and Y are independent. In fact, $X \sim \text{Exp}(2)$ and $Y \sim \text{Exp}(3)$

Example 43

Let $f(x,y) = 24xy, 0 < x < 1, 0 < y < 1, x + y < 1$. Here, we are bounded by a triangle, which cannot be cleanly expressed as the product of two intervals. Thus, X and Y are dependent. Also, note that $f\left(\frac{1}{2}, \frac{3}{4}\right) = 0$ since $\frac{1}{2} + \frac{3}{4} > 1$. However, $f_X\left(\frac{1}{2}\right) > 0$ and $f_Y\left(\frac{3}{4}\right) > 0$ since they are individually valid values. Thus, $f(x,y) \neq f_X(x)f_Y(y)$.

We can demonstrate Poisson splitting using independence. Suppose the number of people who enter a post-office on a given day follows a Poisson distribution with parameter λ. Each person who enters the post office is male with probability p and female with probability $1-p$. We need to show the number of males and females entering the post office are independent Poisson random variables with parameters λp and $\lambda(1-p)$. Let X be the number of males and Y be the number of females. Thus, $X + Y \sim \text{Poi}(\lambda)$. Also, $\mathbb{P}(X + Y = n) = e^{-\lambda}\frac{\lambda^n}{n!}$. Now,

$$\begin{aligned}
\mathbb{P}(X = a, Y = b) &= \mathbb{P}(X = a | X + Y = a + b)\mathbb{P}(X + Y = a + b) && (6.89)\\
&= \binom{a+b}{a} p^a(1-p)^b \cdot e^{-(\lambda)}\frac{\lambda^{a+b}}{(a+b)!} && (6.90)\\
&= \frac{(a+b)!}{a!b!} p^a(1-p)^b \cdot e^{-\lambda}\frac{\lambda^a\lambda^b}{(a+b)!} && (6.91)\\
&= \frac{e^{-\lambda p}p^a}{a!} \cdot \frac{e^{-\lambda(1-p)}((1-p)\lambda)^b}{b!} && (6.92)\\
&= \mathbb{P}\left(\text{Poi}(\lambda p) = a\right)\mathbb{P}\left(\text{Poi}(\lambda(1-p)) = b\right) && (6.93)
\end{aligned}$$

Example 44

Two points are randomly selected independently on a line of length L to be on opposite sides of the midpoint of the line. The two points are independent random variables X and Y where $X \sim \text{Uniform}\left(0, \frac{L}{2}\right)$ and $Y \sim \text{Uniform}\left(\frac{L}{2}, L\right)$. Find the probability the distance between the two points is greater than $\frac{L}{3}$.

Solution

We want $\mathbb{P}\left(Y - X > \frac{L}{3}\right)$. The density functions of X and Y are

$$f_X(x) = \begin{cases} \frac{2}{L}, & 0 < x < \frac{L}{2} \\ 0, & \text{otherwise} \end{cases} \tag{6.94}$$

$$f_Y(y) = \begin{cases} \frac{2}{L}, & \frac{L}{2} < x < L \\ 0, & \text{otherwise} \end{cases} \tag{6.95}$$

Then,

$$f(x,y) = \begin{cases} \frac{4}{L^2}, & 0 < x < \frac{L}{2}, \frac{L}{2} < y < L \\ 0, & \text{otherwise} \end{cases} \tag{6.96}$$

$$\mathbb{P}\left(Y - X > \frac{L}{3}\right) = 1 - \mathbb{P}\left(0 < Y - X < \frac{L}{3}\right) \tag{6.97}$$

$$= 1 - \int_{\frac{L}{6}}^{\frac{L}{2}} \int_{\frac{L}{2}}^{x+\frac{L}{3}} \frac{4}{L^2} dy dx \qquad = 1 - \frac{2}{9} = \frac{7}{9} \tag{6.98}$$

Example 45

Buffon's Needle Problem. A table has parallel lines ruled D apart. A needle of length $L < D$ is thrown randomly onto the table. What is the probability it crosses a line?

Solution

The needle's position is described by two quantities: X = the distance from the needle's midpoint to the nearest lines. X ranges from 0 to $\frac{D}{2}$ and Θ = the acute angle between the needle and the line. Both are uniformly distributed and independent with $X \sim \text{Uniform}\left(0, \frac{D}{2}\right)$ and $\Theta \sim \text{Uniform}\left(0, \frac{\pi}{2}\right)$. Thus,

$$f_X(x) = \frac{2}{D}, 0 < x < \frac{D}{2} \tag{6.99}$$

$$f_\Theta(\theta) = \frac{2}{\theta}, 0 < \theta < \frac{\pi}{2} \tag{6.100}$$

The needle extends a distance of $\frac{L}{2}\cos\theta$ on each side. Thus, the needle crosses a line when the perpendicular projection exceeds X

$$\mathbb{P}\left(X \leq \frac{L}{2}\cos\theta\right) = \int_0^{\frac{\pi}{2}} \int_0^{\frac{L}{2}\cos\theta} \frac{4}{\pi D} dx d\theta \tag{6.101}$$

$$= \frac{2L}{\pi D} \tag{6.102}$$

We can use this result to approximate π!

We can define independence for n random variables as well.

Definition 6.2 (Independence of n random variables)**.** Random variables $X_1, X_2, \cdots, X_n$ are independent if for any subset $I \subseteq \{1, 2, \ldots, n\}$

$$\mathbb{P}(X_i \in A_i \,\forall\, i \in I) = \prod_{i \in I} \mathbb{P}(A_i) \tag{6.103}$$

That is, the events $\{X_1 \in A_1\}, \{X_2 \in A_2\}, \cdots, \{X_n \in A_n\}$ are independent.

In the jointly continuous case, this is equivalent to

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n) \,\forall\, x_1, x_2, \ldots, x_n \tag{6.104}$$

where $f(x_1, x_2, \ldots, x_n)$ is the joint PDF and $f(x_i)$ is the marginal PDF of X_i. Further,

$$\mathbb{P}((X_1, X_2, \ldots, X_n) \in C) = \int\int \cdots \int_C f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n \tag{6.105}$$

Example 46

If X_1, X_2, X_3 are independent random variables that are uniformly distributed over $(0, 1)$, compute the probability that the largest of the three is greater than the sum of the other two.

Solution

We want, $\mathbb{P}\left(\max(X_1, X_2, X_3) > X_1 + X_2 + X_3 - \max(X_1, X_2, X_3)\right)$

This is given by

$$\begin{aligned}
P &= \sum_{i=1}^{3} \mathbb{P}\left(X_i > X_j + X_k | X_i = \max(X_1, X_2, X_3)\right) \mathbb{P}(X_i > X_{i+1}, X_{i+2}) && (6.106)\\
&= 3\mathbb{P}(X_1 \geq X_2, X_3 | X_1 \geq X_2, X_3)\mathbb{P}(X_1 \geq X_2, X_3) && (6.107)\\
&= 3 \int\int\int_{x \geq y+z} dxdydz && (6.108)\\
&= 3 \int_{y=0}^{y=1} \int_{z=0}^{z=1} (1 - y - z) dzdy && (6.109)\\
&= \frac{1}{2} && (6.110)
\end{aligned}$$

6.3 Sums of Independent Random Variables

We now want to know the density of $X+Y$ if X and Y are independent jointly-continuous random variables. Let $G(a)$ be the CDF of $X + Y$. Thus,

$$\begin{aligned}
G(a) &= \mathbb{P}(X + Y \leq a) && (6.111)\\
&= \int\int_{x+y \leq a} f_{X,Y}(x, y) dxdy && (6.112)\\
& && (6.113)
\end{aligned}$$

For each fixed x, $x + y \leq a \implies y \leq a - x$. Thus,

$$G(a) = \int_{-\infty}^{\infty} f_X(x) \left(\int_{-\infty}^{a-x} f_Y(y) dy \right) dx \tag{6.114}$$
$$= \int_{-\infty}^{\infty} f_X(x) F_Y(a - x) \tag{6.115}$$

where F_Y is the CDF of Y. We then use Leibniz's rule to differentiate under the integral sign to get

$$f_{X+Y}(a) = G'(a) \tag{6.116}$$
$$= \frac{d}{da} \int_{-\infty}^{\infty} f_X(x) F_Y(a - x) dx \tag{6.117}$$
$$= \int_{-\infty}^{\infty} f_X(x) \frac{d}{da} \left(F_Y(a - x) \right) dx \tag{6.118}$$
$$= \int_{-\infty}^{\infty} f_X(x) f_Y(a - x) dx \tag{6.119}$$

This is known as the convolution of f_X and f_Y denoted $f_X(x) * f_Y(y)$. Thus,

$$f_{X+Y} = f_X(x) * f_Y(y) = \int_{-\infty}^{\infty} f_X(x) f_Y(a - x) dx \tag{6.120}$$

There are important properties about convolution

- Commutativity: $f_X * f_Y = f_Y * f_X$. Thus $f_{X+Y} = f_{Y+X}$
- Associativity: $(f_X * f_Y) * f_Z = f_X * (f_Y * f_Z)$. Thus, $f_{(X+Y)+Z} = f_{X+(Y+Z)}$

Example 47

If $X, Y \sim \text{Uniform}(0, 1)$. Find the density of $X + Y$

Solution

$$f_X(x) = 1, 0 < x < 1 \tag{6.121}$$

$$f_Y(a-x) = 1, 0 < a - x < 1 \implies a - 1 < x < a \tag{6.122}$$

Thus,

6.3.1 Closure of Sums of Independent Variables

If you add independent random variables from the same family, the sum stays in the same distributional family

Definition 6.3 (Sum of Gamma Variables). $X \sim \Gamma(s, \lambda), Y \sim \Gamma(t, \lambda)$ are independent,

$$X + Y \sim \Gamma(s + t, \lambda) \tag{6.123}$$

If X is the waiting time for s events at rate λ and Y is the additional waiting time for t more events at the same rate, then $X + Y$ is the total wait for $s + t$ events which is $\Gamma(s + t, \lambda)$. The rate must be the same for both, otherwise the sum is not gamma

Definition 6.4 (Sum of Normal Variables). If $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ are independent,

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \tag{6.124}$$

This generalises to n random variables too, you just add their means and variances.

Definition 6.5 (Sum of Poisson Random Variables). If $X \sim \text{Poi}(\lambda_1), Y \sim \text{Poi}(\lambda_2)$ are independent,

$$X + Y \sim \text{Poi}(\lambda_1 + \lambda_2) \tag{6.125}$$

Definition 6.6 (Sum of Binomial Random Variables)**.** If $X \sim \text{Bin}(n, p), Y \sim \text{Bin}(m, p)$ are independent,

$$X + Y \sim \text{Bin}(n + m, p) \tag{6.126}$$

Definition 6.7 (Sum of Negative Binomial Random Variables)**.** If $X \sim \text{NegBin}(r, p), Y \sim \text{NegBin}(s, p)$ are independent,

$$X + Y \sim (r + s, p) \tag{6.127}$$

6.4 Conditional Distributions

We now discuss conditional joint distributions

Discrete Case

Definition 6.8 (Conditional PMF)**.** If X and Y are discrete random variables with joint PMF $p(x, y)$, the conditional PMF of X given $Y = y$ is

$$p_{X|Y}(x|y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p(x, y)}{p_Y(y)} \; \forall \, y \tag{6.128}$$

If X and Y are independent, $p_{X|Y}(x|y) = p_X(x)$. The conditional distribution equals the unconditional distribution i.e. learning the value of Y tells you nothing about X. This is exactly the intuitive meaning of independence, now expressed at the level of distributions.

The conditional CDF is defined by summing the conditional PMF

$$F_{X|Y}(a, y) = \mathbb{P}(X \leq a, Y = y) = \sum_{x \leq a} p_{X|Y}(x, y) \tag{6.129}$$

Example 48

If $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$, what is the conditional distribution of $X+Y$ given $X+Y=n$. For $0 \le k \le n$,

$$p_{X|X+Y}(k|n) = \mathbb{P}(X=k|X+Y=n) \tag{6.130}$$

$$= \frac{\mathbb{P}(X=k, Y=n-k)}{\mathbb{P}(X+Y=n)} \tag{6.131}$$

$$= \frac{\frac{e^{-\lambda_1}\lambda_1^k}{k!}\frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!}}{e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!}} \tag{6.132}$$

$$= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^n\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{n-k} \tag{6.133}$$

This is a deeply satisfying result. It says: given that the total count is n, each of those n events is independently assidned to X with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$ (proportional to its rate) or to Y $\frac{\lambda_2}{\lambda_1+\lambda_2}$. This is the "reverse direction" of Poisson splitting

6.4.1 Continuous Case

Definition 6.9 (Conditional PDF)**.** If X and Y are jointly continuous random variables with joint PDF $f(x,y)$, the conditional density of X given $Y=y$ is

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \;\forall\, y \tag{6.134}$$

From this, conditional probabilities follow

$$\mathbb{P}(X \in A|Y=y) = \int_A f_{X|Y}(x|y)dx \tag{6.135}$$

The conditional CDF is

$$F_{X|Y}(a|y) = \mathbb{P}(X \le a|Y=y) = \int_{-\infty}^{a} f_{X|Y}(x,y)dx \tag{6.136}$$

We also have a continuous analogue of $\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B)$

$$f(x,y) = f_Y(y) f_{X|Y}(x|y) \tag{6.137}$$

Example 49

The joint density of X and Y is

$$f(x,y) = \frac{e^{-\frac{x}{y}} e^{-y}}{y}, 0 < x, y < \infty \tag{6.138}$$

Find $f_{X|Y}(x|y)$ and $\mathbb{P}(X > 1|Y = y)$.

Solution

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x,y)}{f_Y(y)} && (6.139) \\ &= \frac{\frac{e^{-\frac{x}{y}} e^{-y}}{y}}{\int_0^\infty \frac{e^{-\frac{x}{y}} e^{-y}}{y} dx} && (6.140) \\ &= \frac{\frac{e^{-\frac{x}{y}} e^{-y}}{y}}{e^{-y}} && (6.141) \\ &= \frac{e^{-\frac{x}{y}}}{y} && (6.142) \end{aligned}$$

Now

$$\begin{aligned} \mathbb{P}(X > 1|Y = y) &= \int_1^\infty \frac{1}{y} e^{-\frac{x}{y}} dx && (6.143) \\ &= e^{-\frac{1}{y}} && (6.144) \end{aligned}$$

Note that

$$f_{X|Y}(x|y) f_Y(y) = \frac{e^{-\frac{x}{y}} e^{-y}}{y} = f(x,y) \tag{6.145}$$

6.5 Bivariate Normal Distribution

Definition 6.10. Two jointly continuous random variables X and Y are said to be bivariate normal if their joint density is given by

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right) \tag{6.146}$$

for $x, y \in \mathbb{R}$, where $\sigma_X, \sigma_Y > 0$, $\rho \in (-1,1)$ and $\mu_X, \mu_Y \in \mathbb{R}$. This is compactly written as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\,\sigma_X\sigma_Y \\ \rho\,\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right) \tag{6.147}$$

The parameters are

$$\mu_X = \mathbb{E}(X) \tag{6.148}$$

$$\mu_Y = \mathbb{E}(Y) \tag{6.149}$$

$$\sigma_X = \mathrm{Var}(X) \tag{6.150}$$

$$\sigma_Y = \mathrm{Var}(Y) \tag{6.151}$$

ρ captures the dependence of x and y and is defined as the correlation coefficient in the next chapter.think of it as controlling how "tilted" the elliptical contours of the density are. When $\rho = 0$, the contours are axis-aligned and the variables are independent; when $|\rho|$ is close to 1, the contours are highly elongated along a diagonal, indicating strong linear dependence.

When $\rho = 0$, we can factor it into two marginal densities and get X and Y as independent normal distributions. When $\rho \neq 0$, the marginal densities can be obtained by integrating out the other variable. The result is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \tag{6.152}$$

Thus, $X \sim N(\mu_X, \sigma_X)$ regardless of the value of ρ. Similarly, $Y \sim N(\mu_Y, \sigma_Y)$. The parameter ρ affects the joint behaviour but not the marginals — both marginals are always normal with the specified means and variances.

Computing $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$ and simplifying yields

$$X|Y = y \sim N\left(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y), \sigma_X^2(1 - \rho^2)\right) \tag{6.153}$$

This says several remarkable things:

- The conditional distribution of X given $Y = y$ is still **normal** — conditioning a bivariate normal on one variable gives a normal distribution for the other. This is not true for general joint distributions.
- The conditional **mean** is $\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y)$. This is a **linear function** of y. When Y is above its mean ($y > \mu_Y$) and $\rho > 0$, the conditional mean of X shifts upward from μ_X, and vice versa. The strength of this shift is controlled by ρ.
- The conditional **variance** is $\sigma_X^2(1 - \rho^2)$, which is **smaller** than the unconditional variance σ_X^2 (since $|\rho| < 1$). Knowing Y always reduces your uncertainty about X, unless $\rho = 0$ in which case the conditional variance equals the unconditional variance — again reflecting independence.

6.6 Distribution of a Function of a Random Variable

Suppose X is continuous with density $f(x)$ and $Y = g(X)$ for some invertible function g. We want to know the density of Y. If g is strictly increasing then the CDF is given by

$$\begin{aligned} F(y) &= \mathbb{P}(Y \leq y) && (6.154)\\ &= \mathbb{P}(g(X) \leq y) && (6.155)\\ &= \mathbb{P}(X \leq g^{-1}(y)) && (6.156)\\ &= F_X(g^{-1}(y)) && (6.157) \end{aligned}$$

We can then differentiate using the chain rule to get

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y) \tag{6.158}$$

If g is strictly decreasing, then

$$\begin{aligned}
\mathbb{P}(g(X) \leq Y) &= \mathbb{P}(X \geq g^{-1}(y)) &(6.159)\\
&= 1 - F_X(g^{-1}(y)) &(6.160)\\
\implies f_Y(y) &= -f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y) &(6.161)
\end{aligned}$$

To handle both cases uniformly, we take the absolute value

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| \tag{6.162}$$

The factor in the absolute value sign is called the Jacobian in one dimension. It accounts for how the transformation g stretches or compresses the probability density. Where g stretches intervals (large derivative), the density gets thinner; where g compresses (small derivative), the density gets thicker. The absolute value ensures the density is non-negative regardless of whether g is increasing or decreasing.

Now if we have X_1 and X_2 to be jointly continuous with joint PDF $f(x_1, x_2)$. Consider the transformation $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$. We want to find the joint density of (Y_1, Y_2). The key tool is the multivariate change of variables formula. Define the transformation $g : \mathbb{R}^2 \to \mathbb{R}^2$ by $(x_1, x_2) = (g_1(x_1, x_2),\ g_2(x_1, x_2))$, and let g^{-1} denote its inverse (assuming g is one-to-one). The **Jacobian** of the inverse transformation is:

$$J = \det\begin{pmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} \end{pmatrix} \tag{6.163}$$

Then the joint density of (Y_1, Y_2) is:

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2) \cdot |J| \tag{6.164}$$

where x_1, x_2 are expressed in terms of y_1, y_2 through the inverse transformation. This is the direct generalization of the one-variable formula. In one dimension, the Jacobian was just $|dg^{-1}/dy|$; in two dimensions, it's the absolute value of the determinant of the matrix of partial derivatives. The determinant captures how the transformation stretches or compresses **area** elements in the plane — just as the one-dimensional derivative captured stretching or compressing of length elements.

Example 50

Let X_1, X_2 be jointly continuous random variables with density f_{X_1,X_2}. Define $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. We want the joint density of (Y_1, Y_2)

Solution

We invert and solve for X_1 and X_2 to give

$$X_1 = \frac{Y_1 + Y_2}{2} \tag{6.165}$$

$$X_2 = \frac{Y_1 - Y_2}{2} \tag{6.166}$$

Then the jacobian is

$$|J| = \left| \det \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix} \right| = \frac{1}{2} \tag{6.167}$$

Thus,

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2} f_{X_1,X_2}\left(\frac{y_1 + y_2}{2}, \frac{y_1 - y_2}{2}\right) \tag{6.168}$$

6.7 Joint Distribution of Maximum and Minimum

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with CDF $F(\cdot)$ and density $f(\cdot)$. Define

$$U = \max(X_1, \ldots, X_n) \tag{6.169}$$

$$V = \min(X_1, \ldots, X_n) \tag{6.170}$$

We want the joint density of (V, U). For $v < u$, the joint CDF is

$$\mathbb{P}(V \leq v, U \leq u) = \mathbb{P}(U \leq u) - \mathbb{P}(V > v, U \leq u) \tag{6.171}$$

$$= [F(u)]^n - [F(u) - F(v)]^n \tag{6.172}$$

since $\mathbb{P}(U \leq u) = \mathbb{P}(\text{all } X_i \leq u) = [F(u)]^n$ and $\mathbb{P}(V > v, U \leq u) = \mathbb{P}(\text{all } X_i \in (v, u]) = [F(u) - F(v)]^n$. Differentiating $\frac{\partial^2}{\partial v\, \partial u}$ gives the joint density

$$f_{V,U}(v, u) = n(n-1)[F(u) - F(v)]^{n-2} f(v) f(u), \quad v < u \tag{6.173}$$

This has a nice combinatorial interpretation. The factor $f(v)f(u)$ accounts for one observation landing at v and one at u. The factor $[F(u) - F(v)]^{n-2}$ is the probability the remaining $n-2$ observations fall in (v, u). The factor $n(n-1)$ counts the ways to choose which observation is the min and which is the max. Understanding the $n \to \infty$ behaviour of such random variables is the topic of **Extreme Value Theory** (EVT).

6.8 Change of Variables Formula: General Case

The change of variables formula generalises to k dimensions. Suppose $X_1, \ldots, X_k$ have joint density $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)$ and define a transformation $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^k$ by

$$Y_i = g_i(X_1, \ldots, X_k), \quad i = 1, \ldots, k \tag{6.174}$$

Assuming $\mathbf{g}$ is invertible with inverse $\mathbf{g}^{-1}$, the joint density of $(Y_1, \ldots, Y_k)$ is

$$f_{Y_1,\ldots,Y_k}(y_1, \ldots, y_k) = \left.\frac{f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)}{|J(x_1, \ldots, x_k)|}\right|_{(x_1,\ldots,x_k)=\mathbf{g}^{-1}(y_1,\ldots,y_k)} \tag{6.175}$$

where J is the Jacobian determinant

$$J(x_1, \ldots, x_k) = \det \begin{pmatrix} \dfrac{\partial g_1}{\partial x_1} & \cdots & \dfrac{\partial g_1}{\partial x_k} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial g_k}{\partial x_1} & \cdots & \dfrac{\partial g_k}{\partial x_k} \end{pmatrix} \tag{6.176}$$

This is the direct generalisation of the one-variable formula. In one dimension, the Jacobian was just $|dg^{-1}/dy|$; in k dimensions, it is the absolute value of the determinant of the matrix of all partial derivatives. The determinant captures how the transformation stretches or compresses volume elements in $\mathbb{R}^k$.

When $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^q$ with $q \neq k$, the Jacobian matrix is not square so we cannot take its determinant directly. In this case, $|J|$ is replaced by $\sqrt{\det(W^T W)}$ where $W = \left(\frac{\partial g_i}{\partial x_j}\right)$ is the $q \times k$ matrix of partial derivatives.

Example 51

Let (X, Y) denote a random point in the plane where $X, Y \sim N(0, 1)$ are independent. Find the joint distribution of R, Θ, the polar coordinate representation of (X, Y).

Solution

The joint density of (X, Y) is

$$f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2} \tag{6.177}$$

Define $R = \sqrt{X^2 + Y^2}$ and $\Theta = \arctan(Y/X)$. The inverse transformation is $X = R\cos\Theta$, $Y = R\sin\Theta$. The Jacobian is

$$J = \det \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} = r\cos^2\theta + r\sin^2\theta = r \tag{6.178}$$

Applying the formula

$$f_{R,\Theta}(r, \theta) = f_{X,Y}(r\cos\theta, r\sin\theta) \cdot r \tag{6.179}$$

$$= \frac{1}{2\pi} e^{-(r^2\cos^2\theta + r^2\sin^2\theta)/2} \cdot r \tag{6.180}$$

$$= \frac{r}{2\pi} e^{-r^2/2}, \quad r > 0,\ 0 \le \theta < 2\pi \tag{6.181}$$

This factors as $h(r) \cdot g(\theta) = (re^{-r^2/2}) \cdot \frac{1}{2\pi}$ over a rectangular support, so R and Θ are independent with

- R has density $re^{-r^2/2}$ for $r > 0$ — the **standard Rayleigh distribution**
- $\Theta \sim \text{Uniform}[0, 2\pi]$

The radial symmetry of the joint normal density makes the angle completely uniform, and the distance from the origin follows the Rayleigh distribution. This also provides an elegant method for generating normal random variables on a computer known as the **Box-Muller transform**.

7 Properties of Expectation

7.1 Expectation of Functions of Random Variables

Let $g(x, y)$ be a function and assume we have two jointly distributed random variables X and Y. We are often interested in computing $\mathbb{E}(g(X, Y))$.

Definition 7.1 (Expectation of a function of random variables)**.** Let $g(x, y)$ be a function. If X and Y are jointly discrete random variables with joint pmf $p(x, y)$, then

$$\mathbb{E}(g(X, Y)) = \sum_x \sum_y g(x, y) p(x, y) \tag{7.1}$$

If X and Y are jointly continuous with joint PDF $f(x, y)$ then

$$\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy \tag{7.2}$$

More generally, for $\mathbf{X} = (X_1, \ldots, X_k)$ with joint density f, $\mathbb{E}[g(\mathbf{X})] = \int g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$. The power of this formula is that we do not need to find the distribution of $g(X, Y)$ first, we can compute the expectation directly from the joint distribution of (X, Y).

Example 52

Let X and Y be independent variables uniformly distributed on $(0, 1)$. Find $\mathbb{E}|X - Y|^\alpha$ for $\alpha > 0$.

Solution

The joint density is $f(x, y) = 1$ on $(0, 1)^2$. Splitting based on whether $x > y$ or $x < y$ and using symmetry

$$\mathbb{E}|X - Y|^\alpha = \int_0^1 \int_0^1 |x - y|^\alpha dxdy \tag{7.3}$$
$$= 2\int_0^1 \int_0^y (y - x)^\alpha dxdy \tag{7.4}$$
$$= 2\int_0^1 \frac{y^{\alpha+1}}{\alpha + 1} dy \tag{7.5}$$
$$= \frac{2}{(\alpha + 1)(\alpha + 2)} \tag{7.6}$$

As a sanity check: for $\alpha = 1$, $\mathbb{E}|X - Y| = \frac{1}{3}$ and for $\alpha = 2$, $\mathbb{E}(X - Y)^2 = \frac{1}{6}$.

A consequence of the above is that

$$\mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}(X_i) \tag{7.7}$$

Thus for two random variables X and Y,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) \tag{7.8}$$

Proof. For the continuous case, let $g(x, y) = x + y$. Then,

$$\mathbb{E}(X + Y) = \int\int (x + y)f(x, y)dxdy \tag{7.9}$$
$$= \int\int xf(x, y)dxdy + \int\int yf(x, y)dxdy \tag{7.10}$$
$$= \int x\left(\int f(x, y)dy\right)dx + \int y\left(\int f(x, y)dx\right)dy \tag{7.11}$$
$$= \int xf_X(x)dx + \int yf_Y(y)dy \tag{7.12}$$
$$= \mathbb{E}(X) + \mathbb{E}(Y) \tag{7.13}$$

■

Further, if X and Y are independent and $g(x)$ and $h(y)$ are two functions, then

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \tag{7.14}$$

Proof. Recall that if X and Y are independent then $f(x,y) = f_X(x)f_Y(y)$. Thus,

$$\begin{aligned}
\mathbb{E}(g(x)h(Y)) &= \int\int g(x)h(y)f(x,y)dxdy && (7.15)\\
&= \int\int g(x)h(y)f_X(x)f_Y(y)dxdy && (7.16)\\
&= \left(\int g(x)f_X(x)dx\right)\left(\int h(y)f_Y(y)dy\right) && (7.17)\\
&= \mathbb{E}(g(X))\mathbb{E}(h(Y)) && (7.18)
\end{aligned}$$

■

We often want to compute $\mathbb{E}(X)$ where X is the number of things that happen. For example, X might be the number of successful trials, the number of matching pairs, etc. In essence, X is a count. We can always write a count as a sum of indicator variables.

Definition 7.2. If there are events $A_1,\ A_2, \cdots, A_n$ we define the indicator random variable as

$$\mathbb{I}_{A_i} = \begin{cases} 1 & \text{if event } A_i \text{ occurs,} \\ 0 & \text{otherwise} \end{cases} \tag{7.19}$$

If X counts how many of each A_i occurs, then

$$X = \sum_{i=1}^{n} \mathbb{I}_{A_i} \tag{7.20}$$

Thus by the linearity of expectation,

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(\mathbb{I}_{A_i}) \tag{7.21}$$

Now,

$$\mathbb{E}(\mathbb{I}_{A_i}) = 1 \cdot \mathbb{P}(A_i) + 0 \cdot \mathbb{P}(A_i^c) \tag{7.22}$$

$$= \mathbb{P}(A_i) \tag{7.23}$$

Thus,

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{P}(A_i) \tag{7.24}$$

This formula is extraordinarily useful because counting random variables often have very complicated distributions. The events $A_1, \cdots, A_n$ might be highly dependent on one another, and finding the actual distribution of X might be intractable. But none of that matters for computing the expectation. All you need are the marginal probabilities $\mathbb{P}(A_i)$ for each individual event, and those are usually much easier to compute.

Example 53

Ten hunters are waiting for ducks to fly by. When a flock of ducks flies overhead, the hunters aim at the same time, but each chooses his target at random, independently of the others. If each hunter independently hits his target with probability p, compute the expected number of ducks that escape unhurt when a flock of size f flies overhead.

Solution

For each duck from $i = 1$ to $i = f$, define the indicator variable $\mathbb{I}_{A_i}$ where A_i is the event the ducks escape unhurt. Thus,

$$\mathbb{E}(X) = \sum_{i=1}^{f} \mathbb{E}(\mathbb{I}_{A_i}) \tag{7.25}$$

$$= \sum_{i=1}^{f} \mathbb{P}(A_i) \tag{7.26}$$

$$= f\mathbb{P}(A_i) \tag{7.27}$$

Now we need to determine $\mathbb{P}(A_i)$. A given hunter fails to kill duck i if he chooses a different duck altogether which has probability $\frac{f-1}{f}$. Or, he could miss which has probability $\frac{1-p}{f}$ where $\frac{1}{f}$ is the probability of choosing that duck. Thus,

$$\frac{f-1}{f} + \frac{1-p}{f} = 1 - \frac{p}{f} \tag{7.28}$$

Since all 10 hunters act independently,

$$\mathbb{P}(A_i) = \left(1 - \frac{p}{f}\right)^{10} \tag{7.29}$$

Then,

$$\mathbb{E}(X) = f\left(1 - \frac{p}{f}\right)^{10} \tag{7.30}$$

We can also compute higher moments via par indicators. We need the identity,

$$\binom{X}{2} = \frac{X(X-1)}{2} = \sum_{i_1 < i_2} \mathbb{I}_{A_{i_1}} \mathbb{I}_{A_{i_2}} \tag{7.31}$$

The left-hand side is the number of pairs of distinct events that both occur. The right-hand side counts the same thing, by summing the product $\mathbb{I}_{A_{i_1}} \mathbb{I}_{A_{i_2}}$ over all pairs (i_1, i_2)

with $i_1 < i_2$. This this product is 1 if and only if both events A_{i_1} and A_{i_2} occur. Then,

$$\mathbb{E}(X(X-1)) = 2\sum_{i_1<i_2} \mathbb{P}(A_{i_1} \cap A_{i_2}) \tag{7.32}$$

From this, we can get

$$\mathbb{E}(X^2) = \mathbb{E}(X(X-1)) + \mathbb{E}(X) \tag{7.33}$$

which we can use to compute variance.

Example 54

Let $X \sim \text{Bin}(n,p)$. We can represent X as the number of successes in n independent Bernoulli trials. Let A_i be the event that trial i is a success. Then, $X = \sum_{i=1}^n \mathbb{I}_{A_i}$ where $\mathbb{P}(A_i) = p$. Thus,

$$\mathbb{E}(X) = \sum_{i=1}^n \mathbb{P}(A_i) = np \tag{7.34}$$

Then for any two pairs $i_1 < i_2$ by independence,

$$\mathbb{P}(A_{i_1} \cap A_{i_2}) = p^2 \tag{7.35}$$

There are $\binom{n}{2}$ such pairs. Thus,

$$\mathbb{E}(X(X-1)) = 2\binom{n}{2}p^2 = n(n-1)p^2 \tag{7.36}$$

Thus,

$$\mathbb{E}(X^2) = n(n-1)p^2 + np \tag{7.37}$$

Hence,

$$\text{Var}(X) = n(n-1)p^2 + np - n^2p^2 = np(1-p) \tag{7.38}$$

7.2 Variance and Covariance

Definition 7.3 (Covariance)**.** The covariance between two random variables X and Y is

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right) \tag{7.39}$$

We can equivalently state this as

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{7.40}$$

Proof.

$$(X-\mathbb{E}(X))(Y-\mathbb{E}(Y)) = XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y) \tag{7.41}$$

Then,

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathbb{E}\left(XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)\right) && (7.42)\\
&= \mathbb{E}(XY) - \mathbb{E}(Y)\mathbb{E}(X) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) && (7.43)\\
&= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) && (7.44)
\end{aligned}$$

■

The intuition is the following. The quantity $X-\mathbb{E}(X)$ measures how much X deviates from its average value, and similarly for $Y-\mathbb{E}(Y)$. The product $(X)(Y)$ is positive when both deviations have the same sign. That is, when X and Y are both above their means or both below their means simultaneously. The product is negative when one is above and the other below. Therefore, the average of this product, which is the covariance, is positive when X and Y tend to deviate together in the same direction, and negative when they tend to deviate in opposite directions. A covariance of zero means there is no consistent linear relationship between the deviations. The magnitude of covariance depends on the units in which X and Y are measured. If you measure X in meters and rescale to centimeters, the covariance changes by a factor of 100. This is why we often prefer correlation, which is a unit-free version, as we will see in the next section.

There are some basic properties of covariance. The first is that covariance is symmetric. This follows from its definition

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X) \tag{7.45}$$

Self-covariance equals the variance.

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) \tag{7.46}$$

This follows since

$$\mathrm{Cov}(X, X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) \tag{7.47}$$

$$= \mathrm{Var}(X) \tag{7.48}$$

Covariance is linear in the first argument.

$$\mathrm{Cov}(aX + b, Y) = a\mathrm{Cov}(X, Y) \tag{7.49}$$

Notice that the constant b disappears. The intuition is that adding a constant to X shifts X but does not change how X fluctuates around its mean. Since covariance only sees fluctuations, the constant has no effect. This implies that the covariance of a random variable with a constant is 0.

$$\mathrm{Cov}(X, a) = 0 \tag{7.50}$$

Covariance is also bilinear.

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j) \tag{7.51}$$

We can apply bilinearity with $Y_j = X_j$ for the inner sum to get

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \operatorname{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \tag{7.52}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j) \tag{7.53}$$

We split this double sum into diagonal terms (where $i = j$) and off-diagonal terms (where $i \neq j$) and we use the symmetry of covariance to combine the off-diagonal terms in pairs:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j) \tag{7.54}$$

If X and Y are independent then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ so $\operatorname{Cov}(X, Y) = 0$. Thus, independence implies zero covariance. However, the reverse is not true.

Example 55

Let $X \sim (-1, 1)$ and $Y = X^2$. Then, $\mathbb{E}(X) = 0$ and $\mathbb{E}(XY) = \mathbb{E}(X^3) = \int_{-1}^{1} \frac{x^3}{2} dx = 0$ Thus, $\operatorname{Cov}(X, Y) = 0$ even though Y depends on X.

The above happens because covariance is blind to non-linear relationships.

7.2.1 The Correlation Coefficient

Definition 7.4 (Correlation Coefficient)**.** The correlation coefficient between two random variables X and Y is,

$$\rho(X, Y) = \operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} \tag{7.55}$$

There are three important properties from this:

- $-1 \leq \rho(X, Y) \leq 1$. X and Y are said to be uncorrelated if $\rho(X, Y) = 0$
- If $\rho(X, Y) = \pm 1$ then $Y = \pm aX + b$ with $a > 0$
- $\rho(aX + b, Y) = \rho(X, Y)$

$\rho(X, Y)$ measures the strength and direction of a linear relationship between X and Y (close to 1: strong linear, positive slope; close to

Example 56

Suppose we toss a fair coin 3 times. Let X be the number of heads and Y be the number of tails. Find $\text{Cov}(X, Y)$.

Solution

We have

$$X \sim \text{Bin}(3, 0.5) \tag{7.56}$$

$$Y = 3 - X \tag{7.57}$$

Thus,

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(X, 3 - X) && (7.58) \\ &= \text{Cov}(X, 3) - \text{Cov}(X, X) && (7.59) \\ &= \text{Cov}(X, 3) - \text{Var}(X) && (7.60) \\ &= 0 - \text{Var}(X) && (7.61) \\ &= -0.75 && (7.62) \end{aligned}$$

Further,

$$\rho(X, Y) = \frac{-0.75}{\sqrt{0.75^2}} = -1 \tag{7.63}$$

Note that we had

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j) \tag{7.64}$$

If the random variables X_i are pairwise independent, that is any two of them are independent of each other. Then for every pair (i, j) with $i \neq j$ we have $\text{Cov}(X_i, X_j) = 0$ so we are left with

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \tag{7.65}$$

The variance formula for sums of indicators is also as follows. Suppose $X_i = \mathbb{I}_{A_i}$. Then,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = 2\sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_i \mathbb{P}(A_i) - \left(\sum_i \mathbb{P}(A_i)\right)^2 \tag{7.66}$$

Proof. We have $\mathbb{E}(X_i^2) = \mathbb{E}(X_i) = \mathbb{P}(A_i)$. Thus,

$$\mathrm{Var}(X_i) = \mathbb{P}(A_i) - (\mathbb{P}(A_i))^2 = \mathbb{P}(A_i)\,(1 - \mathbb{P}(A_i)) \tag{7.67}$$

Also,

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) = \mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i)\mathbb{P}(A_j) \tag{7.68}$$

This is because $X_i X_j$ is an indicator itself which is 1 when A_i and A_j occur. ■

Example 57

A particle starts at the origin in the plane. At each step, it moves one unit of distance in a completely random direction. By "completely random direction" we mean that the angle of the step is uniformly distributed over $(0, 2\pi)$, independent of all previous steps. The position after n n steps is therefore a random point in the plane, and we want to compute the expected squared distance from the origin after n steps

Solution

Let $\theta_1, \cdots, \theta_n$ be the angles of the n steps where each $\theta_i \sim (0, 2\pi)$. The position after k steps is $Z_k = (X_k, Y_k)$ where $Z_0 = (0, 0)$. The horizontal and vertical component of Z_n is

$$X_n = \cos\theta_1 + \cdots \cos\theta_n \tag{7.69}$$

$$Y_n = \sin\theta_1 + \cdots + \sin\theta_n \tag{7.70}$$

We are interested in $\mathbb{E}(X_n^2 + Y_n^2)$. By linearity,

$$\mathbb{E}(X_n^2 + Y_n^2) = \mathbb{E}(X_n^2) + \mathbb{E}(Y_n^2) \tag{7.71}$$

Also,

$$\mathbb{E}(X_n) = \sum_{i=1}^{n} \mathbb{E}(\cos\theta_i) = n\mathbb{E}(\cos\theta_1) \tag{7.72}$$

Now,

$$\mathbb{E}(\cos\theta_1) = \int_0^{2\pi} \frac{\cos\theta}{2\pi} d\theta = 0 \tag{7.73}$$

Similarly, $\mathbb{E}(Y_n) = 0$. Thus,

$$\mathbb{E}(X_n^2) = \mathrm{Var}(X_n) \tag{7.74}$$

$$\mathbb{E}(Y_n^2) = \mathrm{Var}(Y_n) \tag{7.75}$$

Since each θ_i is independent,

$$\mathrm{Var}(X_n) = \sum_{i=1}^{n} \mathrm{Var}(\cos\theta_i) = n\mathrm{Var}(\cos\theta_1) \tag{7.76}$$

Now,

$$\mathrm{Var}(\cos\theta_1) = \mathbb{E}(\cos^2\theta) = \int_0^{2\pi} \frac{\cos^2\theta}{2\pi} d\theta = \frac{1}{2} \tag{7.77}$$

Thus,

$$\mathrm{Var}(X_n) = \frac{n}{2} \tag{7.78}$$

Similarly,

$$\mathbb{E}(Y_n^2) = \mathrm{Var}(Y_n) = \frac{n}{2} \tag{7.79}$$

So after n steps of a random walk in the plane, the expected squared distance from the origin is exactly n

7.3 Conditional Expectation

Recall that for two random variables X and Y, the conditional distribution of X given $Y = y$ is

$$p_{X|Y}(x|y) = \frac{p(x,y)}{p_Y(y)} \tag{7.81}$$

in the discrete case where $p(x, y)$ is the joint PMF and $p_Y(y)$ is the marginal PMF of Y. In the continuous case,

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \tag{7.82}$$

where $f(x, y)$ is the joint density and $f_Y(y)$ is the marginal density of Y. Similarly, we can define conditional expectation.

Definition 7.5 (Conditional Expectation)**.** The conditional expectation of X given $Y = y$ is the expected value of X computed against the conditional distribution. Since this depends on what value y we condition on, we have a function of y

$$g(y) = \mathbb{E}(X|Y = y) = \begin{cases} \sum_x x p_{X|Y}(x|y) & \text{discrete} \\ \int_x x f_{X|Y}(x|y) & \text{continuous} \end{cases} \tag{7.83}$$

Note that $g(y) = \mathbb{E}(X|Y = y)$ is a number. It is the expected value of X given $Y = y$. We can also consider $g(Y)$ which is a result of plugging the random variable Y into g. This is a random variable, not a number. We give this random variable a special notation:

$$g(Y) = \mathbb{E}(X|Y) \tag{7.84}$$

Once you know the value of Y, you know exactly what to expect for X on average. This is $g(y)$. But before you know Y, the conditional expectation is itself uncertain, because it depends on which value of Y comes out. So $\mathbb{E}(X|Y)$ is a random variable that captures this uncertainty.

The tower property states that

$$\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X|Y)] \tag{7.85}$$

Proof. We will consider the continuous case. The discrete cases follows similarly with sums instead of integrals. We have $g(y) = \mathbb{E}(X|Y = y) = \int x f_{X|Y}(x, y)dx$. Thus,

$$\begin{aligned}
\mathbb{E}(g(Y)) &= \int g(y) f_Y(dy) && (7.86)\\
&= \int \left(\int x f_{X|Y}(x, y)dx \right) f_Y(y)dy && (7.87)\\
&= \int \int x \frac{f(x, y)}{f_Y(y)} f_Y(y)dxdy && (7.88)\\
&= \int \int x f(x, y)dxdy && (7.89)\\
&= \int x \left(\int f(x, y)dy \right) dx && (7.90)\\
&= \int x f_X(x)dx && (7.91)\\
&= \mathbb{E}(X) && (7.92)
\end{aligned}$$

■

There is a parallel result for probabilities. For any event A and a random variable Y,

$$\mathbb{P}(A) = \begin{cases} \sum_y \mathbb{P}(A|Y = y)p_Y(y) & \text{discrete } Y \\ \int_{-\infty}^{\infty} \mathbb{P}(A|Y = y)f_Y(y)dy & \text{continuous Y} \end{cases} \tag{7.93}$$

This is the tower property applied to the indicator $X = \mathbb{I}_A$ since $\mathbb{P}(A) = \mathbb{E}(\mathbb{I}_A)$. If X and Y are independent then

$$\mathbb{E}(X|Y) = \mathbb{E}(X) \tag{7.94}$$

This is because conditioning on Y does not change the expected value of X if X does not depend on Y.

Example 58

A miner is trapped in a mine with three doors. Door 1 leads to a tunnel that brings him to safety after 3 hours. Door 2 leads to a tunnel that returns him to the mine after 5 hours. Door 3 leads to a tunnel that returns him to the mine after 7 hours. Each time the miner makes a choice, he picks one of the three doors uniformly at random, independently of his previous choices. We want to find the expected time until he reaches safety.

Solution

Let X be the total time until safety and Y be the choice of door on the first attempt. Thus $\text{supp}(Y) = \{1, 2, 3\}$ where each value has probability $\frac{1}{3}$. Let $g(y) = \mathbb{E}(X|Y = y)$. Thus,

$$g(1) = 3 \tag{7.95}$$

Now, if he picks door 2, he travels in the mine 5 hours and he again has to pick a door. We effectively have a memoryless situation. Thus,

$$g(2) = 5 + \mathbb{E}(X) \tag{7.96}$$

Similarly,

$$g(3) = 7 + \mathbb{E}(X) \tag{7.97}$$

Thus, by the tower property

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(\mathbb{E}(X|Y)) && (7.98)\\
&= \mathbb{E}(g(Y)) && (7.99)\\
&= \frac{1}{3}(3) + \frac{1}{3}(5 + \mathbb{E}(X)) + \frac{1}{3}(7 + \mathbb{E}(X)) && (7.100)
\end{aligned}$$

Thus,

$$\mathbb{E}(X) = 15 \tag{7.101}$$

Example 59

The number of customers entering a department store on a given day is a random variable N N with mean 50. The amount of money spent by each customer is a random variable, with all customers having the same mean of 8 dollars. The amounts spent are independent of each other and independent of N . We want the expected total amount of money spent in the store.

Solution

Let N be the number of customers with $\mathbb{E}(N) = 50$. Let S_i be the ammount spent by the i^{th} customer with $\mathbb{E}(S_i) = 8$. The total amount of money spent is

$$T = S_1 + \cdots + S_N \tag{7.102}$$

If we know n customers came in then

$$\mathbb{E}(T|N = n) = \mathbb{E}(S_1 + \ldots S_n|N = n) \tag{7.103}$$

Since these are independent,

$$\mathbb{E}(T|N = n) = \mathbb{E}(S_1 + \cdots + S_n) = n\mathbb{E}(S_1) = 8n \tag{7.104}$$

Now,

$$g(n) = \mathbb{E}(T|N = n) = 8n \tag{7.105}$$

Then, by the tower property

$$\mathbb{E}(T) = \mathbb{E}(g(N)) = \mathbb{E}(8N) = 8\mathbb{E}(N) = 8 \cdot 50 = 400 \tag{7.106}$$

The above technique is called Wald's identity.

Theorem 7.1 (Wald's Identity)**.** If N is a non-negative integer-valued random variable, $S_1, S_2, \ldots S_n$ are independent and identically distributed random variables with mean μ and N is independent of each S_i, then

$$\mathbb{E}\left(\sum_{i=1}^{n} S_i\right) = \mu\mathbb{E}(N) \tag{7.107}$$

7.4 Moment Generating Functions

Definition 7.6 (Moment Generating Function)**.** The moment generating function of a random variable X, denoted $M_X(t)$, is defined as the expected value of e^{tX} which is a function of the real parameter t

$$M_X(t) = \mathbb{E}\left(e^{tX}\right) \tag{7.108}$$

In the discrete case this is

$$M_X(t) = \sum_{x\in\mathrm{supp}(X)} e^{tx}p(x) \tag{7.109}$$

In the continuous case this is

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx}f(x)dx \tag{7.110}$$

By the Taylor expansion of e^x, we have

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + \cdots = \sum_{n=0}^{\infty} \frac{t^n}{n!}X^n \tag{7.111}$$

Taking the expectation of both sides,

$$\mathbb{E}\left(e^{tX}\right) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}(X^n) \tag{7.112}$$

Thus, the coefficients of the power series expansion of $M_X(t)$ are the different moments of X. We can further extract moments by differentiation,

$$M_X(0) = \mathbb{E}(1) = 1 \tag{7.113}$$

Then by differentiating each term,

$$M'_X(0) = \mathbb{E}(X) \tag{7.114}$$

$$M''_X(0) = \mathbb{E}(X^2) \tag{7.115}$$

$$M_X^(n)(0) = \mathbb{E}(X^n) \tag{7.116}$$

The moment generating function uniquely determines the distribution of X. Theoretically, we can write down the CDF of X given its MGF. Two random variables with the same MGF (in the neighbourhood of 0) must have the same distribution. Further,

Theorem 7.2. If X and Y are independent random variables then

$$M_{X+Y}(t) = M_X(t)M_Y(t) \tag{7.117}$$

Proof. We start with the forward direction,

$$M_{X+Y}(t) = \mathbb{E}\left(e^{t(X+Y)}\right) \tag{7.118}$$

$$= \mathbb{E}(e^{tX} \cdot e^{tY}) \tag{7.119}$$

$$= \mathbb{E}\left(e^{tX}\right)\mathbb{E}e^{tY} \tag{7.120}$$

$$= M_X(t)M_Y(t) \tag{7.121}$$

The reverse proof is ommitted. T ■

he moment problem showed that MGFs uniquely determine the distribution of a random variable but individual moments do not. This is because the individual moments may be the coefficient of a power series that has radius of convergence to be 0.

We now turn to compute the MGFs of some of the distributions we have seen so far.

7.4.1 MGF of Binomial Distribution

Let $X \sim \text{Bin}(n, p)$. Then,

$$
\begin{aligned}
M_X(t) &= \mathbb{E}\left(e^{tX}\right) && (7.122)\\
&= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} e^{tk} && (7.123)\\
&= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} && (7.124)\\
&= (pe^t + 1 - p)^n \text{ (Via binomial theorem)} && (7.125)
\end{aligned}
$$

Then,

$$
\begin{aligned}
M_X'(t) &= n(pe^t + 1 - p)^{n-1} pe^t && (7.126)\\
\implies M_X'(0) &= np = \mathbb{E}(X) && (7.127)
\end{aligned}
$$

Here is a way to see the binomial MGF that makes its structure transparent. A binomial random variable $X \sim Bin(n, p)$ can be written as $X = X_1 + \cdots + X_n$ where each X_i is an independent Bernoulli(p) random variable taking value 1 with probability p and value 0 with probability $1 - p$. The MGF of a single Bernoulli is

$$
M_{X_1}(t) = e^t p + e^0(1-p) = pe^t + 1 - p \qquad (7.128)
$$

Since each Bernoulli trials are independent

$$
M_X(t) = \prod_{i=1}^{N} M_{X_i}(t) = (pe^t + 1 - p)^n \qquad (7.129)
$$

7.4.2 MGF of Poisson Distribution

Let $X \sim \text{Poi}(\lambda)$. Then,

$$M_X(t) = \mathbb{E}\left(e^{tX}\right) \tag{7.130}$$
$$= \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} \tag{7.131}$$
$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} \tag{7.132}$$
$$= e^{-\lambda} e^{\lambda e^t} \tag{7.133}$$
$$= e^{\lambda(e^t - 1)} \tag{7.134}$$

We can see that

$$M'_X(t) = \lambda e^t e^{\lambda(e^t-1)} \tag{7.135}$$
$$\implies M'_X(0) = \lambda \tag{7.136}$$

Example 60

If $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$, what is the distribution of $X + Y$?

Solution

Since X and Y are independent, use

$$M_{X+Y}(t) = M_X(t) M_Y(t) \tag{7.137}$$
$$= e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} \tag{7.138}$$
$$= e^{(\lambda_1+\lambda_2)(e^t-1)} \tag{7.139}$$

Which is the MGF of $\text{Poi}(\lambda_1 + \lambda_2)$

This was a much simpler approach than computing $\mathbb{P}(X + Y = n)$ like we did before

7.4.3 MGF of Standard Normal Distribution

Let $X \sim N(0,1)$. Then,

$$M_X(t) = \mathbb{E}(e^{tX}) \tag{7.140}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tx} e^{-\frac{x^2}{2}} dx \tag{7.141}$$

Complete the square to get

$$M_X(t) = e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} \tag{7.142}$$

$$= e^{\frac{t^2}{2}} \tag{7.143}$$

We can generalise this to general normal distributions. If $Y \sim N(\mu, \sigma^2)$, we can write it as $Y = \mu + \sigma X$ where $X \sim N(0,1)$. Then,

$$M_Y(t) = \mathbb{E}\left(e^{t(\mu+\sigma X)}\right) \tag{7.144}$$

$$= e^{\mu t} M_X(\sigma t) \tag{7.145}$$

$$= e^{\mu t + \frac{\sigma^2 t^2}{2}} \tag{7.146}$$

Example 61

Let X and Y be independent standard normal distributions. Let $U = X + Y$ and $V = X - Y$. We want to know the joint distribution of U and V

Solution

$$M_{U+V}(t) = M_{2X}(t) \tag{7.147}$$
$$= M\left(e^{2Xt}\right) \tag{7.148}$$
$$= M_X(2t) \tag{7.149}$$
$$= e^{t^2} \tag{7.150}$$

Also,

$$M_{X+Y}(t) = M_X(t)M_Y(t) = e^{t^2} \tag{7.151}$$
$$M_{X-Y}(t) = M_X(t)M_{-Y}(t) = e^{t^2} \tag{7.152}$$

Thus, U and V are independent $N(0,2)$ distributions.

7.4.4 Joint Moment Generating Functions

We can generalise MGFs to multiple random variables.

Definition 7.7 (MGF of multiple random variables)**.** Given random variables $X_1, X_2, \ldots, X_n$, the joint MGF is defined as

$$M(t_1, t_2, \ldots, t_n) = \mathbb{E}\left(e^{t_1X_1+t_2X_2+\cdots+t_nX_n}\right) \tag{7.153}$$

The joint MGF carries information about all the mixed moments of the random variables. Specifically,

$$\frac{\partial^{r_1+r_2+\cdots+r_n}}{\partial t_1^{r_1}\partial t_2^{r_2}\ldots\partial t_n^{r_n}}M|_{t_1=t_2=\cdots=t_n=0} = \mathbb{E}(X_1^{r_1}X_2^{r_2}\ldots X_n^{r_n}) \tag{7.154}$$

The joint MGF, when it exists in a neighbourhood of the origin, uniquely determines the joint distribution of $(X_1, \ldots, X_n)$. Two random vectors with the same joint MGF have the same joint distribution. Further, the random variables $X_1, X_2, \ldots, X_n$ are mutually independent if and only if the joint MGF factors

$$M(t_1, t_2, \ldots, t_n) = M_{X_1}(t_1)M_{X_2}(t_2)\ldots M_{X_n}(t_n) \tag{7.155}$$

7.5 The Prediction Problem and Conditional Expectation

Suppose we have two random variables X and Y where Y is unobservable and X is observable. The two random variables are related in some way i.e. they have some joint distribution. We ask given that we can observe X but not Y, what is our best guess for Y based on observing X. We are trying to find a function $g(X)$ that is a function of the observable X that is as close to the unobservable Y. The way we measure closeness is using the mean squared error

$$\text{MSE} = \mathbb{E}[(Y - g(X))^2] \tag{7.156}$$

We want to find g that minimises the MSE. This minimiser is called the best predictor of Y given X. As it turns out, this is exactly

$$g(X) = \mathbb{E}(Y|X) \tag{7.157}$$

The minimum MSE is

$$\text{MSE} = \text{Var}(Y|X = x) \tag{7.158}$$

8 Concentration Inequalities

The expected value of a random variable tells us where the random variable X is centered on average. However, this tells us nothing about how spread out X is, or about the chances that X will be far away from its expected value. For example, we may want to control quantities like

$$\mathbb{P}(X \geq E(X) + a) \tag{8.1}$$

for some threshold $a > 0$. If we have full knowledge of the distribution, say X has some explicit density f_X, then we can compute this directly as

$$\mathbb{P}(X \geq E(X) + a) = \int_{\mathbb{E}(X)+a}^{\infty} f_X(x)dx \tag{8.2}$$

We often do not have the explicit density. We might only know a few summary statistics such as the mean, or the mean and the variance. It turns out, we can say something useful about tail probabilities using concentration inequalities

Definition 8.1 (Concentration Inequalities)**.** Inequalities that give us upper bound on tail probabilities using only minimal information about the distribution

8.1 Markov's Inequality

Definition 8.2 (Markov's Inequality)**.** Let X be a non-negative random variable. Then for any $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{8.3}$$

Proof. Given any $a > 0$, define the indicator random variable

$$\mathbb{I}_{\{X \geq a\}} = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases} \tag{8.4}$$

For non-negative X we have the inequality

$$X \geq a\mathbb{I}_{\{X \geq a\}} \tag{8.5}$$

This is because if $X \geq a$ then $\mathbb{I}_{\{X \geq a\}} = 1$ and so $X \geq a$ which is true by assumption. If $X < a$, then $\mathbb{I}_{\{X \geq a\}} = 0$ and $X \geq 0$ which is true because X is non-negative. Then, take the expectation of both sides

$$\mathbb{E}(X) \geq a\mathbb{P}(X \geq) \tag{8.6}$$

Thus,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{8.7}$$

■

The inequality is sharp i.e. it cannot be improved in some cases. Let $X \sim \text{Bernoulli}(p)$. Then $\mathbb{E}(X) = p$. Take $a = 1$. Then,

$$\mathbb{P}(X \geq 1) = \mathbb{P}(X = 1) = p \tag{8.8}$$

Makrov's inequality says

$$\mathbb{P}(X \geq 1) \leq p \tag{8.9}$$

The Markov inequality is bad for random variables with tails that decay faster than polynomial decays. For example, if $X \sim \text{Exp}(\lambda)$ then $\mathbb{E}(X) = \frac{1}{\lambda}$ and Markov says

$$\mathbb{P}(X \geq a) \leq \frac{1}{\lambda a} \tag{8.10}$$

However,

$$\mathbb{P}(X \geq a) = e^{-\lambda a} \tag{8.11}$$

Exponential decay is much faster than $\frac{1}{a}$ decay. Thus while Markov's inequality is universal (it works for any non-negative random variable with finite mean), it is often a weak bound.

8.2 Chebyshev's Inequality

Definition 8.3 (Chebyshev's Inequality)**.** Let X be a random variable with mean μ and variance σ^2. Then for any $a > 0$

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2} \tag{8.12}$$

This is a two-sided tail bound as it controls the probability that X deviates from its mean by at least a in either direction.

Proof. Chebyshev's inequality is just an application of Markov's inequality to a cleverly chosen non-negative random variable. Define $Y = (X - \mu)^2$. This is non-negative because it is a square. Apply Markov's inequality to Y with threshold a^2. Then,

$$\mathbb{P}(Y \geq a^2) \leq \frac{\mathbb{E}(Y)}{a^2} \tag{8.13}$$

Now,

$$\begin{aligned} \mathbb{P}(Y \geq a^2) &= \mathbb{P}((X - \mu)^2 \geq a^2) && (8.14) \\ &= \mathbb{P}(|X - \mu| \geq a) && (8.15) \end{aligned}$$

Also,

$$\begin{aligned} \mathbb{E}(Y) &= E((X - \mu)^2) && (8.16) \\ &= \mathrm{Var}(X) && (8.17) \\ &= \sigma^2 && (8.18) \end{aligned}$$

Thus,

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2} \tag{8.19}$$

■

Let $a = k\sigma$. By Chebyshev's inequality

$$\mathbb{P}(|X - | \geq k\sigma) \leq \frac{1}{k^2} \tag{8.20}$$

The probability of being more than k standard deviations from the mean is at most $\frac{1}{k^2}$

Example 62

Let X be the number of items produced in a factory during a week with $\mathbb{E}(X) = 50$. What can be said about $\mathbb{P}(X \geq 75)$? If $\mathrm{Var}(X) = 25$, what can be said about $\mathbb{P}(40 \leq X \leq 60)$?

Solution

By Markov's inequality

$$\mathbb{P}(X \geq 75) \leq \frac{50}{75} = \frac{2}{3} \tag{8.21}$$

For part b, by Chebyshev's inequality

$$\begin{aligned}
\mathbb{P}(40 \leq X \leq 60) &= \mathbb{P}(|X - 50| \leq 10) && (8.22)\\
&= 1 - \mathbb{P}(|X - 50| > 10) && (8.23)\\
&\leq 1 - \frac{25}{10^2} && (8.24)\\
&\leq \frac{3}{4} && (8.25)
\end{aligned}$$

8.2.1 The One-Sided Chebyshev Inequality

The standard Chebyshev inequality is two-sided. We may want a one-sided bound, such as knowing the probability of being far above the mean but not below it. That comes from the one-sided Chebyshev inequality.

Definition 8.4 (One-Sided Chebyshev)**.** If X is a random variable with mean 0 and variance σ^2, then for any $a > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2} \tag{8.26}$$

Example 63

The number of items produced during a week is a random variable with $\mathbb{E}(X) = 100$ and $\text{Var}(X) = 400$. Find an upper bound on $\mathbb{P}(X \geq 120)$ using the one-sided Chebyshev.

$$\mathbb{P}(X - 100 \geq 20) \leq \frac{400}{400 + 400} = \frac{1}{2} \tag{8.27}$$

Chebyshev's inequality becomes weak when the variance is large. Consider two random variables X with mean 0 and variance 1 and Y with mean 0 and variance 100. Suppose we want to bound the tail probabilities $\mathbb{P}(|X| \geq 5)$ and $\mathbb{P}(|Y| \geq 5)$ using Chebyshev. For X,

$$\mathbb{P}(|X| \geq 5) \leq \frac{\text{Var}(X)}{5^2} \tag{8.28}$$

$$= \frac{1}{25} \tag{8.29}$$

For Y,

$$\mathbb{P}(|Y| \geq 5) \leq \frac{\text{Var}(Y)}{5} \tag{8.30}$$

$$= \frac{100}{25} \tag{8.31}$$

$$= 4 \tag{8.32}$$

$$\implies \mathbb{P}(|Y| \geq 5) \leq 1 \tag{8.33}$$

This is a useless bound so Chebyshev's inequality is only informative when a^2 is large relative to σ^2

8.2.2 Averaging Reduces Variance

Let $X_1, X_2, \ldots, X_n$ be independent random variables with mean 0 and variance 1. Define the arithmetic mean as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{8.34}$$

We want to determine how $\bar{X}$ behaves. The mean of $\bar{X}$ is

$$\begin{aligned}
\mathbb{E}(\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) && (8.35)\\
&= \frac{1}{n}(0) && (8.36)\\
&= 0 && (8.37)
\end{aligned}$$

The average has the same mean as the individual random variables. The variance is

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) && (8.38)\\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) && (8.39)\\
&= \frac{1}{n}\cdot n \cdot 1 && (8.40)\\
&= \frac{1}{n} && (8.41)
\end{aligned}$$

Thus, the variance of the sample mean is the variance of the individual random variables divided by n. As n grows, the variance of the average shrinks towards 0. The standard deviation of $\bar{X}$ is $\sigma = \frac{1}{\sqrt{n}}$ which decays as n grows. We can apply Chebyshev to $\bar{X}$ which has mean 0 and variance $\frac{1}{n}$. For any $\varepsilon > 0$,

$$\mathbb{P}(|\bar{X}| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} = \frac{1}{n\varepsilon^2} \tag{8.42}$$

Now, for any fixed > 0, as $n \to \infty$, the bound goes to 0. Thus,

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}| \geq \varepsilon) = 0 \tag{8.43}$$

Thus, as the number of samples grow, the probability that the sample mean differs from the true mean by more than ε goes to 0. This is true for any ε, no matter how small. This is known as the weak law of large numbers which we will see soon.

8.3 The Chernoff Inequality

Definition 8.5 (Chernoff Inequality)**.** Let X be a random variable with moment generating function $M_X(t)$. Then for any real number a,

$$\mathbb{P}(X \geq a) \leq e^{-ta} M_X(t)\ \forall t > 0 \tag{8.44}$$

$$\mathbb{P}(X \leq a) \leq e^{-ta} M_X(t)\ \forall t < 0 \tag{8.45}$$

$$\tag{8.46}$$

Proof. The proof of Chernoff's inequality is an application of Markov's inequality to the exponentially transformed variable e^{tX}. Fix $t > 0$, then $X \geq a = tX \geq ta$ (since $t > 0$. This is equal to $e^{tX} \geq e^{ta}$. Then,

$$\mathbb{P}(X \geq a) = \mathbb{P}\left(e^{tX} \geq e^{ta}\right) \tag{8.47}$$

Apply Markov's inequality to get

$$\mathbb{P}\left(e^{tX} \geq e^{ta}\right) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}} = e^{-ta} M_X(t) \tag{8.48}$$

Now for $t < 0$, $tX \geq ta$ is equal to $X \leq a$. Then,

$$\mathbb{P}(X \geq a) = \mathbb{P}(tX \geq ta) = \mathbb{P}\left(e^{tX} \geq e^{ta}\right) \leq e^{-tA} m_X(t) \tag{8.49}$$

■

The key feature of Chernoff is that the bound holds for every choice of $t > 0$. The right-hand side depends on t so we choose t to make the bound as tight as possible:

$$\mathbb{P}(X \geq a) \leq \inf_{t>0} e^{-ta} M_X(t) \tag{8.50}$$

The general recipe for getting a Chernoff bound is

- Compute the MGF $M_X(t)$
- Plug in the Chernoff bound $\mathbb{P}(X \geq a) \leq e^{-ta} M_X(t)$
- Minimise t by taking derivative and setting it equal to 0
- Substitute the optimal t to get the final bound

8.3.1 Chernoff Bound for the Poisson Distribution

Let $X \sim \text{Poi}(\lambda)$. We want to find an exponential tail bound on $\mathbb{P}(X \leq a\lambda)$ for $a > 1$. This is a natural way to express how much above the mean we are interested in. The MGF for a Poisson distribution is

$$M_X(t) = e^{\lambda(e^t - 1)} \tag{8.51}$$

This is finite for all real t. Then, by Chernoff

$$\begin{aligned}
\mathbb{P}(X \geq a\lambda) &\leq e^{-a\lambda t} M_X(t) && (8.52)\\
&= e^{-a\lambda t} e^{\lambda(e^t - 1)} && (8.53)\\
&= e^{\lambda[-at + e^t - 1]} && (8.54)
\end{aligned}$$

We want to minimise the exponent. We do this by maximising the negative of the exponent. Define

$$h(t) = at - e^t + 1 \tag{8.55}$$

Then,

$$h'(t) = a - e^t \tag{8.56}$$

Solving for $h'(t) = 0$ gives $t = \ln(a)$. Substituting this into h gives

$$h(\ln a) = a \ln a - e^{\ln a} + 1 = a \ln(a) - a + 1 \tag{8.57}$$

Thus,

$$\mathbb{P}(X \geq a\lambda) \leq e^{-\lambda(a \ln(a) - a + 1)} \tag{8.58}$$

8.3.2 Chernoff Bound for the Standard Normal

Let $X \sim N(0, 1)$. We know the MGF of X is

$$M_X(t) = e^{\frac{t^2}{2}} \tag{8.59}$$

Then for any $t > 0$,

$$\mathbb{P}(X \geq a) \leq e^{-at} M_X(t) \tag{8.60}$$

$$= e^{-at + \frac{t^2}{2}} \tag{8.61}$$

Thus, we want to minimise the exponent so we maximise $h(t) = at - \frac{t^2}{2} \implies h'(t) = a - t \implies t = a$ Thus,

$$\mathbb{P}(X \geq a) \leq e^{-\frac{a^2}{2}} \tag{8.62}$$

In fact,

$$\mathbb{P}(X \geq a) = \int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt \tag{8.63}$$

This decays so rapidly that most of the contribution to the integral comes from the left endpoint which is $\frac{e^{-\frac{a^2}{2}}}{\sqrt{2\pi}}$

8.3.3 Jensen's Inequality

Definition 8.6 (Convex Functions)**.** A function $f : \mathbb{R} \to \mathbb{R}$ is convex if for any $m \geq 1$, any non-negative weights $t_1, t_2, \ldots, t_m \in [0,1]$ summing to 1 ($t_1 + t_2 + \cdots + t_m = 1$), and any points $x_1, x_2, \ldots, x_m \in \mathbb{R}$ we have

$$f(t_1 x_1 + t_2 x_2 + \cdots + t_m x_m) \leq t_1 f(x_1) + t_2 f(x_2) + \cdots + t_m f(x_m) \tag{8.64}$$

In words: the function value at a weighted average of points is at most the weighted average of function values. The graph of a convex function lies below the chord connecting any two of its points.

Definition 8.7 (Concave Functions)**.** A function $f : \mathbb{R} \to \mathbb{R}$ is concave if for any $m \geq 1$, any non-negative weights $t_1, t_2, \ldots, t_m \in [0,1]$ summing to 1 ($t_1 + t_2 + \cdots + t_m = 1$), and any points $x_1, x_2, \ldots, x_m \in \mathbb{R}$ we have

$$f(t_1 x_1 + t_2 x_2 + \cdots + t_m x_m) \geq t_1 f(x_1) + t_2 f(x_2) + \cdots + t_m f(x_m) \tag{8.65}$$

In words: the function value at a weighted average of points is at least the weighted average of function values. The graph of a convex function lies above the chord connecting any two of its points.

For twice-differentiable functions, there is a clean characterisation in terms of the second derivative

- A twice-differentiable function f is convex on an interval I if $f''(x) \geq 0$ for all $x \in I$
- A twice-differentiable function f is convex on an interval I if $f''(x) \leq 0$ for all $x \in I$

Now we can state Jensen's inequality

Definition 8.8 (Jensen's Inequality)**.** If f is a convex function and X is a random variable, then

$$\mathbb{E}(f(X)) \geq f(\mathbb{E}(X)) \tag{8.66}$$

We know $\mathrm{Var}(X) \geq 0$ which means

$$\mathbb{E}(X^2) - (\mathbb{E}(X))^2 \geq 0 \tag{8.67}$$

which gives

$$\mathbb{E}(X^2) \geq (\mathbb{E}(X))^2 \tag{8.68}$$

This is Jensen's inequality applied to the convex function $f(x) = x^2$. If f was concave then

$$\mathbb{E}(f(X)) \leq f(\mathbb{E}(X)) \tag{8.69}$$

This is because if f is concave, then $-f$ is convex. We can apply Jensen's inequality to $-f$ to get

$$\mathbb{E}(-f(X)) \geq -f(\mathbb{E}(X)) \iff -\mathbb{E}(f(X)) \geq -f(\mathbb{E}(X)) \iff \mathbb{E}(f(X)) \leq f(\mathbb{E}(X)) \tag{8.70}$$

8.4 The Weak Law of Large Numbers

Theorem 8.1 (Weak Law of Large Numbers)**.** If $X_1, X_2, \ldots, X_n$ are independently and identically distributed random variables with mean μ, then for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty \tag{8.71}$$

That is, the probability that the sample mean deviates from the true mean approaches 0 as our sample size increases. We often denote this as

$$\bar{X}_n \to \mu \tag{8.72}$$

So if you flip a coin many times, the proportion of heads gets close to $\frac{1}{2}$. We will omit the proof as we did a rough proof in a previous section. We also have a strong law of large numbers.

Theorem 8.2 (Strong Law of Large Numbers)**.** If $X_1, X_2, \ldots, X_n$ are indpendently and identically distributed with mean μ, then with probability 1

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \text{ as } n \to \infty \tag{8.73}$$

Every random variable X_i is a function $X_i : S \to \mathbb{R}$ where $S$is the sample space. For each individual outcome $\omega \in S$, the values $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$ are a deterministic sequence of real numbers. Thus, the strong law of large numbers says that there is a subset $A \subseteq S$ with $\mathbb{P}(A) = 1$ such that for every $\omega \in A$, the deterministic sequence

$$\frac{X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)}{n} \tag{8.74}$$

converges to μ as n $\to \infty$. The strong law of large numbers says that the convergence to μ holds for almost every outcome of the random experiment. The SLLN is a stronger statement as it implies the WLLN. The converse is not true. This is because the WLLN says that for each individual n, the probability of being far from the mean is small, but the WLLN does not say anything about whether the entire sequence converges.

8.5 The Central Limit Theorem

Theorem 8.3 (Central Limit Theorem)**.** Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2 (both finite). Define

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \tag{8.75}$$

Then the distribution of Z_n converges to the standard normal as $n \to \infty$. Specifically, for any $-\infty < a < \infty$,

$$F_{Z_n}(a) = \mathbb{P}(Z_n \le a) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx = \Phi(a) \tag{8.76}$$

as $n \to \infty$.

We see that the quantity $S_n = X_1 + \cdots + X_n$ has mean $n\mu$ and variance $n\sigma^2$. Thus subtracting the mean and dividing the standard deviation will approach a standard normal distribution which has mean 0 and variance 1. The CLT says that regardless of the original distribution of X_i (as long as the mean and variance are finite), the standardised sum Z_n has approximately the same distribution as a standard normal random variable when n is large. The CLT explains why the normal distribution is so widespread in nature and in statistics. Many quantities in the real world are approximately normal because they arise as sums of many small, roughly independent contributions:

- A person's height is the sum of many small genetic and environmental factors.
- Measurement errors are sums of many small independent errors.
- Stock returns over a long period are sums of many short-term returns.
- The position of a Brownian particle is the sum of many small molecular impacts.

Whenever you see a bell-shaped distribution in nature, the CLT is usually the explanation. The CLT is also the foundation of statistical inference. Confidence intervals, hypothesis tests, and most classical statistical procedures are derived from the CLT. It is the reason we can compute approximate probabilities for sums even when we do not know the underlying distribution.

8.5.1 The CLT for Binomial and Poisson

Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli(p) random variables. Then $S_n = X_1 + \cdots + X_n$is a Bin(n, p) random variable with mean np and variance $np(1-p)$. Then by the CLT,

$$\frac{S_n - np}{\sqrt{np(1-p)}} \to N(0, 1) \tag{8.77}$$

This is the normal approximation to the Binomial which we saw earlier. Now let $X_1, \ldots, X_n$ be independent and identically distributed Poi(1) random variables. Then $S_n = X_1 + \cdots + X_n$ is itself Poisson with parameter n. Each X_i has mean 1 and variance 1 so the CLT gives

$$\frac{S_n - n}{\sqrt{n}} \to N(0,1) \tag{8.78}$$

as $n \to \infty$. In other words, a Poi(λ) random variable with large λ is approximately $N(\lambda, \lambda)$ which is normal with mean λ and variance λ. This is the Normal approximation to the Poisson and it is useful whenever λ is large enough.

Example 64

An astronomer wants to measure the distance from his observatory to a distant star in light years. Each measurement is subject to error from atmospheric conditions and instrument noise. He believes that the measurements are independent and identically distributed random variables with common mean d (the true distance, which he wants to estimate) and common variance 4 (light-years squared). He plans to take N measurements $X_1, X_2, \ldots, X_n$ and use the average $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$ as his estimate of the true distance. How many measurements does he need to take to be 99% sure that his estimate is accurate to within ± 0.5 light years.

Solution

We want

$$\mathbb{P}(|\bar{X}_n - d| < 0.5) \geq 0.99 \tag{8.79}$$

or equivalently

$$\mathbb{P}(|\bar{X}_n - d| \geq 0.5) \leq 0.01 \tag{8.80}$$

We want to find the smallest n for which this holds.

The CLT tells us that

$$\frac{(\bar{X}_n - d)\sqrt{n}}{2} \approx N(0,1) \tag{8.81}$$

where $\sigma = 2$ since $\sigma^2 = 4$. Thus,

$$\mathbb{P}(|\bar{X}_n - d| < 0.5) = \mathbb{P}\left(\left|\frac{(\bar{X}_n - d)\sqrt{n}}{2}\right| < \frac{0.5\sqrt{n}}{2}\right) \tag{8.82}$$

Thus,

$$\left(\left|\frac{(\bar{X}_n - d)\sqrt{n}}{2}\right| < \frac{\sqrt{n}}{4}\right) \tag{8.83}$$

Thus,

$$\mathbb{P}\left(|Z| < \frac{\sqrt{n}}{4}\right) \approx 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1 \tag{8.84}$$

Thus, $n \approx 107$

Example 65

The number of students enrolling in a psychology course is a Poisson random variable with mean 100. The professor will teach two sections if 120 or mroe students enroll. Approximate the probability the professor has to teach two sections.

Solution

We can approximate this as $N(100, 100)$. We apply our continuity correction where $X \geq 120 = X \geq 119.5$. Thus,

$$\mathbb{P}\left(\frac{Z - 100}{10} \geq 1.95\right) \tag{8.85}$$