Probability

Cambridge University Mathematical Tripos: Part IA

4th May 2024

Contents

1 Probability spaces

1.1 Probability spaces and σ-algebras

Definition. Suppose Ω is a set, and $\mathcal{F}$ is a collection of subsets of Ω. We call $\mathcal{F}$ a σ-algebra if
(i) $\Omega \in \mathcal{F}$
(ii) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
(iii) for any countable collection $(A_n)_{n\geq 1}$ with $A_n \in \mathcal{F}$ for all n, we must also have that $\bigcup_n A_n \in \mathcal{F}$

Definition. Suppose that $\mathcal{F}$ is a σ-algebra on Ω. A function $\mathbb{P} : \mathcal{F} \to [0,1]$ is called a probability measure if
(i) $\mathbb{P}(\Omega) = 1$
(ii) for any countable disjoint collection of sets $(A_n)_{n\geq 1}$ in $\mathcal{F}$ ($A_n \in \mathcal{F}$ for all n), then $\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mathbb{P}(A_n)$ (this is called 'countable additivity')
We say that $\mathbb{P}(A)$ is 'the probability of A'. We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, where Ω is the sample space, $\mathcal{F}$ is the σ-algebra, and $\mathbb{P}$ is the probability measure.

Remark. When Ω is countable, we take $\mathcal{F}$ to be all subsets of Ω, i.e. $\mathcal{F} = \mathcal{P}(\Omega)$, its power set.

Definition. The elements of Ω are called outcomes, and the elements of $\mathcal{F}$ are called events.

Note that $\mathbb{P}$ is dependent on $\mathcal{F}$ but not on Ω. We talk about probabilities of *events*, not probabilities of *outcomes*. For example, if you pick a uniform number from the interval $[0,1]$; then the probability of getting any specific outcome is zero—but we can define useful events that have nonzero probabilities.

1.2 Properties of the probability measure

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, since A and A^c are disjoint sets, whose union is Ω
- $\mathbb{P}(\varnothing) = 0$, since it is the complement of Ω
- if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ using the inclusion-exclusion theorem

Example. Consider the following examples of probability spaces and probability measures.

- Rolling a fair 6-sided die:
  - $\Omega = \{1,2,3,4,5,6\}$
  - $\mathcal{F} = \mathcal{P}(\Omega)$
  - $\forall \omega \in \Omega, \mathbb{P}(\{\omega\}) = \frac{1}{6}$, and if $A \subseteq \Omega$ then $\mathbb{P}(A) = \frac{|A|}{6}$
- Equally likely outcomes (more generally). Suppose Ω is some finite set, e.g. $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$. Then we define $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$. In classical probability, this models picking a random element of Ω.

- Picking balls from a bag. Suppose we have n balls with n labels from the set $\{1, \dots, n\}$, indistinguishable by touching. Let us pick $k \leq n$ balls at random from the bag, *without replacement*. Here, 'at random' just means that all possible outcomes are equally likely, and their probability measures should be equal.

  We will take $\Omega = \{A \subseteq \{1, \dots, n\} : |A| = k\}$. Then $|\Omega| = \binom{n}{k}$. Then of course $\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}$, since all outcomes (combinations, in this case) are equally likely.

- Consider a well-shuffled deck of 52 cards, i.e. it is equally likely that each possible permutation of the 52 cards will appear. $\Omega = \{\text{all permutations of 52 cards}\}$, and $|\Omega| = 52!$

  The probability that the top two cards are aces is therefore $\frac{4\times 3\times 50!}{52!} = \frac{1}{221}$, since there are $4 \times 3 \times 50!$ outcomes that produce such an event.

- Consider a string of n random digits from $\{0, \dots, 9\}$. Then $\Omega = \{0, \dots, 9\}^n$, and $|\Omega| = 10^n$. We define $A_k = \{\text{no digit exceeds } k\}$, and $B_k = \{\text{largest digit is } k\}$. Then $\mathbb{P}(B_k) = \frac{|B_k|}{|\Omega|}$. Notice that $B_k = A_k \setminus A_{k-1}$. $|A_k| = (k+1)^n$, so $|B_k| = (k+1)^n - k^n$, so $\mathbb{P}(B_k) = \frac{(k+1)^n - k^n}{10^n}$.

- The birthday problem. There are n people; what is the probability that at least two of them share a birthday? We assume that each year has exactly 365 days, i.e. nobody is born on 29$^{\text{th}}$ of February, and that the probabilities of being born on any given day are equal.

  Let $\Omega = \{1, \dots, 365\}^n$, and $\mathcal{F} = \mathcal{P}(\Omega)$. Since all outcomes are equally likely, we take $\mathbb{P}(\{\omega\}) = \frac{1}{365^n}$. Let $A = \{\text{at least two people share the same birthday}\}$. $A^c = \{\text{all } n \text{ birthdays are different}\}$. Since $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$, it suffices to calculate $\mathbb{P}(A^c)$, which is $\frac{|A^c|}{|\Omega|} = \frac{365!}{(365-n)!365^n}$. So the answer is $\mathbb{P}(A) = 1 - \frac{365!}{(365-n)!365^n}$.

  Note that at $n = 22$, $\mathbb{P}(A) \approx 0.476$ and at $n = 23$, $\mathbb{P}(A) \approx 0.507$. So when there are at least 23 people in a room, the probability that two of them share a birthday is around 50%.

1.3 Combinatorial analysis

Let Ω be a finite set, and suppose $|\Omega| = n$. We want to partition Ω into k disjoint subsets $\Omega_1, \dots, \Omega_k$ with $|\Omega_i| = n_i$ and $\sum_{i=1}^k n_i = n$. How many ways of doing such a partition are there? The result is

$$\underbrace{\binom{n}{n_1}}_{\text{choose first set}} \underbrace{\binom{n-n_1}{n_2}}_{\text{choose second set}} \cdots \underbrace{\binom{n-(n_1+n_2+\cdots+n_{k-1})}{n_k}}_{\text{choose last set}} = \frac{n!}{n_1!n_2!\dots n_k!}$$

So we will write

$$\binom{n}{n_1, \dots, n_k} = \frac{n!}{n_1!n_2!\dots n_k!}$$

Now, let $f : \{1, \dots, k\} \to \{1, \dots, n\}$. f is strictly increasing if $x < y \implies f(x) < f(y)$. f is increasing if $x < y \implies f(x) \leq f(y)$. How many strictly increasing functions f exist? Note that if we know the range of f, the function is completely determined. The range is a subset of $\{1, \dots, n\}$ of size k, i.e. a k-subset of an n-set, which yields $\binom{n}{k}$ choices, and thus there are $\binom{n}{k}$ strictly increasing functions.

How many increasing functions f exist? Let us define a bijection from the set of increasing functions $\{f : \{1, \dots, k\} \to \{1, \dots, n\}\}$ to the set of *strictly* increasing functions $\{g : \{1, \dots, k\} \to \{1, \dots, n+k-1\}\}$. For any increasing function f, we define $g(i) = f(i) + i - 1$. Then g is clearly strictly increasing, and takes values in the range $\{1, \dots, n+k-1\}$. By extension, we can define an increasing function f from any strictly increasing function g. So the total number of increasing functions $f : \{1, \dots, k\} \to \{1, \dots, n\}$ is $\binom{n+k-1}{k}$.

1.4 Stirling's formula

Let (a_n) and (b_n) be sequences. We will write $a_n \sim b_n$ if $\frac{a_n}{b_n} \to 1$ as $n \to \infty$. This is asymptotic equality.

Theorem (Stirling's Formula). $n! \sim n^n \sqrt{2\pi n}\, e^{-n}$ as $n \to \infty$.

Let us first prove the weaker statement $\log(n!) \sim n \log n$.

Proof. Let us define $l_n = \log(n!) = \log 2 + \log 3 + \cdots + \log n$. For $x \in \mathbb{R}$, we write $\lfloor x \rfloor$ for the integer part of x. Note that

$$\log\lfloor x \rfloor \le \log x \le \log\lfloor x+1 \rfloor$$

Let us integrate this from 1 to n.

$$\sum_{k=1}^{n-1} \log k \le \int_1^n \log x \,\mathrm{d}x \le \sum_{k=2}^{n} \log k$$

$$l_{n-1} \le n \log n - n + 1 \le l_n$$

For all n, therefore:

$$n \log n - n + 1 \le l_n \le (n+1)\log(n+1) - (n+1) + 1$$

Dividing through by $n \log n$, we get

$$\frac{l_n}{n \log n} \to 1$$

as $n \to \infty$. □

The following complete proof is non-examinable.

Proof. For any twice-differentiable function f, for any $a < b$ we have

$$\int_a^b f(x)\,\mathrm{d}x = \frac{f(a)+f(b)}{2}(b-a) - \frac{1}{2}\int_a^b (x-a)(b-x)f''(x)\,\mathrm{d}x$$

Now let $f(x) = \log x$, $a = k$ and $b = k+1$. Then

$$\begin{aligned}\int_k^{k+1} \log x \,\mathrm{d}x &= \frac{\log k + \log(k+1)}{2} + \frac{1}{2}\int_k^{k+1} \frac{(x-k)(k+1-x)}{x^2}\,\mathrm{d}x \\ &= \frac{\log k + \log(k+1)}{2} + \frac{1}{2}\int_0^1 \frac{x(1-x)}{(x+k)^2}\,\mathrm{d}x\end{aligned}$$

Let us take the sum for $k = 1, \dots, n-1$ of the equality.

$$\int_1^n \log x \, \mathrm{d}x = \frac{\log((n-1)!) + \log(n!)}{2} + \frac{1}{2} \sum_{k=1}^{n-1} \int_0^1 \frac{x(1-x)}{(x+k)^2} \, \mathrm{d}x$$

$$n \log n - n + 1 = \log(n!) - \frac{\log n}{2} + \sum_{k=1}^{n-1} a_k; \quad a_k = \frac{1}{2} \int_0^1 \frac{x(1-x)}{(x+k)^2} \, \mathrm{d}x$$

$$\log(n!) = n \log n - n + \frac{\log n}{2} + 1 - \sum_{k=1}^{n-1} a_k$$

$$n! = n^n \, e^{-n} \sqrt{n} \, \exp\left(1 - \sum_{k=1}^{n-1} a_k\right)$$

Now, note that

$$a_k \leq \frac{1}{2} \int_0^1 \frac{x(1-x)}{k^2} \, \mathrm{d}x = \frac{1}{12k^2}$$

So the sum of all a_k converges. We set

$$A = \exp\left(1 - \sum_{k=1}^{\infty} a_k\right)$$

and then

$$n! = n^n \, e^{-n} \sqrt{n} A \, \exp\left(\underbrace{\sum_{k=n}^{\infty} a_k}_{\text{converges to zero}}\right)$$

Therefore,

$$n! \sim n^n \sqrt{n} \, e^{-n} A$$

To finish the proof, we must show that $A = \sqrt{2\pi}$. We can utilise the fact that $n! \sim n^n \sqrt{n} \, e^{-n} A$.

$$\begin{aligned} 2^{-2n} \binom{2n}{n} &= 2^{-2n} \cdot \frac{2n!}{(n!)^2} \\ &\sim 2^{-2n} \frac{(2n)^{2n} \cdot \sqrt{2n} \cdot A \cdot e^{-2n}}{n^n \, e^{-n} \sqrt{n} A \, n^n \, e^{-n} \sqrt{n} A} \\ &= \frac{\sqrt{2}}{A\sqrt{n}} \end{aligned}$$

Using a different method, we will prove that $2^{-2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}$, which then forces $A = \sqrt{2\pi}$. Consider

$$I_n = \int_0^{\frac{\pi}{2}} (\cos\theta)^n \, \mathrm{d}\theta; \quad n \geq 0$$

So $I_0 = \frac{\pi}{2}$ and $I_1 = 1$. By integrating by parts,

$$I_n = \frac{n-1}{n} I_{n-2}$$

Therefore,

$$I_{2n} = \frac{2n-1}{2n} I_{2n-2} = \frac{(2n-1)(2n-3)\dots(3)(1)}{(2n)(2n-2)\dots(2)} I_0$$

Multiplying the numerator and denominator by the denominator, we have

$$I_{2n} = \frac{(2n)!}{(n! \cdot 2^n)^2} \cdot \frac{\pi}{2} = 2^{-2n} \frac{2n}{n} \cdot \frac{\pi}{2}$$

In the same way, we can deduce that

$$I_{2n+1} = \frac{(2n)(2n-2)\dots(2)}{(2n+1)(2n-1)\dots(3)(1)} I_1 = \frac{1}{2n+1} \left(2^{-2n} \binom{2n}{n}\right)^{-1}$$

From $I_n = \frac{n-1}{n} I_{n-2}$, we get that

$$\frac{I_n}{I_{n-2}} \to 1$$

as $n \to \infty$. We now want to show that $\frac{I_{2n}}{I_{2n+1}} \to 1$. We see from the definition of I_n that I is a decreasing function of n. Therefore,

$$\frac{I_{2n}}{I_{2n+1}} \leq \frac{I_{2n-1}}{I_{2n+1}} \to 1$$

and also

$$\frac{I_{2n}}{I_{2n+1}} \geq \frac{I_{2n}}{I_{2n-2}} \to 1$$

So

$$\frac{I_{2n}}{I_{2n+1}} \to 1$$

which means that

$$\frac{2^{-2n}\binom{2n}{n}\frac{\pi}{2}}{\left(2^{-2n}\binom{2n}{n}\right)^{-1}\frac{1}{2n+1}} \to 1 \implies \left(2^{-2n}\binom{2n}{n}\right)^2 \frac{\pi}{2}(2n+1) \to 1$$

Therefore,

$$\left(2^{-2n}\binom{2n}{n}\right)^2 \sim \frac{2}{\pi(2n+1)} \sim \frac{1}{\pi n}$$

Finally,

$$A = \sqrt{2\pi}$$

completes the proof. □

1.5 Countable subadditivity

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(A_n)_{n\geq 1}$ be a (not necessarily disjoint) sequence of events in $\mathcal{F}$. Then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Proof. Let us define a new sequence $B_1 = A_1$ and $B_n = A_n \setminus (A_1 \cup A_2 \cup \cdots \cup A_{n-1})$. So by construction, the sequence B_n is a disjoint sequence of events in $\mathcal{F}$. Note further that the union of all B_n is equal to the union of all A_n. So

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right)$$

By the countable additivity axiom,

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

But $B_n \subseteq A_n$. So $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$. Therefore,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

□

1.6 Continuity of probability measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(A_n)_{n\geq 1}$ be an increasing sequence in $\mathcal{F}$, i.e. $A_n \in \mathcal{F}$, and $A_n \subseteq A_{n+1}$. Then $\mathbb{P}(A_n) \leq \mathbb{P}(A_{n+1})$. We want to show that

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n} A_n\right)$$

Proof. Let $B_1 = A_1$, and for all $n \geq 2$, let $B_n = A_n \setminus (A_1 \cup A_2 \cup \cdots \cup A_{n-1})$. Then the union over B_i up to n is equal to the union over A_i up to n. So

$$\mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{k=1}^{n} B_k\right) = \sum_{k=1}^{n} \mathbb{P}(B_k) \to \sum_{k=1}^{\infty} \mathbb{P}(B_k) = \mathbb{P}\left(\bigcup_{n} B_n\right) = \mathbb{P}\left(\bigcup_{n} A_n\right)$$

□

We can say that probability measures are continuous; an increasing sequence of events has a probability which tends to some limit. Similarly, if (A_n) is decreasing, then the limit probability is the probability of the intersection of all A_n.

2 Inclusion-exclusion

2.1 Inclusion-exclusion formula

Suppose that $A, B \in \mathcal{F}$. Then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Now let also $C \in \mathcal{F}$. Then

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

Let $A_1, \ldots, A_n$ be events in $\mathcal{F}$. Then

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = &\sum_{i=1}^{n} \mathbb{P}(A_i) \\ &- \sum_{1\le i_1<i_2\le n} \mathbb{P}\left(A_{i_1} \cap A_{i_2}\right) \\ &+ \sum_{1\le i_1<i_2<i_3\le n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap A_{i_3}\right) \\ &- \cdots \\ &+ (-1)^{n+1}\mathbb{P}(A_1 \cap \cdots \cap A_n)\end{aligned}$$

Or more concisely,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\le i_1<i_2<\cdots<i_k\le n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right)$$

Proof. The case for $n = 2$ has been verified, so we can use induction on n. Now, let us assume this holds for $n - 1$ events.

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \mathbb{P}\left(\left(\bigcup_{i=1}^{n-1} A_i\right) \cup A_n\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\left(\bigcup_{i=1}^{n-1} A_i\right) \cap A_n\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{i=1}^{n-1} (A_i \cap A_n)\right)\end{aligned}$$

Let $B_i = A_i \cap A_n$ for all i. By the inductive hypothesis, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{i=1}^{n-1} B_n\right) \\ &= \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n-1} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) \\ &\quad - \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n-1} \mathbb{P}\left(B_{i_1} \cap \cdots \cap B_{i_k}\right) \\ &\quad + \mathbb{P}(A_n)\end{aligned}$$

which gives the claim as required. □

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with $|\Omega| < \infty$ and $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$. Let $A_1, \ldots, A_n \in \mathcal{F}$. Then

$$|A_1 \cup \cdots \cup A_n| = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \left|A_{i_1} \cap \cdots \cap A_{i_k}\right|$$

2.2 Bonferroni inequalities

Truncating the sum in the inclusion-exclusion formula at the rth term yields an estimate for the probability. The Bonferroni inequalities state that if r is odd, it is an overestimate, and if r is even, it is an underestimate.

$$r \text{ odd} \implies \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{k=1}^{r} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right)$$

$$r \text{ even} \implies \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{k=1}^{r} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right)$$

Proof. Again, we will use induction. The $n = 2$ case is trivial. Suppose that r is odd. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) \tag{$*$}$$

where $B_i = A_i \cap A_n$. Since r is odd,

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) \leq \sum_{k=1}^{r} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n-1} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right)$$

Since $r - 1$ is even, we can apply the inductive hypothesis to $\mathbb{P}\left(\bigcup_{i=1}^{n-1} B_i\right)$.

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} B_i\right) \geq \sum_{k=1}^{r-1} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n-1} \mathbb{P}\left(B_{i_1} \cap \cdots \cap B_{i_k}\right)$$

We can substitute both bounds into $(*)$ to get an overestimate. □

2.3 Counting using inclusion-exclusion

We can apply the inclusion-exclusion formula to count various things. How many functions $f : \{1, \ldots, n\} \to \{1, \ldots, m\}$ are surjective? Let Ω be the set of such functions, and $A = \{f \in \Omega : f \text{ is a surjection}\}$. For all $i \in \{1, \ldots, m\}$, we define $A_i = \{f \in \Omega : i \notin \{f(1), f(2), \ldots, f(n)\}\}$. Then $A = {A_1}^c \cap {A_2}^c \cap \cdots \cap {A_m}^c = (A_1 \cup A_2 \cup \cdots \cup A_m)^c$. Then

$$|A| = |\Omega| - |A_1 \cup \cdots \cup A_m| = m^n - |A_1 \cup \cdots \cup A_m|$$

Now, let us use the inclusion-exclusion formula.

$$|A_1 \cup \cdots \cup A_m| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \left|A_{i_1} \cap \cdots \cap A_{i_k}\right|$$

Note that $A_{i_1} \cap \cdots \cap A_{i_k}$ is the set of functions where k distinct numbers are not included in the function's range. There are $(m-k)^n$ such functions.

$$\begin{aligned}
|A_1 \cup \cdots \cup A_m| &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n} (m-k)^n \\
&= \sum_{k=1}^{n} (-1)^{k+1} \binom{m}{k} (m-k)^n \\
|A| &= m^n - \sum_{k=1}^{n} (-1)^{k+1} \binom{m}{k} (m-k)^n \\
|A| &= \sum_{k=0}^{n} (-1)^{k} \binom{m}{k} (m-k)^n
\end{aligned}$$

2.4 Counting derangements

A derangement is a permutation which has no fixed point, i.e. $\forall i, \sigma(i) \neq i$. We will let Ω be the set of permutations of $\{1, \ldots, n\}$, i.e. S_n. Let A be the set of derangements in Ω. Let us pick a permutation σ at random from Ω. What is the probability that it is a derangement? We define $A_i = \{f \in \Omega : f(i) = i\}$, then $A = A_1^c \cap \cdots \cap A_n^c = \left(\bigcup_{i=1}^n A_i\right)^c$, so $\mathbb{P}(A) = 1 - \mathbb{P}\left(\bigcup_{i=1}^n A_i\right)$. By the inclusion-exclusion formula,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) \\
&= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \frac{(n-k)!}{|\Omega|} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \frac{n!}{k!(n-k)!} \cdot \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \frac{1}{k!}
\end{aligned}$$

So

$$\mathbb{P}(A) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} = \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

As $n \to \infty$, this value tends to $e^{-1} \approx 0.3678$.

3 Independence and dependence of events

3.1 Independence of events

Definition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A, B \in \mathcal{F}$. A and B are called independent if

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

We write $A \perp B$, or $A \perp\!\!\!\perp B$. A countable collection of events (A_n) is said to be independent if for all distinct $i_1, \dots, i_k$, we have

$$\mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) = \prod_{j=1}^{k} \mathbb{P}\left(A_{i_j}\right)$$

Remark. To show that a collection of events is independent, it is insufficient to show that events are pairwise independent. For example, consider tossing a fair coin twice, so $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$. $\mathbb{P}(\{\omega\}) = \frac{1}{4}$. Consider the events A, B, C where

$$A = \{(0,0),(0,1)\}; \quad B = \{(0,0),(1,0)\}; \quad C = \{(1,0),(0,1)\}$$

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$$

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{(0,0)\}) = \frac{1}{4} = \mathbb{P}(A) \cdot \mathbb{P}(B)$$
$$\mathbb{P}(A \cap C) = \mathbb{P}(\{(0,1)\}) = \frac{1}{4} = \mathbb{P}(A) \cdot \mathbb{P}(C)$$
$$\mathbb{P}(B \cap C) = \mathbb{P}(\{(1,0)\}) = \frac{1}{4} = \mathbb{P}(B) \cdot \mathbb{P}(C)$$

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\varnothing) = 0$$

Claim. If $A \perp B$, then $A \perp B^c$.

Proof.

$$\begin{aligned}
\mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}(A) - \mathbb{P}(A) \cdot \mathbb{P}(B) \\
&= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\
&= \mathbb{P}(A)\,\mathbb{P}(B^c)
\end{aligned}$$

as required. □

3.2 Conditional probability

Definition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. We define the conditional probability of A given B, written $\mathbb{P}(A \mid B)$, as

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

Note that if A and B are independent, then

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

Claim. Suppose that (A_n) is a disjoint sequence in $\mathcal{F}$. Then

$$\mathbb{P}\left(\bigcup A_n \;\middle|\; B\right) = \sum_n \mathbb{P}(A_n \mid B)$$

This is the countable additivity property for conditional probability.

Proof.

$$\begin{aligned}\mathbb{P}\left(\bigcup A_n \;\middle|\; B\right) &= \frac{\mathbb{P}\left(\left(\bigcup A_n\right) \cap B\right)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}\left(\bigcup (A_n \cap B)\right)}{\mathbb{P}(B)}\end{aligned}$$

By countable additivity, since the $(A_n \cap B)$ are disjoint,

$$\begin{aligned}&= \sum_n \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} \\ &= \sum_n \mathbb{P}(A_n \mid B)\end{aligned}$$

□

We can think of $\mathbb{P}(\,\cdot \mid B)$ as a new probability measure for the same Ω.

3.3 Law of total probability

Claim. Suppose (B_n) is a disjoint collection of events in $\mathcal{F}$, such that $\bigcup B = \Omega$, and for all n, we have $\mathbb{P}(B_n) > 0$. If $A \in \mathcal{F}$ then

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \mid B_n) \cdot \mathbb{P}(B_n)$$

Proof.

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap \Omega) \\ &= \mathbb{P}\left(A \cap \left(\bigcup B_n\right)\right) \\ &= \mathbb{P}\left(\bigcup (A \cap B_n)\right)\end{aligned}$$

By countable additivity,

$$\begin{aligned}&= \sum_n \mathbb{P}(A \cap B_n) \\ &= \sum_n \mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)\end{aligned}$$

□

3.4 Bayes' formula

Claim. Suppose (B_n) is a disjoint sequence of events with $\bigcup B_n = \Omega$ and $\mathbb{P}(B_n) > 0$ for all n. Then

$$\mathbb{P}(B_n \mid A) = \frac{\mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)}{\sum_k \mathbb{P}(A \mid B_k)\,\mathbb{P}(B_k)}$$

Proof.

$$\begin{aligned}\mathbb{P}(B_n \mid A) &= \frac{\mathbb{P}(B_n \cap A)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)}{\mathbb{P}(A)}\end{aligned}$$

By the law of total probability,

$$= \frac{\mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)}{\sum_k \mathbb{P}(A \mid B_k)\,\mathbb{P}(B_k)}$$

□

Note that on the right hand side, the numerator appears somewhere in the denominator. This formula is the basis of Bayesian statistics. It allows us to reverse the direction of a conditional probability—knowing the probabilities of the events (B_n), and given a model of $\mathbb{P}(A \mid B_n)$, we can calculate the posterior probabilities of B_n given that A occurs.

3.5 Bayes' formula for medical tests

Consider the probability of getting a false positive on a test for a rare condition. Suppose 0.1% of the population have condition A, and we have a test which is positive for 98% of the affected population, and 1% of those unaffected by the disease. Picking an individual at random, what is the probability that they suffer from A, given that they have a positive test?

We define A to be the set of individuals suffering from the condition, and P is the set of individuals testing positive. Then by Bayes' formula,

$$\mathbb{P}(A \mid P) = \frac{\mathbb{P}(P \mid A)\,\mathbb{P}(A)}{\mathbb{P}(P \mid A)\,\mathbb{P}(A) + \mathbb{P}(P \mid A^c)\,\mathbb{P}(A^c)} = \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.01 \cdot 0.999} \approx 0.09 = 9\%$$

Why is this so low? We can rewrite this instance of Bayes' formula as

$$\mathbb{P}(A \mid P) = \frac{1}{1 + \frac{\mathbb{P}(P|A^c)\mathbb{P}(A^c)}{\mathbb{P}(P|A)\mathbb{P}(A)}}$$

Here, $\mathbb{P}(A^c) \approx 1, \mathbb{P}(P \mid A) \approx 1$. So

$$\mathbb{P}(A \mid P) \approx \frac{1}{1 + \frac{\mathbb{P}(P|A^c)}{\mathbb{P}(A)}}$$

So this is low because $\mathbb{P}(P \mid A^c) \gg \mathbb{P}(A)$. Suppose that there is a population of 1000 people and about 1 suffers from the disease. Among the 999 not suffering from A, about 10 will test positive. So there will be about 11 people who test positive, and only 1 out of 11 (9%) of those actually has the disease.

3.6 Probability changes under extra knowledge

Consider these three statements:

(a) I have two children, (at least) one of whom is a boy.

(b) I have two children, and the eldest one is a boy.

(c) I have two children, one of whom is a boy born on a Thursday.

What is the probability that I have two boys, given a, b or c? Since no further information is given, we will assume that all outcomes are equally likely. We define:

- BG is the event that the elder sibling is a boy, and the younger is a girl;
- GB is the event that the elder sibling is a girl, and the younger is a boy;
- BB is the event that both children are boys; and
- GG is the event that both children are girls.

Now, we have

(a) $\mathbb{P}(BB \mid BB \cup BG \cup GB) = \frac{1}{3}$

(b) $\mathbb{P}(BB \mid BB \cup BG) = \frac{1}{2}$

(c) Let us define GT to be the event that the elder sibling is a girl, and the younger is a boy born on a Thursday, and define TN to be the event that the elder sibling is a boy born on a Thursday

and the younger is a boy not born on a Thursday, and other events are defined similarly. So

$$\begin{aligned}\mathbb{P}(TT \cup TN \cup NT \mid GT \cup TG \cup TT \cup TN \cup NT) &= \frac{\mathbb{P}(TT \cup TN \cup NT)}{\mathbb{P}(GT \cup TG \cup TT \cup TN \cup NT)} \\ &= \frac{\frac{1}{2}\frac{1}{7}\frac{1}{2}\frac{1}{7} + 2 \cdot \frac{1}{2}\frac{1}{7}\frac{6}{2}\frac{6}{7}}{2 \cdot \frac{1}{2}\frac{1}{2}\frac{1}{7} + \frac{1}{2}\frac{1}{7}\frac{1}{2}\frac{1}{7} + 2 \cdot \frac{1}{2}\frac{1}{7}\frac{6}{2}\frac{6}{7}} \\ &= \frac{13}{27} \approx 48\%\end{aligned}$$

3.7 Simpson's paradox

Consider admissions by men and women from state and independent schools to a university given by the tables

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

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

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

This is seemingly a paradox; both women and men are more likely to be admitted if they come from a state school, but when looking at all applicants, they are more likely to be admitted if they come from an independent school. This is called Simpson's paradox; it arises when we aggregate data from disparate populations. Let A be the event that an individual is admitted, B be the event that an individual is a man, and C be the event that an individual comes from a state school. We see that

$$\begin{gathered}\mathbb{P}(A \mid B \cap C) > \mathbb{P}(A \mid B \cap C^c) \\ \mathbb{P}(A \mid B^c \cap C) > \mathbb{P}(A \mid B^c \cap C^c) \\ \mathbb{P}(A \mid C) < \mathbb{P}(A \mid C^c)\end{gathered}$$

First, note that

$$\begin{aligned}\mathbb{P}(A \mid C) &= \mathbb{P}(A \cap B \mid C) + \mathbb{P}(A \cap B^c \mid C) \\ &= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} + \frac{\mathbb{P}(A \cap B^c \cap C)}{\mathbb{P}(C)} \\ &= \frac{\mathbb{P}(A \mid B \cap C)\,\mathbb{P}(B \cap C)}{\mathbb{P}(C)} + \frac{\mathbb{P}(A \mid B^c \cap C)\,\mathbb{P}(B^c \cap C)}{\mathbb{P}(C)} \\ &= \mathbb{P}(A \mid B \cap C)\,\mathbb{P}(B \mid C) + \mathbb{P}(A \mid B^c \cap C)\,\mathbb{P}(B^c \mid C) \\ &> \mathbb{P}(A \mid B \cap C^c)\,\mathbb{P}(B \mid C) + \mathbb{P}(A \mid B^c \cap C^c)\,\mathbb{P}(B^c \mid C)\end{aligned}$$

Let us also assume that $\mathbb{P}(B \mid C) = \mathbb{P}(B \mid C^c)$. Then

$$\begin{aligned}\mathbb{P}(A \mid C) &> \mathbb{P}(A \mid B \cap C^c)\,\mathbb{P}(B \mid C^c) + \mathbb{P}(A \mid B^c \cap C^c)\,\mathbb{P}(B^c \mid C^c) \\ &= \mathbb{P}(A \mid C^c)\end{aligned}$$

So we needed to further assume that $\mathbb{P}(B \mid C) = \mathbb{P}(B \mid C^c)$ in order for the 'intuitive' result to hold. The assumption was not valid in the example, so the result did not hold.

4 Discrete distributions

4.1 Discrete distributions

In a discrete probability distribution on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, Ω is either finite or countable, i.e. $\Omega = \{\omega_1, \omega_2, \dots\}$, and as stated before, $\mathcal{F}$ is the power set of Ω. If we know $\mathbb{P}(\{\omega_i\})$, then this completely determines $\mathbb{P}$. Indeed, let $A \subseteq \Omega$, then

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i\,:\,\omega_i \in A} \{\omega_i\}\right) = \sum_{i\,:\,\omega_i \in A} \mathbb{P}(\{\omega_i\})$$

by countable additivity. We will see later that this is not true if Ω is uncountable. We write $p_i = \mathbb{P}(\{\omega_i\})$, and we then call this a discrete probability distribution. It has the following key properties:

- $p_i \geq 0$
- $\sum_i p_i = 1$

4.2 Bernoulli distribution

We model the outcome of a test with two outcomes (e.g. the toss of a coin) with the Bernoulli distribution. Let $\Omega = \{0, 1\}$. We will denote $p = p_1$, then clearly $p_0 = 1 - p$.

4.3 Binomial distribution

The binomial distribution B has parameters $N \in \mathbb{Z}^+$, $p \in [0, 1]$. This distribution models a sequence of N independent Bernoulli distributions of parameter p. We then count the amount of 'successes', i.e. trials in which the result was 1. $\Omega = \{0, 1, \dots, N\}$.

$$\mathbb{P}(\{k\}) = p_k = \binom{N}{k} p^k (1-p)^{N-k}$$

4.4 Multinomial distribution

The multinomial distribution is a generalisation of the binomial distribution. M has parameters $N \in \mathbb{Z}^+$ and $p_1, p_2, \dots \in [0, 1]$ where $\sum_{i=1}^k p_i = 1$. This models a sequence of N independent trials in which a number from 1 to N is selected, where the probability of selecting i is p_i. $\Omega = \left\{(n_1, \dots, n_k) \in \mathbb{N}^k : \sum_{i=1}^k n_i = N\right\}$, in other words, ordered partitions of N. Therefore

$$\begin{aligned}\mathbb{P}(n_1 \text{ outcomes had value } 1, \dots, n_k \text{ outcomes had value } k) &= \mathbb{P}((n_1, \dots, n_k)) \\ &= \binom{N}{n_1, \dots, n_k} p_1^{n_1} \dots p_k^{n_k}\end{aligned}$$

4.5 Geometric distribution

Consider a Bernoulli distribution of parameter p. The geometric distribution models running this trial many times independently until the first 'success' (i.e. the first result of value 1). Then $\Omega = \{1, 2, \dots\} = \mathbb{Z}^+$. Then

$$p_k = (1-p)^{k-1}p$$

We can compute the infinite geometric series $\sum p_k$ which gives 1. We could alternatively model the distribution using $\Omega' = \{0, 1, \dots\} = \mathbb{N}$ which records the amount of failures before the first success. Then

$$p'_k = (1-p)^k p$$

Again, the sum converges to 1.

4.6 Poisson distribution

This is used to model the number of occurrences of an event in a given interval of time. $\Omega = \{0, 1, 2, \dots\} = \mathbb{N}$. This distribution has one parameter $\lambda \in \mathbb{R}$. We have

$$p_k = e^{-\lambda}\frac{\lambda^k}{k!}$$

Then

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

Suppose customers arrive into a shop during the time interval $[0, 1]$. We will subdivide $[0, 1]$ into N intervals $\left[\frac{i-1}{N}, \frac{i}{N}\right]$. In each interval, a single customer arrives with probability p, independent of other time intervals. In this example,

$$\mathbb{P}\,(k \text{ customers arrive}) = \binom{N}{k}p^k(1-p)^{N-k}$$

Let $p = \frac{\lambda}{N}$ for $\lambda > 0$. We will show that as $N \to \infty$, this binomial distribution converges to the Poisson distribution.

$$\begin{aligned}\binom{N}{k}p^k(1-p)^{N-k} &= \frac{N!}{k!(N-k)!}\left(\frac{\lambda}{n}\right)^k\cdot\left(1-\frac{\lambda}{n}\right)^{N-k}\\ &= \frac{\lambda_k}{k!}\cdot\frac{N!}{N^k(N-k)!}\cdot\left(1-\frac{\lambda}{N}\right)^{N-k}\\ &\to \frac{\lambda_k}{k!}\cdot 1\cdot e^{-\lambda}\end{aligned}$$

which matches the Poisson distribution.

5 Discrete random variables

5.1 Random variables

Definition. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A random variable X is a function $X : \Omega \to \mathbb{R}$ satisfying

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for any given x.

Suppose $A \subseteq \mathbb{R}$. Then typically we write

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

as shorthand. Given $A \in \mathcal{F}$, we define the indicator of A to be

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

Because $A \in \mathcal{F}$, 1_A is a random variable. Suppose X is a random variable. We define the probability distribution function of X to be

$$F_X : \mathbb{R} \to [0,1]; \quad F_X(x) = \mathbb{P}(X \leq x)$$

Definition. $(X_1, \dots, X_n)$ is called a random variable in $\mathbb{R}^n$ if $(X_1, \dots, X_n) : \Omega \to \mathbb{R}^n$, and for all $x_1, \dots, x_n \in \mathbb{R}$ we have

$$\{X_1 \leq x_1, \dots, X_n \leq x_n\} = \{\omega : X_1(\omega) \leq x_1, \dots, X_n(\omega) \leq x_n\} \in \mathcal{F}$$

This definition is equivalent to saying that $X_1, \dots, X_n$ are all random variables in $\mathbb{R}$. Indeed,

$$\{X_1 \leq x_1, \dots, X_n \leq x_n\} = \{X_1 \leq x_1\} \cap \cdots \cap \{X_n \leq x_n\}$$

which, since $\mathcal{F}$ is a σ-algebra, is an element of $\mathcal{F}$.

Definition. A random variable X is called discrete if it takes values in a countable set. Suppose X takes values in the countable set S. For every $x \in S$, we write

$$p_x = \mathbb{P}(X = x) = \mathbb{P}(\{\omega : X(\omega) = x\})$$

We call $(p_x)_{x \in S}$ the probability mass function of X, or the distribution of X. If (p_x) is Bernoulli for example, then we say that X is a Bernoulli (or such) random variable, or that X has the Bernoulli distribution.

Definition. Suppose $X_1, \dots, X_n$ are discrete random variables taking variables in $S_1, \dots, S_n$. We say that the random variables $X_1, \dots, X_n$ are independent if

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) \quad \forall x_1 \in S_1, \dots, x_n \in S_n$$

As an example, suppose we toss a p-biased coin n times independently. Let $\Omega = \{0,1\}^n$. For every $\omega \in \Omega$,

$$p_\omega = \prod_{k=1}^{n} p^{\omega_k}(1-p)^{1-\omega_k}; \quad \text{where we write } \omega = (\omega_1, \dots, \omega_n)$$

We define a set of discrete random variables $X_k(\omega) = \omega_k$. Then X_k gives the output of the kth toss. We have

$$\mathbb{P}(X_k = 1) = \mathbb{P}(\omega_k = 1) = p; \quad \mathbb{P}(X_k = 0) = \mathbb{P}(\omega_k = 0) = 1 - p$$

So X_k has the Bernoulli distribution with parameter p. We can also show that the X_i are independent. Let $x_1, \dots, x_n \in \{0, 1\}$. Then

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) &= \mathbb{P}(\omega = (x_1, \dots, x_n)) \\
&= p_{(x_1,\dots,x_n)} \\
&= \prod_{k=1}^{N} p^{x_k}(1-p)^{1-x_k} \\
&= \prod_{k=1}^{N} \mathbb{P}(X_k = x_k)
\end{aligned}$$

as required. Now, we define $S_n(\omega) = X_1(\omega) + \cdots + X_n(\omega)$. This is the number of heads in N tosses. So $S_n : \Omega \to \{0, \dots, N\}$, and

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

So S_n has the binomial distribution with parameters n and p.

5.2 Expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space such that Ω is countable. Let $X : \Omega \to \mathbb{R}$ be a random variable, which is necessarily discrete. We say that X is non-negative if $X \geq 0$. We define the expectation of X to be

$$\mathbb{E}[X] = \sum_{\omega} X(\omega) \cdot \mathbb{P}(\{\omega\})$$

We will write

$$\Omega_X = \{X(\omega) : \omega \in \Omega\}$$

So

$$\Omega = \bigcup_{x \in \Omega_X} \{X = x\}$$

So we have partitioned Ω using X. Note that

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega} X(\omega)\mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} \sum_{\omega \in \{X=x\}} X(\omega)\mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} \sum_{\omega \in \{X=x\}} x\mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} x\mathbb{P}(\{X = x\})
\end{aligned}$$

which matches the more familiar definition of the expectation; the average of the values taken by X, weighted by the probability of the event occurring. So

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x p_x$$

5.3 Expectation of binomial distribution

Let $X \sim \text{Bin}(N, p)$. Then

$$\forall k = 0, \dots, N, \quad \mathbb{P}(X = k) = \binom{N}{k} p^k (1-p)^{N-k}$$

So using the second definition,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{N} k\mathbb{P}(X = k) \\
&= \sum_{k=0}^{N} k\binom{n}{k} p^k (1-p)^{N-k} \\
&= \sum_{k=0}^{N} \frac{k \cdot N!}{k! \cdot (N-k)!} p^k (1-p)^{N-k} \\
&= \sum_{k=1}^{N} \frac{(N-1)! \cdot N \cdot p}{(k-1)! \cdot (N-k)!} p^{k-1} (1-p)^{N-k} \\
&= Np \sum_{k=1}^{N} \binom{N-1}{k-1} p^{k-1} (1-p)^{N-k} \\
&= Np \sum_{k=0}^{N-1} \binom{N-1}{k} p^k (1-p)^{N-1-k} \\
&= Np(p + 1 - p)^{N-1} \\
&= Np
\end{aligned}$$

5.4 Expectation of Poisson distribution

Let $X \sim \text{Poi}(\lambda)$, so

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

Hence

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

5.5 Expectation of a general random variable

Let X be a general (not necessarily non-negative) discrete random variable. Then we define

$$X^+ = \max(X, 0); \quad X^- = \max(-X, 0)$$

Then $X = X^+ - X^-$. Note that X^+ and X^- are non-negative random variables, which has a well-defined expectation. So if at least one of $\mathbb{E}[X^+], \mathbb{E}[X^-]$ is finite, we define

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$$

If both are infinite, then we say that the expectation of X is not defined. Whenever we write $\mathbb{E}[X]$, it is assumed to be well-defined. If $\mathbb{E}[|X|] < \infty$, we say that X is integrable. When $\mathbb{E}[X]$ is well-defined, we have again that

$$\mathbb{E}[X] = \sum_{x \in \Omega_x} x \cdot \mathbb{P}(X = x)$$

5.6 Properties of the expectation

The following properties follow immediately from the definition.

(i) If $X \geq 0$, then $\mathbb{E}[X] \geq 0$.

(ii) If $X \geq 0$ and $\mathbb{E}[X] = 0$, then $\mathbb{P}(X = 0) = 1$.

(iii) If $c \in \mathbb{R}$, then $\mathbb{E}[cX] = c\mathbb{E}[X]$, and $\mathbb{E}[c + X] = c + \mathbb{E}[X]$.

(iv) If X, Y are two integrable random variables, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

(v) More generally, let $c_1, \ldots, c_n \in \mathbb{R}$ and $X_1, \ldots, X_n$ integrable random variables. Then

$$\mathbb{E}[c_1 X_1 + \cdots + c_n X_n] = c_1 \mathbb{E}[X_1] + \cdots + c_n \mathbb{E}[X_n]$$

So the expectation is a linear operator over finitely many inputs.

5.7 Countable additivity for the expectation

Suppose $X_1, X_2, \ldots$ are non-negative random variables. Then

$$\mathbb{E}\left[\sum_n X_n\right] = \sum_n \mathbb{E}[X_n]$$

The non-negativity constraint allows us to guarantee that the sums are well-defined; they could be infinite, but at least their values are well-defined. We will construct a proof assuming that Ω is countable, however the result holds regardless of the choice of Ω.

Proof.

$$\begin{aligned}
\mathbb{E}\left[\sum_n X_n\right] &= \sum_\omega \sum_n X_n(\omega) \mathbb{P}(\{\omega\}) \\
&= \sum_n \sum_\omega X_n(\omega) \mathbb{P}(\{\omega\}) \\
&= \sum_n \mathbb{E}[X_n]
\end{aligned}$$

□

We are allowed to rearrange the sums since all relevant terms are non-negative.

5.8 Expectation of indicator function

If $X = 1(A)$ where $A \in \mathcal{F}$, then $\mathbb{E}[X] = \mathbb{P}(A)$. This is obvious from the second definition of the expectation.

5.9 Expectation under function application

If $g : \mathbb{R} \to \mathbb{R}$, we can define $g(X)$ to be the random variable given by

$$g(X)(\omega) = g(X(\omega))$$

Then

$$\mathbb{E}[g(X)] = \sum_{x \in \Omega_X} g(x) \cdot \mathbb{P}(X = x)$$

Proof. Let $Y = g(X)$. Then

$$\mathbb{E}[Y] = \sum_{y \in \Omega_Y} y \cdot \mathbb{P}(Y = y)$$

Note that

$$\begin{aligned}
\{Y = y\} &= \{\omega : Y(\omega) = y\} \\
&= \{\omega : g(X(\omega)) = y\} \\
&= \{\omega : X(\omega) \in g^{-1}(\{y\})\} \\
&= \{X \in g^{-1}(\{y\})\}
\end{aligned}$$

where $g^{-1}(\{y\})$ is the set of all x such that $g(x) \in \{y\}$. So

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

□

5.10 Calculating expectation with cumulative probabilities

If $X \geq 0$ and takes integer values, then

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

Proof. Since X takes non-negative integer values,

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

This represents the fact that any integer is the sum of that many ones, e.g. $4 = 1+1+1+1+0+0+\ldots$ to infinity. Taking the expectation of the above formula, using that $\mathbb{E}[1(A)] = \mathbb{P}(A)$ and countable additivity, we have the result as claimed. □

5.11 Inclusion-exclusion formula with indicators

We can provide another proof of the inclusion-exclusion formula, using some basic properties of indicator functions.

- $1(A^c) = 1 - 1(A)$
- $1(A \cap B) = 1(A) \cdot 1(B)$
- Following from the above, $1(A \cup B) = 1 - (1 - 1(A))(1 - 1(B))$.

More generally,

$$1(A_1 \cup \cdots \cup A_n) = 1 - \prod_{i=1}^{n} (1 - 1(A_i))$$

which gives the inclusion-exclusion formula. Taking the expectation of both sides, we can see that

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i_1 < i_2} \mathbb{P}\left(A_{i_1} \cap A_{i_2}\right) + \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap \cdots \cap A_n)$$

which is the result as previously found.

6 Variance and covariance

6.1 Variance

Let X be a random variable, and $r \in \mathbb{N}$. If it is well-defined, we call $\mathbb{E}[X^r]$ the rth moment of X. We define the variance of X by

$$\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

If the variance is small, X is highly concentrated around $\mathbb{E}[X]$. If the variance is large, X has a wide distribution including values not necessarily near $\mathbb{E}[X]$. We call $\sqrt{\operatorname{Var}(X)}$ the standard deviation of X, denoted with σ. The variance has the following basic properties:

- $\operatorname{Var}(X) \geq 0$, and if $\operatorname{Var}(X) = 0$, $\mathbb{P}(X = \mathbb{E}[X]) = 1$.
- If $c \in \mathbb{R}$, then $\operatorname{Var}(cX) = c^2 \operatorname{Var}(X)$, and $\operatorname{Var}(X + c) = \operatorname{Var}(X)$.
- $\operatorname{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$. This follows since

$$\begin{aligned}
\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2
\end{aligned}$$

- $\operatorname{Var}(X) = \min_{c \in \mathbb{R}} \mathbb{E}\left[(X - c)^2\right]$, and this minimum is achieved at $c = \mathbb{E}[X]$. Indeed, if we let $f(c) = \mathbb{E}\left[(X - c)^2\right]$, then $f(c) = \mathbb{E}\left[X^2\right] - 2c\mathbb{E}[X] + c^2$. Minimising f, we get $f(\mathbb{E}[X]) = \operatorname{Var}(X)$ as required.

As an example, consider $X \sim \operatorname{Bin}(n, p)$. Then $\mathbb{E}[X] = np$, as we found before. Note that we can also represent this binomial distribution as the sum of n Bernoulli distributions of parameter p to get the same result. The variance of X is

$$\operatorname{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

In fact, in order to compute $\mathbb{E}\left[X^2\right]$ it is easier to find $\mathbb{E}\left[X(X-1)\right]$.

$$\begin{aligned}
\mathbb{E}\left[X(X-1)\right] &= \sum_{k=2}^{n} k\cdot(k-1)\cdot\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k} \\
&= \sum_{k=2}^{n} \frac{k(k-1)n!p^k(1-p)^{n-k}}{(n-k)!k!} \\
&= \sum_{k=2}^{n} \frac{n!p^k(1-p)^{n-k}}{((n-2)-(k-2))!(k-2)!} \\
&= n(n-1)p^2\sum_{k=2}^{n}\binom{n-2}{k-2}p^{k-2}(1-p)^{n-k} \\
&= n(n-1)p^2
\end{aligned}$$

Hence,

$$\mathrm{Var}\left(X\right) = \mathbb{E}\left[X(X-1)\right] + \mathbb{E}\left[X\right] - \mathbb{E}\left[X\right]^2 = n(n-1)p^2 + np - (np)^2 = np(1-p)$$

As a second example, if $X \sim \mathrm{Poi}(\lambda)$, we have $\mathbb{E}\left[X\right] = \lambda$. Because of the factorial term, it is easier to use $X(X-1)$ than X^2.

$$\begin{aligned}
\mathbb{E}\left[X(X-1)\right] &= \sum_{k=2}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda_{k-2}}{(k-2)!}\cdot\lambda^2 \\
&= \lambda^2
\end{aligned}$$

Hence,

$$\mathrm{Var}\left(X\right) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

6.2 Covariance

Definition. Let X and Y be random variables. Their covariance is defined

$$\mathrm{Cov}\left(X,Y\right) = \mathbb{E}\left[(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right])\right]$$

It is a measure of how dependent X and Y are.

Immediately we can deduce the following properties.

- $\mathrm{Cov}\left(X,Y\right) = \mathrm{Cov}\left(Y,X\right)$
- $\mathrm{Cov}\left(X,X\right) = \mathrm{Var}\left(X\right)$
- $\mathrm{Cov}\left(X,Y\right) = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\cdot\mathbb{E}\left[Y\right]$. Indeed, $(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right]) = XY - X\mathbb{E}\left[Y\right] - Y\mathbb{E}\left[X\right] + \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$ and the result follows.
- Let $c\in\mathbb{R}$. Then $\mathrm{Cov}\left(cX,Y\right) = c\,\mathrm{Cov}\left(X,Y\right)$, and $\mathrm{Cov}\left(c+X,Y\right) = \mathrm{Cov}\left(X,Y\right)$.

- $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)$. Indeed, we have

  $\operatorname{Var}(X+Y) = \mathbb{E}\left[(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^2\right]$ which gives

  $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] + \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] + 2\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$ as required.
- For all $c \in \mathbb{R}$, $\operatorname{Cov}(c, X) = 0$
- If X, Y, Z are random variables, then $\operatorname{Cov}(X+Y, Z) = \operatorname{Cov}(X,Z) + \operatorname{Cov}(Y,Z)$. More generally, for $c_1, \dots, c_n, d_1, \dots, d_m$ real numbers, and for $X_1, \dots, X_n, Y_1, \dots, Y_m$ random variables, we have

$$\operatorname{Cov}\left(\sum_{i=1}^{n} c_i X_i, \sum_{j=1}^{m} d_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} c_i d_j \operatorname{Cov}\left(X_i, Y_j\right)$$

  In particular, if we apply this to $X_i = Y_i$, and $c_i = d_i = 1$, then we have

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

6.3 Expectation of functions of a random variable

Recall that X and Y are independent if for all x and y,

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

We would like to prove that given positive functions $f, g : \mathbb{R} \to \mathbb{R}_+$, if X and Y are independent we have

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]$$

Proof.

$$\begin{aligned}
\mathbb{E}[f(X)g(Y)] &= \sum_{(x,y)} f(x)g(y)\mathbb{P}(X = x, Y = y) \\
&= \sum_{(x,y)} f(x)g(y)\mathbb{P}(X = x)\,\mathbb{P}(Y = y) \\
&= \sum_{x} f(x)\mathbb{P}(X = x) \cdot \sum_{y} g(y)\mathbb{P}(Y = y) \\
&= \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]
\end{aligned}$$

□

The same result holds for general functions, provided the required expectations exist.

6.4 Covariance of independent variables

Suppose X and Y are independent. Then

$$\operatorname{Cov}(X, Y) = 0$$

This is because

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right] \\ &= \mathbb{E}[X-\mathbb{E}[X]]\cdot\mathbb{E}[Y-\mathbb{E}[Y]] \\ &= 0\cdot 0 \\ &= 0\end{aligned}$$

In particular, we can deduce that

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$$

Note, however, that the covariance being equal to zero does not imply independence. For instance, let X_1, X_2, X_3 be independent Bernoulli random variables with parameter $\frac{1}{2}$. Let us now define $Y_1 = 2X_1 - 1$, $Y_2 = 2X_2 - 1$, and $Z_1 = X_3 Y_1$, $Z_2 = X_3 Y_2$. Now, we have

$$\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = \mathbb{E}[Z_1] = \mathbb{E}[Z_2] = 0$$

We can find that

$$\operatorname{Cov}(Z_1, Z_2) = \mathbb{E}[Z_1 \cdot Z_2] = \mathbb{E}\left[X_3^2 Y_1 Y_2\right] = \mathbb{E}\left[X_3^2\right]\cdot 0 \cdot 0 = 0$$

However, Z_1 and Z_2 are in fact not independent. Since Y_1, Y_2 are never zero,

$$\mathbb{P}(Z_1 = 0, Z_2 = 0) = \mathbb{P}(X_3 = 0) = \frac{1}{2}$$

But also

$$\mathbb{P}(Z_1 = 0) = \mathbb{P}(Z_2 = 0) = \mathbb{P}(X_3 = 0) = \frac{1}{2} \implies \mathbb{P}(Z_1 = 0)\cdot\mathbb{P}(Z_2 = 0) = 0$$

So the events are not independent.

7 Inequalities for random variables

7.1 Markov's inequality

The following useful inequality, and the others derived from it, hold in the discrete and the continuous case.

Theorem. Let $X \geq 0$ be a non-negative random variable. Then for all $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof. Observe that $X \geq a \cdot 1(X \geq a)$. This can be seen to be true simply by checking both cases, $X < a$ and $X \geq a$. Taking expectations, we get

$$\mathbb{E}[X] \geq \mathbb{E}[a \cdot 1(X \geq a)] = \mathbb{E}[a \cdot \mathbb{P}(X \geq a)] = a \cdot \mathbb{P}(X \geq a)$$

and the result follows. □

7.2 Chebyshev's inequality

Theorem. Let X be a random variable with finite expectation. Then for all $a > 0$,

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq a\right) \leq \frac{\operatorname{Var}(X)}{a^2}$$

Proof. Note that $\mathbb{P}\left(|X - \mathbb{E}[X]| \geq a\right) = \mathbb{P}\left(|X - \mathbb{E}[X]|^2 \geq a^2\right)$. Then we can apply Markov's inequality to this non-negative random variable to get

$$\mathbb{P}\left(|X - \mathbb{E}[X]|^2 \geq a^2\right) \leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{a^2} = \frac{\operatorname{Var}(X)}{a^2}$$

□

7.3 Cauchy–Schwarz inequality

Theorem. If X and Y are random variables, then

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}$$

Proof. It suffices to prove this statement for X and Y which have finite second moments, i.e. $\mathbb{E}\left[X^2\right]$ and $\mathbb{E}\left[X^2\right]$ are finite. Clearly if they are infinite, then the upper bound is infinite which is trivially true. We need to show that $|\mathbb{E}[XY]|$ is finite. Here we can apply the additional assumption that X and Y are non-negative, since we are taking the absolute value:

$$XY \leq \frac{1}{2}\left(X^2 + Y^2\right) \implies \mathbb{E}[XY] \leq \frac{1}{2}\left(\mathbb{E}\left[X^2\right] + \mathbb{E}\left[Y^2\right]\right)$$

Now, we can assume $\mathbb{E}\left[X^2\right] > 0$ and $\mathbb{E}\left[Y^2\right] > 0$. If this were not the case, the result is trivial since if at least one of them were equal to zero, the corresponding random variable would be identically zero. Let $t \in \mathbb{R}$ and consider

$$0 \leq (X - tY)^2 = X^2 - 2tXY + t^2Y^2$$

Hence

$$\mathbb{E}\left[X^2\right] - 2t\mathbb{E}[XY] + t^2\mathbb{E}\left[Y^2\right] \geq 0$$

We can view this left hand side as a function $f(t)$. The minimum value of this function is achieved at $t_* = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$. Then

$$f(t_*) \geq 0 \implies \mathbb{E}\left[X^2\right] - \frac{2\mathbb{E}[XY]}{\mathbb{E}[Y^2]} + \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} \geq 0$$

Hence,

$$\mathbb{E}[XY]^2 \leq \mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]$$

and the result follows. □

Note that we also have

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}$$

This is because we can redefine $X \mapsto |X|$ and $Y \mapsto |Y|$, giving

$$|\mathbb{E}[|X| \cdot |Y|]| \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}$$
$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}$$

7.4 Equality in Cauchy–Schwarz

In what cases do we get equality in the Cauchy–Schwarz inequality? Recall that the inequality states

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}$$

Recall that in the proof, we considered the random variable $(X - tY)^2$ where X and Y were non-negative, and had finite second moments. The expectation of this random variable was called $f(t)$, and we found that $f(t)$ was minimised when $t = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$. We have equality exactly when $f(t) = 0$ for this value of t. But $(X - tY)^2$ is a non-negative random variable, with expectation zero, so it must be zero with probability 1. So we have equality if and only if X is exactly tY.

7.5 Jensen's inequality

Definition. A function $f : \mathbb{R} \to \mathbb{R}$ is called convex if $\forall x, y \in \mathbb{R}$ and for all $t \in [0, 1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

This can be visualised as linearly interpolating the values of the function at two points, x and y. The linear interpolation of those points is always greater than the function applied to the linear interpolation of the input points.

Theorem. Let X be a random variable, and let f be a convex function. Then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

We can remember the direction of this inequality by considering the variance: $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ which is non-negative. Further, $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$ hence $\mathbb{E}\left[X^2\right] \geq \mathbb{E}[X]^2$. Squaring is an example of a convex function, so Jensen's inequality holds in this case. We will first prove a basic lemma about convex functions.

Lemma. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Then f is the supremum of all the lines lying below it. More formally, $\forall m \in \mathbb{R}$, $\exists a, b \in \mathbb{R}$ such that $f(m) = am + b$ and $f(x) \geq ax + b$ for all x.

Proof. Let $m \in \mathbb{R}$. Let $x < m < y$. Then we can express m as $tx + (1-t)y$ for some t in the interval $[0, 1]$. By convexity,

$$f(m) \leq tf(x) + (1-t)f(y)$$

And hence,

$$\begin{aligned} tf(m) + (1-t)f(m) &\leq tf(x) + (1-t)f(y) \\ t(f(m) - f(x)) &\leq (1-t)(f(y) - f(m)) \\ \frac{f(m) - f(x)}{m - x} &\leq \frac{f(y) - f(m)}{y - m} \end{aligned}$$

So the slope of the line joining m to a point on its left is smaller than the slope of the line joining m to a point on its right. So we can produce a value $a \in \mathbb{R}$ given by

$$a = \sup_{x<m} \frac{f(m) - f(x)}{m - x}$$

such that

$$\frac{f(m) - f(x)}{m - x} \leq a \leq \frac{f(y) - f(m)}{y - m}$$

for all $x < m < y$. We can rearrange this to give

$$f(x) \geq a(x - m) + f(m) = ax + (f(m) - am)$$

for all x. □

We may now prove Jensen's inequality.

Proof. Set $m = \mathbb{E}[X]$. Then from the lemma above, there exists $a, b \in \mathbb{R}$ such that

$$f(m) = am + b \implies f(\mathbb{E}[X]) = a\mathbb{E}[X] + b \tag{$*$}$$

and for all x, we have

$$f(x) \geq ax + b$$

We can now apply this inequality to X to get

$$f(X) \geq aX + b$$

Taking the expectation, by $(*)$ we get

$$\mathbb{E}[f(X)] \geq a\mathbb{E}[X] + b = f(\mathbb{E}[X])$$

as required. □

Like the Cauchy–Schwarz inequality, we would like to consider the cases of equality. Let X be a random variable, and f be a convex function such that if $m = \mathbb{E}[X]$, then $\exists a, b \in \mathbb{R}$ such that

$$f(m) = am + b; \quad \forall x \neq m,\ f(x) > ax + b$$

We know that $f(X) \geq aX + b$, since f is convex. Then $f(X) - (aX + b) \geq 0$ is a non-negative random variable. Taking expectations,

$$\mathbb{E}[f(X) - (aX + b)] \geq 0$$

But $\mathbb{E}[aX + b] = am + b = f(m) = f(\mathbb{E}[X])$. We assumed that $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$, hence $\mathbb{E}[aX + b] = \mathbb{E}[f(X)]$ and $\mathbb{E}[f(X) - (aX + b)] = 0$. But since $f(X) \geq aX + b$, this forces $f(X) = aX + b$ everywhere. By our assumption, for all $x \neq m$, $f(x) > ax + b$. This forces $X = m$ with probability 1.

7.6 Arithmetic mean and geometric mean inequality

Let f be a convex function. Suppose $x_1, \dots, x_n \in \mathbb{R}$. Then, from Jensen's inequality,

$$\frac{1}{n}\sum_{k=1}^{n} f(x_k) \geq f\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)$$

Indeed, we can define a random variable X to take values $x_1, \dots, x_n$ all with equal probability. Then, $\mathbb{E}[f(X)]$ gives the left hand side, and $f(\mathbb{E}[X])$ gives the right hand side. Now, let $f(x) = -\log x$. This is a convex function as required. Hence

$$-\frac{1}{n}\sum_{k=1}^{n} \log x_k \geq -\log\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)$$

$$\left(\prod_{k=1}^{n} x_k\right)^{\frac{1}{n}} \leq \frac{1}{n}\sum_{k=1}^{n} x_k$$

Hence the geometric mean is less than or equal to the arithmetic mean.

8 Combinations of random variables

8.1 Conditional expectation and law of total expectation

Recall that if $B \in \mathcal{F}$ with $\mathbb{P}(B) \geq 0$, we defined

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

Now, let X be a random variable, and let B be an event as above with nonzero probability. We can then define

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X \cdot 1(B)]}{\mathbb{P}(B)}$$

The numerator is notably zero when $1(B) = 0$, so in essence we are excluding the case where X is not B.

Theorem (law of total expectation)**.** Suppose $X \geq 0$. Let (Ω_n) be a partition of Ω into disjoint events, so $\Omega = \bigcup_n \Omega_n$. Then

$$\mathbb{E}[X] = \sum_n \mathbb{E}[X \mid \Omega_n] \cdot \mathbb{P}(\Omega_n)$$

Proof. We can write $X = X \cdot 1(\Omega)$, where

$$1(\Omega) = \sum_n 1(\Omega_n)$$

Taking expectations, we get

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_n X \cdot 1(\Omega_n)\right]$$

By countable additivity of expectation, we have

$$\mathbb{E}[X] = \sum_n \mathbb{E}[X \cdot 1(\Omega_n)] = \sum_n \mathbb{E}[X \mid \Omega_n] \cdot \mathbb{P}(\Omega_n)$$

as required. □

8.2 Joint distribution

Definition. Let $X_1, \dots, X_n$ be discrete random variables. Their joint distribution is defined as

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

for all $x_i \in \Omega_i$.

Now, we have

$$\mathbb{P}(X_1 = x_1) = \mathbb{P}\left(\{X_1 = x_1\} \cap \bigcup_{i=2}^{n} \bigcup_{x_i} \{X_i = x_i\}\right) = \sum_{x_2,\dots,x_n} \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

In general,

$$\mathbb{P}(X_i = x_i) = \sum_{x_1,x_2,\dots,x_{i-1},x_{i+1},\dots,x_n} \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

We call $(\mathbb{P}(X_i = x_i))_i$ the marginal distribution of X_i. Let X, Y be random variables. The conditional distribution of X given $Y = y$ where $y \in \Omega_y$ is defined to be

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

We can find

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y \mathbb{P}(X = x \mid Y = y)\,\mathbb{P}(Y = y)$$

which is the law of total probability.

8.3 Convolution

Let X and Y be independent, discrete random variables. We would like to find $\mathbb{P}(X + Y = z)$. Clearly this is

$$\begin{aligned}
\mathbb{P}(X + Y = z) &= \sum_y \mathbb{P}(X + Y = z, Y = y) \\
&= \sum_y \mathbb{P}(X = z - y, Y = y) \\
&= \sum_y \mathbb{P}(X = z - y) \cdot \mathbb{P}(Y = y)
\end{aligned}$$

This last sum is called the convolution of the distributions of X and Y. Similarly,

$$\mathbb{P}(X+Y=z) = \sum_x \mathbb{P}(X=x)\cdot\mathbb{P}(Y=z-x)$$

As an example, let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ be independent. Then

$$\begin{aligned}
\mathbb{P}(X+Y=n) &= \sum_{r=0}^{n} \mathbb{P}(X=r)\,\mathbb{P}(Y=n-r) \\
&= \sum_{r=0}^{n} e^{-\lambda}\frac{\lambda^r}{r!}\cdot e^{-\mu}\frac{\mu^{n-r}}{(n-r)!} \\
&= e^{-(\lambda+\mu)}\sum_{r=0}^{n}\frac{\lambda^r\mu^{n-r}}{r!(n-r)!} \\
&= \frac{e^{-(\lambda+\mu)}}{n!}\sum_{r=0}^{n}\frac{\lambda^r\mu^{n-r}\cdot n!}{r!(n-r)!} \\
&= \frac{e^{-(\lambda+\mu)}}{n!}\sum_{r=0}^{n}\binom{n}{r}\lambda^r\mu^{n-r} \\
&= \frac{e^{-(\lambda+\mu)}}{n!}(\lambda+\mu)^n
\end{aligned}$$

which is the probability mass function of a Poisson random variable with parameter $\lambda+\mu$. In other words, $X+Y \sim \text{Poi}(\lambda+\mu)$.

8.4 Conditional expectation

Let X and Y be discrete random variables. Then the conditional expectation of X given that $Y=y$ is

$$\begin{aligned}
\mathbb{E}[X \mid Y=y] &= \frac{\mathbb{E}[X\cdot 1(Y=y)]}{\mathbb{P}(Y=y)} \\
&= \frac{1}{\mathbb{P}(Y=y)}\sum_x x\cdot\mathbb{P}(X=x, Y=y) \\
&= \sum_x x\cdot\mathbb{P}(X=x\mid Y=y)
\end{aligned}$$

Observe that for every $y\in\Omega_y$, this expectation is purely a function of y. Let $g(y)=\mathbb{E}[X\mid Y=y]$. Now, we define the conditional expectation of X given Y as $\mathbb{E}[X\mid Y]=g(Y)$. Note that $\mathbb{E}[X\mid Y]$ is a

random variable, dependent only on Y. We have

$$\begin{aligned}\mathbb{E}[X \mid Y] &= g(Y) \cdot 1 \\ &= g(Y) \sum_y 1(Y = y) \\ &= \sum_y g(Y) \cdot 1(Y = y) \\ &= \sum_y g(y) \cdot 1(Y = y) \\ &= \sum_y \mathbb{E}[X \mid Y = y] \cdot 1(Y = y)\end{aligned}$$

This is perhaps a clearer way to see that it depends only on Y. As an example, let us consider tossing a p-biased coin n times independently. We write X_i for the indicator function that the ith toss was a head. Let $Y_n = X_1 + \cdots + X_n$. What is $\mathbb{E}[X_1 \mid Y_n]$? Let $g(y) = \mathbb{E}[X_1 \mid Y_n = y]$. Then $\mathbb{E}[X_1 \mid Y_n] = g(Y_n)$. We therefore need to find g. Let $y \in \{0, \dots, n\}$, then

$$\begin{aligned}g(y) &= \mathbb{E}[X_1 \mid Y_n = y] \\ &= \mathbb{P}(X_1 = 1 \mid Y_n = y)\end{aligned}$$

Clearly if $y = 0$, then $\mathbb{P}(X_1 = 1 \mid Y_n = 0) = 0$. Now, suppose $y \neq 0$. We have

$$\begin{aligned}\mathbb{P}(X_1 = 1 \mid Y_n = y) &= \frac{\mathbb{P}(X_1 = 1, Y_n = y)}{\mathbb{P}(Y_n = y)} \\ &= \frac{\mathbb{P}(X_1 = 1, X_2 + \cdots + X_n = y - 1)}{\mathbb{P}(Y_n = y)} \\ &= \frac{\mathbb{P}(X_1 = 1) \cdot \mathbb{P}(X_2 + \cdots + X_n = y - 1)}{\mathbb{P}(Y_n = y)} \\ &= \frac{p \cdot \binom{n-1}{y-1} \cdot p^{y-1}(1-p)^{n-y}}{\mathbb{P}(Y_n = y)} \\ &= \frac{\binom{n-1}{y-1} \cdot p^y (1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} \\ &= \frac{\binom{n-1}{y-1}}{\binom{n}{y}} \\ &= \frac{y}{n}\end{aligned}$$

Hence

$$g(y) = \frac{y}{n}$$

We can then find that

$$\mathbb{E}[X_1 \mid Y_n] = g(Y_n) = \frac{Y_n}{n}$$

which is indeed a random variable dependent only on Y_n.

8.5 Properties of conditional expectation

The following properties hold.

- For all $c \in \mathbb{R}$, $\mathbb{E}[cX \mid Y] = c\mathbb{E}[X \mid Y]$, and $\mathbb{E}[c \mid Y] = c$.
- Let $X_1, \dots, X_n$ be random variables. Then $\mathbb{E}\left[\sum_{i=1}^n X_i \mid Y\right] = \sum_{i=1}^n \mathbb{E}[X_i \mid Y]$.
- $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$.

The last property is not obvious from the definition, so it warrants its own proof. We can see by the standard properties of the expectation that

$$\begin{aligned}
\mathbb{E}[X \mid Y] &= \sum_y 1(Y = y)\mathbb{E}[X \mid Y = y] \\
\therefore \mathbb{E}[\mathbb{E}[X \mid Y]] &= \sum_y \mathbb{E}[1(Y = y)]\,\mathbb{E}[X \mid Y = y] \\
&= \sum_y \mathbb{P}(Y = y)\,\mathbb{E}[X \mid Y = y] \\
&= \sum_y \mathbb{P}(Y = y)\,\frac{\mathbb{E}[X \cdot 1(Y = y)]}{\mathbb{P}(Y = y)} \\
&= \sum_y \mathbb{E}[X \cdot 1(Y = y)] \\
&= \mathbb{E}\left[\sum_y X \cdot 1(Y = y)\right] \\
&= \mathbb{E}\left[X \sum_y 1(Y = y)\right] \\
&= \mathbb{E}[X]
\end{aligned}$$

Alternatively, we could expand the inner expectation as a sum:

$$\sum_y \mathbb{E}[X \mid Y = y] \cdot \mathbb{P}(Y = y) = \sum_x \sum_y x \cdot \mathbb{P}(X = x \mid Y = y) \cdot \mathbb{P}(Y = y)$$

and the result follows as required. The final property relates conditional probability to independence. Let X and Y be independent. Then $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$. Indeed,

$$\begin{aligned}
\mathbb{E}[X \mid Y] &= \sum_y 1(Y = y)\mathbb{E}[X \mid Y = y] \\
&= \sum_y 1(Y = y)\mathbb{E}[X] \\
&= \mathbb{E}[X]
\end{aligned}$$

Proposition. Suppose Y and Z are independent random variables. Then

$$\mathbb{E}[\mathbb{E}[X \mid Y] \mid Z] = \mathbb{E}[X]$$

Proof. Let $\mathbb{E}[X \mid Y] = g(Y)$ be a random variable that is a function only of Y. Since Y and Z are independent, $f(Y)$ is also independent of Z for any function f. Then $\mathbb{E}[g(Y) \mid Z] = \mathbb{E}[g(Y)] = \mathbb{E}[X]$. □

Proposition. Suppose $h : \mathbb{R} \to \mathbb{R}$ is a function. Then

$$\mathbb{E}[h(Y) \cdot X \mid Y] = h(Y) \cdot \mathbb{E}[X \mid Y]$$

We can 'take out what is known', since we know what Y is.

Proof. Note that

$$\mathbb{E}[h(Y) \cdot X \mid Y = y] = \mathbb{E}[h(y) \cdot X \mid Y = y] = h(y) \cdot \mathbb{E}[X \mid Y = y]$$

Then

$$\mathbb{E}[h(Y) \cdot X \mid Y] = h(y) \cdot \mathbb{E}[X \mid Y]$$

as required. □

Corollary. $\mathbb{E}[\mathbb{E}[X \mid Y] \mid Y] = \mathbb{E}[X \mid Y]$, and $\mathbb{E}[X \mid X] = X$.

Let $X_i = 1(i\text{th toss is a head})$, and $Y_n = X_1 + \cdots + X_n$. We found before that $\mathbb{E}[X_1 \mid Y_n] = \frac{Y_n}{n}$. By symmetry, for all i we have $\mathbb{E}[X_i \mid Y_n] = \mathbb{E}[X_1 \mid Y_n]$. Hence

$$\mathbb{E}[Y_n \mid Y_n] = \mathbb{E}\left[\sum_{i=1}^{n} X_i \mid Y_n\right] = \sum_{i=1}^{n} \mathbb{E}[X_i \mid Y_n] = n \cdot \mathbb{E}[X_1 \mid Y_n]$$

which yields the same result.

9 Random walks

9.1 Definition

A random process, also known as a stochastic process, is a sequence of random variables X_n for $n \in \mathbb{N}$. A random walk is a random process that can be expressed as

$$X_n = x + Y_1 + \cdots + Y_n$$

where the Y_i are independent and identically distributed, and x is a deterministic number. We will focus on the simple random walk on $\mathbb{Z}$, which is defined by taking

$$\mathbb{P}(Y_i = 1) = p; \quad \mathbb{P}(Y_i = -1) = 1 - p = q$$

This can be thought of as a specific case of a Markov chain; it has the property that the path to X_i does not matter, all that matters is the value that we are at, at any point in time.

9.2 Gambler's ruin estimate

What is the probability that X_n reaches some value a before it falls to 0? We will write $\mathbb{P}_x$ for the probability measure $\mathbb{P}$ with the condition that $X_0 = x$, i.e.

$$\mathbb{P}_x(A) = \mathbb{P}(A \mid X_0 = x)$$

We define $h(x) = \mathbb{P}_x((X_n) \text{ hits } a \text{ before hitting } 0)$. We can define a recurrence relation. By the law of total probability, we have, for $0 < x < a$,

$$\begin{aligned} h(x) &= \mathbb{P}_x((X_n) \text{ hits } a \text{ before hitting } 0 \mid Y_1 = 1) \cdot \mathbb{P}_x(Y_1 = 1) \\ &\quad + \mathbb{P}_x((X_n) \text{ hits } a \text{ before hitting } 0 \mid Y_1 = -1) \cdot \mathbb{P}_x(Y_1 = -1) \\ &= p \cdot h(x+1) + q \cdot h(x-1) \end{aligned}$$

Note that

$$h(0) = 0; \quad h(a) = 1$$

There are two cases; $p = q = \frac{1}{2}$ and $p \neq q$. If $p = q = \frac{1}{2}$, then

$$h(x) - h(x+1) = h(x-1) - h(x)$$

We can then solve this to find

$$h(x) = \frac{x}{a}$$

If $p \neq q$, then

$$h(x) = ph(x+1) + qh(x-1)$$

We can try a solution of the form λ^x. Substituting gives

$$p\lambda^2 - \lambda + q = 0 \implies \lambda = 1, \frac{q}{p}$$

The general solution can be found by using the boundary conditions.

$$h(x) = A + B\left(\frac{q}{p}\right)^x \implies h(x) = \frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^a - 1}$$

This is known as the 'gambler's ruin' estimate, since it determines whether a gambler will reach a target before going bankrupt.

9.3 Expected time to absorption

Let us define T to be the first time that $x = 0$ or $x = a$. Then $T = \min\{n \geq 0 : X_n \in \{0, a\}\}$. We want to find $\mathbb{E}_x[T] = \tau_x$. We can apply a condition on the first step, and use the law of total expectation to give

$$\tau_x = p\mathbb{E}_x[T \mid Y_1 = 1] + q\mathbb{E}_x[T \mid Y_1 = -1]$$

Hence

$$\tau_x = p(\tau_{x+1} + 1) + q(\tau_{x-1} + 1)$$

We can deduce that, for $0 < x < a$,

$$\tau_x = 1 + p\tau_{x+1} + q\tau_{x-1}$$

and $\tau_0 = \tau_a = 0$. If $p = q = \frac{1}{2}$, then we can try a solution of the form Ax^2.

$$Ax^2 = 1 + \frac{1}{2}A(x+1)^2 + \frac{1}{2}A(x-1)^2$$

This gives a general solution of the form

$$A = -1 \implies \tau_x = -x^2 + Bx + C \implies \tau_x = x(a-x)$$

If $p \neq q$, then we will try a solution of the form Cx, giving

$$C = \frac{1}{q-p}$$

The general solution has the form

$$\tau_x = \frac{x}{q-p} + A + B\left(\frac{q}{p}\right)^x \implies \tau_x = \frac{x}{q-p} - \frac{q}{q-p} \cdot \frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^a - 1}$$

10 Probability generating functions

10.1 Definition

Let X be a random variable with values in the positive integers, $\mathbb{N}$. Let $p_r = \mathbb{P}(X = r)$ be the probability mass function. Then the probability generating function is defined to be

$$p(z) = \sum_{r=0}^{\infty} p_r z^r = \mathbb{E}\left[z^X\right] \text{ for } |z| \leq 1$$

When $|z| \leq 1$, the probability generating function converges absolutely, since $\left|\sum_{r=0}^{\infty} p_r z^r\right| \leq \sum_{r=0}^{\infty} p_r = 1$. So $p(z)$ is well-defined and has a radius of convergence of at least 1.

Theorem. The probability generating function of X uniquely determines the distribution of X.

Proof. Suppose (p_r) and (q_r) are two probability mass functions with

$$\sum_{r=0}^{\infty} p_r z^r = \sum_{r=0}^{\infty} q_r z^r, \forall |z| \leq 1$$

We will show that $p_r = q_r$ for all r. First, set $z = 0$, then clearly $p_0 = q_0$. Then by induction, suppose that $p_r = q_r$ for all $r \leq n$. Then we would like to show that $p_{n+1} = q_{n+1}$. We know that

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

Hence, dividing by z^{n+1}, and taking the limit as $z \to 0$, we have $p_{n+1} = q_{n+1}$ as required. □

10.2 Finding moments and probabilities

Theorem.

$$\lim_{z \to 1^-} p'(z) = p'(1^-) = \mathbb{E}[X]$$

Proof. We will first assume that $\mathbb{E}[X]$ is finite; we will then extend the proof to the infinite case. Let $0 < z < 1$, then since the series $p(z)$ is absolutely convergent, we can interchange the sum and the derivative operators, giving

$$p'(z) = \sum_{r=0}^{\infty} r p_r z^{r-1}$$

We can make an upper bound for this sum:

$$\sum_{r=0}^{\infty} r p_r z^{r-1} \leq \sum_{r=0}^{\infty} r p_r = \mathbb{E}[X]$$

Since $0 < z < 1$, we see that $p'(z)$ is an increasing function of z. This implies that there exists a limit of $p'(z)$ as $z \to 1^-$, which is upper bounded by $\mathbb{E}[X]$. Now, let $\varepsilon > 0$ and let N be an integer large enough such that

$$\sum_{r=0}^{N} r p_r \geq \mathbb{E}[X] - \varepsilon$$

We have further that, since $0 < z < 1$,

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

So

$$\lim_{z \to 1^-} p'(z) \geq \sum_{r=1}^{N} r p_r \geq \mathbb{E}[X] - \varepsilon$$

which is true for any ε. Therefore $\lim_{z \to 1^-} p'(z) = \mathbb{E}[X]$. Now, in the case that $\mathbb{E}[X]$ is infinite, for any M we can find a sufficiently large N such that

$$\sum_{r=0}^{N} r p_r \geq M$$

From above, we know that

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

Since this is true for any M, this limit is equal to ∞. □

In exactly the same way, we can prove that

$$p''(1^-) = \mathbb{E}[X(X-1)]$$

and in general,

$$p^{(k)}(1^-) = \mathbb{E}[X(X-1)\cdots(X-k+1)]$$

In particular,

$$\operatorname{Var}(X) = p''(1^-) + p'(1^-) - p'(1^-)^2$$

Further,

$$\mathbb{P}(X = n) = \frac{1}{n!} \left.\frac{\mathrm{d}^n}{\mathrm{d}z^n} p(z)\right|_{z=0}$$

10.3 Sums of random variables

Suppose that $X_1, \dots, X_n$ are independent random variables with probability generating functions $q_1, \dots, q_n$ respectively. Then

$$p(z) = \mathbb{E}\left[z^{X_1 + \cdots + X_n}\right]$$

Recall that if X and Y are independent, then for all functions f and g, we have $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)]$. Therefore,

$$p(z) = \mathbb{E}\left[z^{X_1} z^{X_2} \cdots z^{X_n}\right] = \mathbb{E}\left[z^{X_1}\right] \cdots \mathbb{E}\left[z^{X_n}\right] = q_1(z) \cdots q_n(z)$$

So the probability generating function factorises into its independent parts. In particular, if all the X_i are independent and identically distributed, then

$$p(z) = q(z)^n$$

10.4 Common probability generating functions

Suppose that $X \sim \operatorname{Bin}(n, p)$. Then

$$\begin{aligned} p(z) &= \mathbb{E}\left[z^X\right] \\ &= \sum_{r=0}^{n} z^r \binom{n}{r} p^r (1-p)^{n-r} \\ &= \sum_{r=0}^{n} \binom{n}{r} (pz)^r (1-p)^{n-r} \\ &= (pz + 1 - p)^n \end{aligned}$$

Now, let $X \sim \operatorname{Bin}(n, p)$, $Y \sim \operatorname{Bin}(m, p)$ be independent random variables. Then the probability generating function of $X + Y$ is

$$(pz + 1 - p)^n \cdot (pz + 1 - p)^m = (pz + 1 - p)^{n+m}$$

which is the probability generating function of a binomial distribution where the number of trials is $n + m$. Now, suppose that $X \sim \operatorname{Geo}(p)$. Then

$$\begin{aligned} p(z) &= \mathbb{E}\left[z^X\right] \\ &= \sum_{r=0}^{n} z^r (1-p)^r p \\ &= \frac{p}{1 - z(1-p)} \end{aligned}$$

Now, suppose that $X \sim \text{Poi}(\lambda)$. Then

$$\begin{aligned} p(z) &= \mathbb{E}\left[z^X\right] \\ &= \sum_{r=0}^{n} z^r e^{-\lambda} \frac{\lambda^r}{r!} \\ &= e^{\lambda(z-1)} \end{aligned}$$

10.5 Random sums of random variables

Consider the sum of a random number of random variables. Let $X_1, \ldots$ be independent and identically distributed, and let N be an independent random variable with values in $\mathbb{N}$. Now, we can define the random variables S_n to be

$$S_n = X_1 + \cdots + X_n$$

Then

$$S_N = X_1 + \cdots + X_N$$

is a random variable dependent on N. For all $\omega \in \Omega$,

$$\begin{aligned} S_N(\omega) &= X_1(\omega) + \cdots + X_{N(\omega)}(\omega) \\ &= \sum_{i=1}^{N(\omega)} X_i(\omega) \end{aligned}$$

Now, let q be the probability generating function of N, and p be the probability generating function of X_1 (or equivalently, any X_i). Then let

$$\begin{aligned} r(z) &= \mathbb{E}\left[z^{S_N}\right] \\ &= \sum_n \mathbb{E}\left[z^{X_1+\cdots+X_N} \cdot 1(N=n)\right] \\ &= \sum_n \mathbb{E}\left[z^{X_1+\cdots+X_n} \cdot 1(N=n)\right] \\ &= \sum_n \mathbb{E}\left[z^{X_1+\cdots+X_n}\right] \mathbb{E}\left[1(N=n)\right] \\ &= \sum_n \mathbb{E}\left[z^{X_1+\cdots+X_n}\right] \mathbb{P}\left(N=n\right) \\ &= \sum_n \mathbb{E}\left[z^{X_1}\right]^n \mathbb{P}\left(N=n\right) \\ &= \sum_n p(z)^n \mathbb{P}\left(N=n\right) \\ &= q(p(z)) \end{aligned}$$

Here is an alternative proof using conditional expectation.

$$\begin{aligned} r(z) &= \mathbb{E}\left[z^{S_N}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[z^{S_N} \mid N\right]\right] \end{aligned}$$

We can see that

$$\begin{aligned}\mathbb{E}\left[z^{S_N} \mid N = n\right] &= \mathbb{E}\left[z^{S_n} \mid N = n\right] \\ &= \mathbb{E}\left[z^{X_1}\right]^n \\ &= p(z)^n\end{aligned}$$

Therefore,

$$\begin{aligned}r(z) &= \mathbb{E}\left[p(z)^N\right] \\ &= q(p(z))\end{aligned}$$

Using this expression for r, we can find that

$$\mathbb{E}[S_N] = r'(1^-) = q'(p(1^-)) \cdot p'(1^-) = q'(1^-) \cdot p'(1^-) = \mathbb{E}[N]\,\mathbb{E}[X_1]$$

Similarly,

$$\operatorname{Var}(S_N) = \mathbb{E}[N] \operatorname{Var}(X_1) + \operatorname{Var}(N) (\mathbb{E}[X_1])^2$$

11 Branching processes

11.1 Introduction

Let $(X_n : n \geq 0)$ be a random process, where X_n is the number of individuals in generation n, and $X_0 = 1$. The individual in generation 0 produces a random number of offspring with distribution

$$g_k = \mathbb{P}(X_1 = k)$$

Then every individual in generation 1 produces an independent number of offspring with the same distribution. This is called a branching process. We can write a recursive formula for X_n. First, let $(Y_{k,n} : k \geq 1, n \geq 0)$ be an independent and identically distributed sequence with distribution $(g_k)_k$. So $Y_{k,n}$ is the number of offspring of the kth individual in generation n.

$$X_{n+1} = \begin{cases} Y_{1,n} + \cdots + Y_{X_n,n} & \text{when } X_n \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

11.2 Expectation of generation size

Theorem.

$$\mathbb{E}[X_n] = \mathbb{E}[X_1]^n$$

Proof. Inductively,

$$\begin{aligned}
\mathbb{E}[X_{n+1}] &= \mathbb{E}[\mathbb{E}[X_{n+1} \mid X_n]] \\
\mathbb{E}[X_{n+1} \mid X_n = m] &= \mathbb{E}\left[Y_{1,n} + \cdots + Y_{X_n,n} \mid X_n = m\right] \\
&= \mathbb{E}\left[Y_{1,n} + \cdots + Y_{m,n} \mid X_n = m\right] \\
&= m\mathbb{E}\left[Y_{1,n}\right] \\
&= m\mathbb{E}[X_1] \\
\therefore \mathbb{E}[X_{n+1} \mid X_n] &= X_n \cdot \mathbb{E}[X_1] \\
\therefore \mathbb{E}[X_{n+1}] &= \mathbb{E}[X_n \cdot \mathbb{E}[X_1]] \\
&= \mathbb{E}[X_n] \cdot \mathbb{E}[X_1]
\end{aligned}$$

□

11.3 Probability generating functions

Theorem. Let $G(z) = \mathbb{E}\left[z^{X_1}\right]$ be the probability generating function of X_1, and $G_n(z) = \mathbb{E}\left[z^{X_n}\right]$ be the probability generating function of X_n. Then

$$G_{n+1}(z) = G(G_n(z)) = G(G(\cdots G(z) \cdots)) = G_n(G(z))$$

Proof.

$$\begin{aligned}
G_{n+1}(z) &= \mathbb{E}\left[z^{X_{n+1}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[z^{X_{n+1}} \mid X_n\right]\right] \\
\mathbb{E}\left[z^{X_{n+1}} \mid X_n = m\right] &= \mathbb{E}\left[z^{Y_{1,n}+\cdots+Y_{m,n}} \mid X_n = m\right] \\
&= \mathbb{E}\left[z^{X_1}\right]^m \\
&= G(z)^m \\
\therefore \mathbb{E}\left[\mathbb{E}\left[z^{X_{n+1}} \mid X_n\right]\right] &= \mathbb{E}\left[G(z)^{X_n}\right] \\
&= G_n(G(z))
\end{aligned}$$

□

11.4 Probability of extinction

We define the extinction probability q as the probability that $X_n = 0$ for some $n \geq 1$, and $q_n = \mathbb{P}(X_n = 0)$. It is clear that $X_n = 0$ implies that $X_{n+1} = 0$. So the sequence of events $(A_n) = (\{X_n = 0\})$ is an increasing sequence of events. So by the continuity of the probability measure, $\mathbb{P}(A_n)$ converges to $\mathbb{P}\left(\bigcup A_n\right)$ as $n \to \infty$. Note that the event $\bigcup A_n$ is the event that there will be extinction. Therefore, $q_n \to q$ as $n \to \infty$.

Claim. $q_{n+1} = G(q_n)$ and $q = G(q)$.

Proof. Using the above theorem on q,

$$\begin{aligned} q_{n+1} &= \mathbb{P}(X_{n+1} = 0) \\ &= G_{n+1}(0) \\ &= G(G_n(0)) \\ &= G(q_n) \end{aligned}$$

Since G is continuous, taking the limit as $n \to \infty$ and using that $q_n \to q$ gives $G(q) = q$. □

We can form another proof for the first part of the above claim.

Proof. Instead of conditioning on the previous generation, let us condition on the first generation, i.e. $X_1 = m$. Note that after the first generation, we will have m independent subtrees on the family tree. Each tree is identically distributed to the entire tree as a whole. Hence,

$$X_{n+1} = X_n^{(1)} + \cdots + X_n^{(m)}$$

where the $X_i^{(j)}$ are independent and identically distributed random processes each with the same offspring distribution. By the law of total probability,

$$\begin{aligned} q_{n+1} &= \mathbb{P}(X_{n+1} = 0) \\ &= \sum_m \mathbb{P}(X_{n+1} = 0 \mid X_1 = m) \cdot \mathbb{P}(X_1 = m) \\ &= \sum_m \mathbb{P}\left(X_n^{(1)} = 0, \dots, X_n^{(m)} = 0\right) \cdot \mathbb{P}(X_1 = m) \\ &= \sum_m \mathbb{P}\left(X_n^{(1)} = 0\right)^m \cdot \mathbb{P}(X_1 = m) \\ &= \sum_m q_n^m \cdot \mathbb{P}(X_1 = m) \\ &= G(q_n) \end{aligned}$$

□

Theorem. The extinction probability q is the minimal non-negative solution to $G(t) = t$. Further, supposing that $\mathbb{P}(X_1 = 1) < 1$, we have that $q < 1$ if and only if $\mathbb{E}[X_1] > 1$.

Proof. First, we will prove the minimality of q. Let t be the smallest non-negative solution to $G(t) = t$. We will prove inductively that $q_n \leq t$ for all n, and then by taking limits we have that $q \leq t$. Since q is a solution, this will imply that $q = t$. Now, as a base case, $q_0 = 0 = \mathbb{P}(X_0 = 0) \leq t$. Inductively let us suppose that $q_n \leq t$. We know that $q_{n+1} = G(q_n)$. G is an increasing function on $[0, 1]$, and since $q_n \leq t$ we have $q_{n+1} = G(q_n) \leq G(t) = t$.

Now, we can take $\mathbb{P}(X_1 = 1) < 1$. Let us use the notation $g_r = \mathbb{P}(X_1 = r)$ for simplicity. Consider the function $H(z) = G(z) - z$. Let us assume further that $g_0 + g_1 < 1$, since otherwise we cannot possibly ever increase the amount of individuals in future generations, as $\mathbb{E}[X_1] = \mathbb{P}(X_1 = 1) < 1$.

In this case, $G(z) = g_0 + g_1 z = 1 - \mathbb{E}[X_1] + \mathbb{E}[X_1] \cdot z$, and solving $G(z) = z$ we would get only $z = 1$ since $\mathbb{E}[X_1] < 1$. Now,

$$H''(z) = \sum_{r=2}^{\infty} r(r-1) g_r z^{r-2} > 0 \quad \forall z \in (0,1)$$

This implies that $H'(z)$ is a strictly increasing function in $(0,1)$. Hence, $H(z)$ has at most one root different from 1 in $(0,1)$, which follows from Rolle's theorem; indeed, if it had two roots different from 1, then H' would be zero once in (z_1, z_2) and once in $(z_2, 1)$, which contradicts the fact that H' is strictly increasing.

Let us first consider the case where H has no other root apart from 1. In this case, $H(1) = 0$ and $H(0) = g_0 \geq 0 \implies H(z) \geq 0$ for all $z \in [0,1]$. We can find that

$$H'(1^-) = \lim_{z \to 1^-} \frac{H(z) - H(1)}{z - 1} = \frac{H(z)}{z-1} < 0$$

since the numerator is positive, and the denominator is negative. We know that $H'(1^-) = G'(1^-) - 1$, and $H'(1^-) \leq 0 \implies G'(1^-) \leq 1$, and $G'(1^-) = \mathbb{E}[X_1]$. So when $q = 1$, then $\mathbb{E}[X_1] \leq 1$.

In the other case, H has one other root $r < 1$ as well as 1. We have that $H(r) = 0$ and $H(1) = 0$. By Rolle's theorem, there exists $z \in (r,1)$ such that $H'(z) = 0$. Further, $H'(x) = G'(x) - 1$ therefore $G'(z) = 1$. Now,

$$G'(x) = \sum_{r=1}^{\infty} r g_r x^{r-1} \implies H''(x) = G''(x) = \sum_{r=2}^{\infty} r(r-1) g_r x^{r-2}$$

Under the assumption that $g_0 + g_1 < 1$, we have that $G''(x) > 0$ for all $x \in (0,1)$, hence G' is strictly increasing for all $x \in (0,1)$. Therefore, $G'(1^-) > G'(z) = 1$ giving $\mathbb{E}[X_1] > 1$. So if $q < 1$, then $\mathbb{E}[X_1] > 1$. □

12 Continuous random variables

12.1 Probability distribution function

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then, as defined before, $X : \Omega \to \mathbb{R}$ is a random variable if

$$\forall x \in \mathbb{R}, \{X \leq x\} = \{\omega : X(\omega) \leq x\} \in \mathcal{F}$$

We define the probability distribution function $F : \mathbb{R} \to [0,1]$ as

$$F(x) = \mathbb{P}(X \leq x)$$

Theorem. The following properties hold.

(i) If $x \leq y$, then $F(x) \leq F(y)$.

(ii) For all $a < b$, $\mathbb{P}(a < x \leq b) = F(b) - F(a)$.

(iii) F is a right continuous function, and left limits always exist. In other words,

$$F(x^+) = \lim_{y \to x^+} F(y) = F(x); \quad F(x^-) = \lim_{y \to x^-} F(y) \leq F(x)$$

(iv) For all $x \in \mathbb{R}$, $F(x^-) = \mathbb{P}(X < x)$.

(v) We have $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$.

Proof. (i) The first statement is immediate from the definition of the probability measure.

(ii) We can deduce

$$\begin{aligned}\mathbb{P}(a < X \leq b) &= \mathbb{P}(\{a < X\} \cap \{X \leq b\}) \\ &= \mathbb{P}(X \leq b) - \mathbb{P}(\{X \leq b\} \cap \{X \leq a\}) \\ &= \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) \\ &= F(b) - F(a)\end{aligned}$$

(iii) For right continuity, we want to prove $\lim_{n\to\infty} F\left(x + \frac{1}{n}\right) = F(x)$. We will define $A_n = \left\{x < X \leq x + \frac{1}{n}\right\}$. Then the A_n are decreasing events, and the intersection of all A_n is the empty set $\varnothing$. Hence, by continuity of the probability measure, $\mathbb{P}(A_n) \to 0$ as $n \to \infty$. But $\mathbb{P}(A_n) = \mathbb{P}\left(x < X \leq x + \frac{1}{n}\right) = F\left(x + \frac{1}{n}\right) - F(x)$, hence $F(x + \frac{1}{n}) \to F(x)$ as required. Now, we want to show that left limits always exist. This is clear since F is an increasing function, and is always bounded above by 1.

(iv) We know $F(x^-) = \lim_{n\to\infty} F\left(x - \frac{1}{n}\right)$. Consider $B_n = \left\{X \leq x - \frac{1}{n}\right\}$. Then the B_n is an increasing sequence of events, and their union is $\{X < x\}$. Hence $\mathbb{P}(B_n)$ converges to $\mathbb{P}(X < x)$, so $F(x^-) = \mathbb{P}(X < x)$.

(v) This is evident from the properties of the probability measure.

□

12.2 Defining a continuous random variable

For a discrete random variable, F is a step function, which of course is right continuous with left limits.

Definition. A random variable X is called *continuous* if F is a continuous function. In this case, clearly left limits and right limits give the same value, and $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.

In this course, we will consider only *absolutely* continuous random variables. A continuous random variable is absolutely continuous if F is differentiable. We will make the convention that $F'(x) = f(x)$, where $f(x)$ is called the probability density function of X. The following immediate properties hold.

(i) $f \geq 0$

(ii) $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$

(iii) $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$

(iv) For $S \subseteq \mathbb{R}$, $\mathbb{P}(X \in S) = \int_S f(x)\,\mathrm{d}x$

Here is an intuitive explanation of the probability density function. Suppose Δx is a small quantity. Then

$$\mathbb{P}(x < X \leq x + \Delta x) = \int_x^{x+\Delta x} f(y)\,\mathrm{d}y \approx f(x) \cdot \Delta x$$

So we can think of $f(x)$ as the continuous analogy to $\mathbb{P}(X = x)$.

12.3 Expectation

Consider a continuous random variable $X : \Omega \to \mathbb{R}$, with probability distribution function $F(x)$ and probability density function $f(x) = F'(x)$. We define the expectation of such a *non-negative* random variable as

$$\mathbb{E}[X] = \int_0^\infty x f(x)\,\mathrm{d}x$$

In this case, the expectation is either non-negative and finite, or positive infinity. Now, let X be a general continuous random variable, that is not necessarily non-negative. Suppose $g \geq 0$. Then,

$$\mathbb{E}[g(X)] = \int_{-\infty}^\infty g(x) f(x)\,\mathrm{d}x$$

We can define $X_+ = \max(X, 0)$ and $X_- = \max(-X, 0)$. If at least one of $\mathbb{E}[X_+]$ or $\mathbb{E}[X_-]$ is finite, then clearly

$$\mathbb{E}[X] \coloneqq \mathbb{E}[X_+] - \mathbb{E}[X_-] = \int_{-\infty}^\infty x f(x)\,\mathrm{d}x$$

It is easy to verify that the expectation is a linear function, due to the linearity property of the integral.

12.4 Computing the expectation

Claim. Let $X \geq 0$. Then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq x)\,\mathrm{d}x$$

Proof. Using the definition of the expectation,

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x f(x)\,\mathrm{d}x \\
&= \int_0^\infty \left(\int_0^x \mathrm{d}y \right) f(x)\,\mathrm{d}x \\
&= \int_0^x \mathrm{d}y \int_y^\infty f(x)\,\mathrm{d}x \\
&= \int_0^\infty \mathrm{d}y \left(1 - \int_{-\infty}^y f(x)\,\mathrm{d}x \right) \\
&= \int_0^\infty \mathrm{d}y\, \mathbb{P}(X \geq y)
\end{aligned}$$

□

Here is an alternative proof.

Proof. For every $\omega \in \Omega$, we can write

$$X(\omega) = \int_0^\infty 1(X(\omega) \geq x)\,\mathrm{d}x$$

Taking expectations, we get

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty 1(X(\omega) \geq x)\,\mathrm{d}x\right]$$

We will interchange the integral and the expectation, although this step is not justified or rigorous.

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty \mathbb{E}[1(X(\omega) \geq x)]\,\mathrm{d}x \\ &= \int_0^\infty \mathbb{P}(X \geq x)\,\mathrm{d}x\end{aligned}$$

□

12.5 Variance

We define the variance of a continuous random variable as

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

12.6 Uniform distribution

Consider the uniform distribution defined by $a, b \in \mathbb{R}$.

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

We write $X \sim U[a,b]$. For some $x \in [a,b]$, we can write

$$\mathbb{P}(X \leq x) = \int_a^x f(y)\,\mathrm{d}y = \frac{x-a}{b-a}$$

Hence, for $x \in [a,b]$,

$$F(x) = \begin{cases} 1 & x > b \\ \frac{x-a}{b-a} & x \in [a,b] \\ 0 & x < a \end{cases}$$

Then,

$$\mathbb{E}[X] = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{a+b}{2}$$

12.7 Exponential distribution

The exponential distribution is defined by $f(x) = \lambda e^{-\lambda x}$ for $\lambda > 0, x > 0$. We write $X \sim \mathrm{Exp}(\lambda)$.

$$F(x) = \mathbb{P}(X \leq x) = \int_0^x \lambda e^{-\lambda y}\,\mathrm{d}y = 1 - e^{-\lambda x}$$

Further,

$$\mathbb{E}[X] = \int_0^\infty \lambda x e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}$$

We can view the exponential distribution as a limit of geometric distributions. Suppose that $T \sim \mathrm{Exp}(\lambda)$, and let $T_n = \lfloor nT \rfloor$ for all $n \in \mathbb{N}$. We have

$$\mathbb{P}(T_n \geq k) = \mathbb{P}\left(T \geq \frac{k}{n}\right) = e^{-\lambda k/n} = \left(e^{-\lambda/n}\right)^k$$

Hence T_n is a geometric distribution with parameter $p_n = e^{-\lambda/n}$. As $n \to \infty$, $p_n \sim \frac{\lambda}{n}$, and $\frac{T_n}{n} \sim T$. Hence the exponential distribution is the limit of a scaled version of the geometric distribution. A key property of the exponential distribution is that it has no memory. If $T \sim \mathrm{Exp}(\lambda)$, $\mathbb{P}(T > t + s \mid T > s) = \mathbb{P}(T > t)$. In fact, the distribution is uniquely characterised by this property.

Proposition. Let T be a positive continuous random variable not identically zero or infinity. Then T has the memoryless property $\mathbb{P}(T > t + s \mid T > s) = \mathbb{P}(T > t)$ if and only if $T \sim \mathrm{Exp}(\lambda)$ for some $\lambda > 0$.

Proof. Clearly if $T \sim \mathrm{Exp}(\lambda)$, then $\mathbb{P}(T > t + s \mid T > s) = e^{-\lambda t} = \mathbb{P}(T > t)$ as required. Now, given that T has this memoryless property, for all s and t, we have $\mathbb{P}(T > t + s) = \mathbb{P}(T > t)\,\mathbb{P}(T > s)$. Let $g(t) = \mathbb{P}(T > t)$; we would like to show that $g(t) = e^{-\lambda t}$. Then g satisfies $g(t+s) = g(t)g(s)$. Then for all $m \in \mathbb{N}$, $g(mt) = (g(t))^m$. Setting $t = 1$, $g(m) = g(1)^m$. Now, $g(m/n)^n = g(mn/n) = g(m)$ hence $g(m/n) = g(1)^{m/n}$. So for all rational numbers $q \in \mathbb{Q}$, $g(q) = g(1)^q$.

Now, $g(1) = \mathbb{P}(T > 1) \in (0, 1)$. Indeed, $g(1) \neq 0$ since in this case, for any rational number q we would have $g(q) = 0$ contradicting the assumption that T was not identically zero, and $g(1) \neq \infty$ because in this case T would be identically infinity. Now, let $\lambda = -\log \mathbb{P}(T > 1) > 0$. We have now proven that $g(t) = e^{-\lambda t}$ for all $t \in \mathbb{Q}$.

Let $t \in \mathbb{R}_+$. Then for all $\varepsilon > 0$, there exist $r, s \in \mathbb{Q}$ such that $r \leq t \leq s$ and $|r - s| \leq \varepsilon$. In this case, $e^{-\lambda s} = \mathbb{P}(T > s) \leq \mathbb{P}(T > t) \leq \mathbb{P}(T > r) = e^{-\lambda r}$. Sending $\varepsilon \to 0$ finishes the proof, showing that $g(t) = e^{-\lambda t}$ for all positive reals. □

12.8 Functions of continuous random variables

Theorem. Suppose that X is a continuous random variable with density f. Let g be a monotonic continuous function (either strictly increasing or strictly decreasing), such that g^{-1} is differentiable. Then $g(X)$ is a continuous random variable with density $fg^{-1}(x)\left|\frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x)\right|$.

Proof. Suppose that g is strictly increasing. We have

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

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(g(X) \leq x) = F'(g^{-1}(x)) \cdot \frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x) = f(g^{-1}(x))\frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x)$$

Note that since g is strictly increasing, so is g^{-1}. Now, suppose the g is strictly decreasing. Since the random variable is continuous,

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

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}\left(g(X) \leq x\right) = -F'(g^{-1}(x)) \cdot \frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x) = f(g^{-1}(x))\left|\frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x)\right|$$

Likewise, in this case, g is strictly decreasing. □

12.9 Normal distribution

The normal distribution is characterised by $\mu \in \mathbb{R}$ and $\sigma > 0$. We define

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$f(x)$ is indeed a probability density function:

$$I = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x$$

Applying the substitution $x \mapsto \frac{x-\mu}{\sigma}$, we have

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2}{2}\right\}\mathrm{d}x$$

We can evaluate this integral by considering I^2.

$$I^2 = \frac{2}{\pi} \int_0^{\infty} \int_0^{\infty} e^{\frac{-(u^2+v^2)}{2}}\,\mathrm{d}u\,\mathrm{d}v$$

Using polar coordinates $u = r\cos\theta$ and $v = r\sin\theta$, we have

$$I^2 = \frac{2}{\pi} \int_0^{\infty} \mathrm{d}r \int_0^{\frac{\pi}{2}} \mathrm{d}\theta\, r e^{-\frac{r^2}{2}} = 1 \implies I = \pm 1$$

But clearly $I > 0$, so $I = 1$. Hence f really is a probability density function. Now, if $X \sim \mathrm{N}(\mu, \sigma^2)$,

$$\begin{aligned}
\mathbb{E}\left[X\right] &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x \\
&= \underbrace{\int_{-\infty}^{\infty} \frac{x-\mu}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x}_{\text{odd function around } \mu \text{ hence } 0} + \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x}_{I=1 \text{ by above}} \\
&= \mu
\end{aligned}$$

We can also compute the variance, using the substitution $u = \frac{x-\mu}{\sigma}$, giving

$$\begin{aligned}\operatorname{Var}(X) &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x \\ &= \sigma^2 \int_{-\infty}^{\infty} \frac{u^2}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} \mathrm{d}u \\ &= \sigma^2\end{aligned}$$

In particular, when $\mu = 0$ and $\sigma^2 = 1$, we call the distribution $\mathrm{N}(\mu, \sigma^2) = \mathrm{N}(0,1)$ the standard normal distribution. We define

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \,\mathrm{d}u\,; \quad \phi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Hence $\Phi(x) = \mathbb{P}(X \le x)$ if X has the standard normal distribution. Since $\phi(x) = \phi(-x)$, we have $\Phi(x) + \Phi(-x) = 1$, hence $\mathbb{P}(X \le x) = 1 - \mathbb{P}(X \le -x)$.

13 Multivariate density functions

13.1 Standardising normal distributions

Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$. Let $a \neq 0, b \in \mathbb{R}$, and let $g(x) = ax + b$. We define $Y = g(X) = aX + b$. We can find the density f_Y of Y, by noting that g is a monotonic function and the inverse has a derivative. We can then use the theorem in the last lecture to show that

$$\begin{aligned}f_Y(y) &= f_X(g^{-1}(y)) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)\right| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\frac{y-b}{a} - \mu)^2}{2\sigma^2}\right) \cdot \frac{1}{2a} \\ &= \frac{1}{\sqrt{2\pi a^2\sigma^2}} \exp\left(-\frac{(y - a\mu + b)^2}{2a^2\sigma^2}\right)\end{aligned}$$

Hence $Y \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$. In particular, $\frac{X-\mu}{\sigma}$ is exactly the standard normal distribution.

Definition. Suppose X is a continuous random variable. Then the median of X, denoted by m, is the number satisfying

$$\mathbb{P}(X \le m) = \mathbb{P}(X \ge m) = \frac{1}{2}$$

If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\mathbb{P}(X \le \mu) = \Phi(0) = \frac{1}{2}$ hence μ is the median of the normal distribution.

13.2 Multivariate density functions

Suppose $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ is a random variable. We say that X has density f if

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f(y_1, \dots, y_n) \,\mathrm{d}y_1 \dots \mathrm{d}y_n$$

Then,

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

This generalises the fact that for all (reasonable) $B \subseteq \mathbb{R}^n$,

$$\mathbb{P}((X_1, \dots, X_n) \in B) = \int_B f(y_1, \dots, y_n) \, \mathrm{d}y_1 \dots \mathrm{d}y_n$$

13.3 Independence of events

In the continuous case, we can no longer use the definition

$$\mathbb{P}(X = a, Y = b) = \mathbb{P}(X = a)\,\mathbb{P}(Y = b)$$

since the probability of a random variable being a specific value is always zero. Instead, we define that $X_1, \dots, X_n$ are independent if for all $x_1, \dots, x_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n)$$

Theorem. Suppose $X = (X_1, \dots, X_n)$ has density f.

(a) Suppose $X_1, \dots, X_n$ are independent with densities $f_1, \dots, f_n$. Then $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$.

(b) Suppose that f factorises as $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$ for some non-negative functions $f_1, \dots, f_n$. Then $X_1, \dots, X_n$ are independent with densities proportional to $f_1, \dots, f_n$. (In order to have a density function, we require that it integrates to 1, so we choose a scaling factor such that this requirement holds.)

In other words, f factorises if and only if it is comprised of independent events.

Proof. (a) We know that

$$\begin{aligned}
\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) &= \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n) \\
&= \int_{-\infty}^{x_1} f_1(y_1) \, \mathrm{d}y_1 \cdots \int_{-\infty}^{x_n} f_n(y_n) \, \mathrm{d}y_n \\
&= \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} \prod_{i=1}^{n} f_i(y_i) \, \mathrm{d}y_i
\end{aligned}$$

So the density of $(X_1, \dots, X_n)$ is the product of the (f_i).

(b) Suppose f factorises. Let $B_1, \dots, B_n \subseteq \mathbb{R}$. Then

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \int_{B_1} \cdots \int_{B_n} f_1(x_1) \cdots f_n(x_n) \, \mathrm{d}y_1 \cdots \mathrm{d}y_n$$

Now, let $B_j = \mathbb{R}$ for all $j \neq i$. Then

$$\mathbb{P}(X_i \in B_i) = \mathbb{P}\left(X_i \in B_i, X_j \in B_j \; \forall j \neq i\right) = \int_{B_i} f_i(y_i) \, \mathrm{d}y_i \cdot \prod_{j \neq 1} \int_{B_j} f_j(x_j) \, \mathrm{d}y_j$$

Since f is a density function,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = 1$$

But f is the product of the f_i, so

$$\prod_j \int_{-\infty}^{\infty} f_j(y)\, \mathrm{d}y = 1 \implies \prod_{j \neq i} \int_{-\infty}^{\infty} f_j(y)\, \mathrm{d}y = \frac{1}{\int_{-\infty}^{\infty} f_i(y)\, \mathrm{d}y}$$

Hence,

$$\mathbb{P}(X_i \in B_i) = \frac{\int_{B_i} f_i(y)\, \mathrm{d}y}{\int_{-\infty}^{\infty} f_i(y)\, \mathrm{d}y}$$

This shows that the density of X_i is

$$\frac{f_i}{\int_{-\infty}^{\infty} f_i(y)\, \mathrm{d}y}$$

The X_i are independent, since

$$\begin{aligned}
\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) &= \frac{\int_{-\infty}^{x_1} f_1(y_1)\, \mathrm{d}y_1 \cdots \int_{-\infty}^{x_n} f_n(y_n)\, \mathrm{d}y_n}{\int_{-\infty}^{\infty} f_1(y_1)\, \mathrm{d}y_1 \cdots \int_{-\infty}^{\infty} f_n(y_n)\, \mathrm{d}y_n} \\
&= \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n)
\end{aligned}$$

□

13.4 Marginal density

Suppose that $(X_1, \ldots, X_n)$ has density f. Now we can compute the marginal density as follows.

$$\begin{aligned}
\mathbb{P}(X_1 \leq x) &= \mathbb{P}(X_1 \leq x, X_2 \in \mathbb{R}, \ldots, X_n \in \mathbb{R}) \\
&= \int_{-\infty}^{x} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \int_{-\infty}^{x} \mathrm{d}x_1 \underbrace{\left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, \mathrm{d}x_2 \cdots \mathrm{d}x_n \right)}_{\text{marginal density of } X_1}
\end{aligned}$$

13.5 Sum of random variables

Recall that in the discrete case, for independent random variables X and Y we have

$$\begin{aligned}
\mathbb{P}(X + Y = z) &= \sum_y \mathbb{P}(X + Y = z, Y = y) \\
&= \sum_y \mathbb{P}(X = z - y)\, \mathbb{P}(Y = y) \\
&= \sum_y p_x(z - y) p_y(y)
\end{aligned}$$

which was called the convolution. In the continuous case,

$$\begin{aligned}
\mathbb{P}(X+Y\le z) &= \iint_{\{x+y\le z\}} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z} f_X(x)f_Y(y-x)\,\mathrm{d}y\right)\mathrm{d}x \ \ (\text{using } y\mapsto y+x) \\
&= \int_{-\infty}^{z}\mathrm{d}y\,\underbrace{\left(\int_{-\infty}^{\infty} f_Y(y-x)f_X(x)\,\mathrm{d}x\right)}_{g(y)}
\end{aligned}$$

Hence the density of $X+Y$ is $g(y)$, where

$$g(y)=\int_{-\infty}^{\infty} f_Y(y-x)f_X(x)\,\mathrm{d}x$$

Definition. Let f, g be density functions. Then the convolution of f and g is

$$(f\star g)(y)=\int_{-\infty}^{\infty} f_Y(y-x)f_X(x)\,\mathrm{d}x$$

Here is a non-rigorous argument, which can be used as a heuristic.

$$\begin{aligned}
\mathbb{P}(X+Y\le z) &= \int_{-\infty}^{\infty}\mathbb{P}(X+Y\le z, Y\in\mathrm{d}y) \\
&= \int_{-\infty}^{\infty}\mathbb{P}(X+Y\le z, Y\in\mathrm{d}y) \\
&= \int_{-\infty}^{\infty}\mathbb{P}(X\le z-y)\,\mathbb{P}(Y\in\mathrm{d}y) \\
&= \int_{-\infty}^{\infty}\mathbb{P}(X\le z-y)\,f_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}F_X(z-y)f_Y(y)\,\mathrm{d}y \\
\frac{\mathrm{d}}{\mathrm{d}z}\mathbb{P}(X+Y\le z) &= \int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}z}F_X(z-y)f_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}f_X(z-y)f_Y(y)\,\mathrm{d}y
\end{aligned}$$

13.6 Conditional density

We will now define the conditional density of a continuous random variable, given the value of another continuous random variable. Let X and Y be continuous random variables with joint density

$f_{X,Y}$ and marginal densities f_X and f_Y. Then we define the conditional density of X given that $Y = y$ is defined as

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

Then we can find the law of total probability in the continuous case.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)\,\mathrm{d}y \end{aligned}$$

13.7 Conditional expectation

We want to define $\mathbb{E}[X \mid Y]$ to be some function $g(Y)$ for some function g. We will define

$$g(y) = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x$$

which is the analogous expression to $\mathbb{E}[X \mid Y = y]$ from the discrete case. Then we just set $\mathbb{E}[X \mid Y] = g(Y)$ to be the conditional expectation.

13.8 Transformations of multidimensional random variables

Theorem. Let X be a continuous random variable with values in $D \subseteq \mathbb{R}^d$, with density f_X. Now, let g be a bijection D to $g(D)$ which has a continuous derivative, and $\det g'(x) \neq 0$ for all $x \in D$. Then the random variable $Y = g(X)$ has density

$$f_Y(y) = f_X(x) \cdot |J| \text{ where } x = g^{-1}(y)$$

where J is the Jacobian

$$J = \det\left(\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j=1}^{d}\right)$$

No proof will be given for this theorem. As an example, let X and Y be independent continuous random variables with the standard normal distribution. The point (X, Y) in $\mathbb{R}^2$ has polar coordinates (R, Θ). What are the densities of R and Θ? We have $X = R\cos\Theta$ and $Y = R\sin\Theta$. The Jacobian is

$$J = \det\begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} = r$$

Hence,

$$\begin{aligned} f_{R,\Theta}(r,\theta) &= f_{X,Y}(r\cos\theta, r\sin\theta)|J| \\ &= f_{X,Y}(r\cos\theta, r\sin\theta)r \\ &= f_X(r\cos\theta) f_Y(r\sin\theta) r \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{r^2\cos^2\theta}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{r^2\sin^2\theta}{2}} \cdot r \\ &= \frac{1}{2\pi} e^{-\frac{r^2}{2}} \cdot r \end{aligned}$$

for all $r > 0$ and $\theta \in [0, 2\pi]$. Note that the joint density factorises into marginal densities:

$$f_{R,\Theta}(r,\theta) = \underbrace{\frac{1}{2\pi}}_{f_\Theta} \underbrace{re^{-\frac{r^2}{2}}}_{f_R}$$

so the random variables R and Θ are independent, where $\Theta \sim U[0, 2\pi]$ and R has density $re^{\frac{-r^2}{2}}$ on $(0, \infty)$.

13.9 Order statistics of a random sample

Let $X_1, \dots, X_n$ be independent and identically distributed random variables with distribution function F and density function f. We can put them in increasing order:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

and let $Y_i = X_{(i)}$. The (Y_i) are the order statistics.

$$\begin{aligned}
\mathbb{P}(Y_1 \leq x) &= \mathbb{P}(\min(X_1, \dots, X_n) \leq x) \\
&= 1 - \mathbb{P}(\min(X_1, \dots, X_n) > x) \\
&= 1 - \mathbb{P}(X_1 > x) \cdots \mathbb{P}(X_n > x) \\
&= 1 - (1 - F(x))^n
\end{aligned}$$

Further,

$$\begin{aligned}
f_{Y_1}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}(1 - (1 - F(x))^n) \\
&= n(1 - F(x))^{n-1} f(x)
\end{aligned}$$

We can compute an analogous result for the maximum.

$$\begin{aligned}
\mathbb{P}(Y_n \leq x) &= (F(x))^n \\
f_{Y_n}(x) &= n(F(x))^{n-1} f(x)
\end{aligned}$$

What are the densities of the other random variables? First, let $x_1 < x_2 < \cdots < x_n$. Then, we can first find the joint distribution $\mathbb{P}(Y_1 \leq x_1, \dots, Y_n \leq x_n)$. Note that this is simply the sum over all possible permutations of the (X_i) of $\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n)$. But since the variables are independent and identically distributed, these probabilities are the same. Hence,

$$\begin{aligned}
\mathbb{P}(Y_1 \leq x_1, \dots, Y_n \leq x_n) &= n! \cdot \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n, X_1 < \cdots < X_n) \\
&= n! \int_{-\infty}^{x_1} \int_{u_1}^{x_2} \cdots \int_{u_{n-1}}^{x_n} f(u_1) \cdots f(u_n) \, \mathrm{d}u_1 \cdots \mathrm{d}u_n \\
\therefore f_{Y_1,\dots,Y_n}(x_1, \dots, x_n) &= n! f(x_1) \cdots f(x_n)
\end{aligned}$$

when $x_1 < x_2 < \cdots < x_n$, and the joint density is zero otherwise. Note that this joint density does not factorise as a product of densities, since we must always consider the indicator function that $x_1 < x_2 < \cdots < x_n$.

13.10 Order statistics on exponential distribution

Let $X \sim \text{Exp}(\lambda)$, $Y \sim \text{Exp}(\mu)$ be independent continuous random variables. Let $Z = \min(X, Y)$.

$$\mathbb{P}(Z \geq z) = \mathbb{P}(X \geq z, Y \geq z) = \mathbb{P}(X \geq z)\,\mathbb{P}(Y \geq z) = e^{-\lambda z} \cdot e^{-\mu z} = e^{-(\lambda+\mu)z}$$

Hence Z has the exponential distribution with parameter $\lambda + \mu$. More generally, if $X_1, \dots, X_n$ are independent continuous random variables with $X_i \sim \text{Exp}(\lambda_i)$, then $Z = \min(X_1, \dots, X_n)$ has distribution $\text{Exp}\left(\sum_{i=1}^{n} \lambda_i\right)$. Now, let $X_1, \dots, X_n$ be independent identically distributed random variables with distribution $\text{Exp}(\lambda)$, and let Y_i be their order statistics. Then

$$Z_1 = Y_1; \quad Z_2 = Y_2 - Y_1; \quad Z_i = Y_i - Y_{i-1}$$

So the Z_i are the 'durations between consecutive results' from the X_i. What is the density of these Z_i? First, note that

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = A \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}; \quad A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Note that $\det A = 1$, and $Z = AY$, and note further that

$$y_j = \sum_{i=1}^{j} z_i$$

Now,

$$\begin{aligned} f_{(Z_1,\dots,Z_n)}(z_1, \dots, z_n) &= f_{(Y_1,\dots,Y_n)}(y_1, \dots, y_n) \underbrace{|A|}_{=1} \\ &= n! f(y_1) \cdots f(y_n) \\ &= n!(\lambda e^{-\lambda y_1}) \cdots (\lambda e^{-\lambda y_n}) \\ &= n! \lambda^n e^{-\lambda(n z_1 + (n-1) z_2 + \cdots + z_n)} \\ &= \prod_{i=1}^{n} (n - i + 1) \lambda e^{-\lambda(n-i+1) z_i} \end{aligned}$$

The density function of the vector Z factorises into functions of the z_i, so $Z_1, \dots, Z_n$ are independent and $Z_i \sim \text{Exp}(\lambda(n - i + 1))$.

14 Moment generating functions

14.1 Moment generating functions

Consider a continuous random variable X with density f. Then the moment generating function of X is defined as

$$m(\theta) = \mathbb{E}\left[e^{\theta X}\right] = \int_{-\infty}^{\infty} e^{\theta x} f(x)\,\mathrm{d}x$$

whenever this integral is finite. Note that $m(0) = 1$.

Theorem. The moment generating function uniquely determines the distribution of a continuous random variable, provided that it is defined on some open interval (a, b) of values of θ.

No proof will be given.

Theorem. Suppose the moment generating function is defined on an open interval of values of θ. Then

$$\frac{\mathrm{d}^r}{\mathrm{d}\theta^r} m(\theta)\bigg|_{\theta=0} = \mathbb{E}[X^r]$$

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Then

$$m(\theta) = \mathbb{E}\left[e^{\theta(X_1+\cdots+X_n)}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\theta X_i}\right]$$

Proof. Since the X_i are independent, we can move the product outside of the expectation. □

14.2 Gamma distribution

Let X be a random variable with density

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

where $\lambda > 0$, $n \in \mathbb{N}$, $x \geq 0$. We can say that $X \sim \Gamma(n, \lambda)$. First, we check that f is indeed a density.

$$\begin{aligned}
I_n &= \int_0^\infty f(x)\,\mathrm{d}x \\
&= \int_0^\infty \lambda e^{-\lambda x} \frac{\lambda^n x^{n-1}}{(n-1)!}\,\mathrm{d}x \\
&= \int_0^\infty \frac{e^{-\lambda x}\lambda^{n-1}(n-1)x^{n-2}}{(n-1)!}\,\mathrm{d}x \\
&= \int_0^\infty \frac{e^{-\lambda x}\lambda^{n-1}x^{n-2}}{(n-2)!}\,\mathrm{d}x \\
&= I_{n-1} = \cdots = I_1
\end{aligned}$$

Note that for $n = 1$, $f(x) = \lambda e^{-\lambda x}$ which is the density of the exponential distribution. Therefore, $I_n = 1$ as required, so f really is a density. Now,

$$m(\theta) = \int_0^\infty \frac{e^{\theta x} e^{-\lambda x}\lambda^n x^{n-1}}{(n-1)!}\,\mathrm{d}x$$

If $\lambda > \theta$, then we have a finite integral. If $\lambda \leq \theta$, then the exponential term $e^{\theta x}$ will dominate and we will have an infinite integral. So, let $\lambda > \theta$.

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

The integral on the right hand side is the probability distribution function of a random variable $Y \sim \Gamma(n, \lambda - \theta)$, which gives 1 since the integral is taken over the entire domain. Hence,

$$m(\theta) = \left(\frac{\lambda}{\lambda - \theta}\right)^n$$

Now, let $X \sim \Gamma(n, \lambda)$ and $Y \sim \Gamma(m, \lambda)$ be independent continuous random variables. Then

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

So by the uniqueness property we saw earlier, we get that $X + Y \sim \Gamma(n + m, \lambda)$. In particular, this implies that if $X_1, \ldots, X_n$ are independent and identically distributed with the distribution $\mathrm{Exp}(\lambda) = \Gamma(1, \lambda)$, then

$$X_1 + \cdots + X_n \sim \Gamma(n, \lambda)$$

We could alternatively consider $\Gamma(\alpha, \lambda)$ for $\alpha > 0$ by replacing $(n-1)!$ with

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha - 1} \, \mathrm{d}x$$

which agrees with this factorial function for integer values of α.

14.3 Moment generating function of the normal distribution

Recall that

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Now,

$$m(\theta) = \int_0^\infty e^{\theta x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\theta x - \frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x$$

Note that

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

Hence,

$$\begin{aligned} m(\theta) &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\theta\mu + \frac{\theta^2\sigma^2}{2} - \frac{\left(x - (\mu + \theta\sigma^2)\right)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \exp\left(\theta\mu + \frac{\theta^2\sigma^2}{2}\right) \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x - (\mu + \theta\sigma^2)\right)^2}{2\sigma^2}\right) \mathrm{d}x \end{aligned}$$

Note that the integral on the right hand side has the form of the probability distribution function of a variable $Y \sim \mathrm{N}(\mu + \theta\sigma^2, \sigma^2)$, hence it integrates to 1.

$$m(\theta) = \exp\left(\theta\mu + \frac{\theta^2\sigma^2}{2}\right)$$

Recall that if $X \sim \mathrm{N}(\mu, \sigma^2)$, then $aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$. We can then deduce that

$$\mathbb{E}\left[e^{\theta(aX+b)}\right] = \exp\left(\theta(a\mu + b) + \frac{\theta^2 a^2\sigma^2}{2}\right)$$

Now, suppose that $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Y \sim \mathrm{N}(\nu, \tau^2)$ are independent. Then

$$\begin{aligned}\mathbb{E}\left[e^{\theta(X+Y)}\right] &= \mathbb{E}\left[e^{\theta X}\right]\mathbb{E}\left[e^{\theta Y}\right] \\ &= \exp\left(\theta\mu + \frac{\theta^2\sigma^2}{2}\right)\exp\left(\theta\nu + \frac{\theta^2\tau^2}{2}\right) \\ &= \exp\left(\theta(\mu+\nu) + \frac{\theta^2(\sigma^2+\tau^2)}{2}\right)\end{aligned}$$

Hence $X + Y \sim \mathrm{N}(\mu + \nu, \sigma^2 + \tau^2)$.

14.4 Cauchy distribution

Suppose that a continuous random variable X has density

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

where $x \in \mathbb{R}$. Now,

$$m(\theta) = \mathbb{E}\left[e^{\theta X}\right] = \int_{-\infty}^{\infty} \frac{e^{\theta x}}{\pi(1+x^2)} = \begin{cases}\infty & \theta \neq 0 \\ 1 & \theta = 0\end{cases}$$

Suppose $X \sim f$. Then $X, 2X, 3X, \ldots$ have the same moment generating function, but they do not have the same distribution. This is because $m(\theta)$ is not finite on an open interval.

14.5 Multivariate moment generating functions

Let $X = (X_1, \ldots, X_n)$ be a random variable with values in $\mathbb{R}^n$. Then the moment generating function of X is defined as

$$m(\theta) = \mathbb{E}\left[e^{\theta^\intercal X}\right] = \mathbb{E}\left[e^{\theta_1 X_1 + \cdots + \theta_n X_n}\right]; \quad \theta = \begin{pmatrix}\theta_1 \\ \vdots \\ \theta_n\end{pmatrix}$$

Theorem. If the moment generating function is finite for a range of values of θ, it uniquely determines the distribution of X. Also,

$$\left.\frac{\partial^r m}{\partial \theta_i^r}\right|_{\theta=0} = \mathbb{E}\left[X_i^r\right]$$

and

$$\left.\frac{\partial^{r+s} m}{\partial\theta_i^r \partial\theta_j^s}\right|_{\theta=0} = \mathbb{E}\left[X_i^r X_j^s\right]$$

Further,

$$m(\theta) = \prod_{i=1}^{n} \mathbb{E}\left[e^{\theta_i X_i}\right]$$

if and only if $X_1, \ldots, X_n$ are independent.

No proof is provided.

15 Limit theorems

15.1 Convergence in distribution

Definition. Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables and let X be another random variable. We say that X_n converges to X in distribution, written $X_n \xrightarrow{d} X$, if

$$F_{X_n}(x) \to F_X(x)$$

for all $x \in \mathbb{R}$ that are continuity points of F_X.

Theorem (Continuity property for moment generating functions)**.** Let X be a continuous random variable with $m(\theta) < \infty$ for some $\theta \neq 0$. Suppose that $m_n(\theta) \to m(\theta)$ for all $\theta \in \mathbb{R}$, where $m_n(\theta) = \mathbb{E}\left[e^{\theta X_n}\right]$, and $m(\theta) = \mathbb{E}\left[e^{\theta X}\right]$. Then $X_n \xrightarrow{d} X$.

15.2 Weak law of large numbers

Theorem. Let $(X_n : n \in \mathbb{N})$ be a sequence of independent and identically distributed random variables, with $\mu = \mathbb{E}[X_1] < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \to 0$$

as $n \to \infty$.

We will give a proof assuming that the variance of X_1 is finite.

Proof. By Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n}-\mu\right|>\varepsilon\right) &= \mathbb{P}\left(|S_n-n\mu|>\varepsilon n\right)\\ &\leq \frac{\operatorname{Var}(S_n)}{\varepsilon^2 n^2}\\ &= \frac{n\sigma^2}{\varepsilon^2 n^2}\\ &\to 0\end{aligned}$$

□

15.3 Types of convergence

Definition. A sequence (X_n) converges to X *in probability*, written $X_n \xrightarrow{\mathbb{P}} X$ as $n \to \infty$ if for all $\varepsilon > 0$,

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0; \quad n \to \infty$$

Definition. A sequence (X_n) converges to X *almost surely* (with probability 1), if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$$

This second definition is a stronger form of convergence. If a sequence (X_n) converges to zero almost surely, then $X_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$.

Proof. We want to show that given any $\varepsilon > 0$, $\mathbb{P}(|X_n| > \varepsilon) \to 0$ as $n \to \infty$, or equivalently, $\mathbb{P}(|X_n| \leq \varepsilon) \to 1$.

$$\mathbb{P}\left(|X_n| \leq \varepsilon\right) \geq \mathbb{P}\left(\underbrace{\bigcap_{m=n}^{\infty} \{|X_m| \leq \varepsilon\}}_{A_n}\right)$$

Note that A_n is an increasing sequence of events, and

$$\bigcup_n A_n = \{|X_m| \leq \varepsilon \text{ for all } m \text{ sufficiently large}\}$$

Hence, as $n \to \infty$,

$$\mathbb{P}(A_n) \to \mathbb{P}\left(\bigcup A_n\right)$$

Therefore,

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n| \leq \varepsilon\right) \geq \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup A_n\right) \geq \mathbb{P}\left(\lim_{n\to\infty} X_n = 0\right)$$

Since X_n converges to zero almost surely, this event on the right hand side has probability 1, so in particular the limit on the left has probability 1, as required. □

15.4 Strong law of large numbers

Theorem. Let $(X_n)_{n\in\mathbb{N}}$ be an independent and identically distributed sequence of random variables, with $\mu = \mathbb{E}[X_1]$ finite. Let $S_n = X_1 + \cdots + X_n$. Then

$$\frac{S_n}{n} \to \mu \text{ as } n \to \infty \text{ almost surely}$$

In other words,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \to \mu\right) = 1$$

The following proof, made under the assumption of a finite fourth moment, is non-examinable. A proof can be formulated without this assumption, but it is more complicated.

Proof. Let $Y_i = X_i - \mu$. Then $\mathbb{E}[Y_i] = 0$, and $\mathbb{E}\left[Y_i^4\right] \le 2^4(\mathbb{E}\left[X_i^4\right] + \mu^4) < \infty$. It then suffices to show that

$$\frac{S_n}{n} \to 0 \text{ a.s.}$$

where $S_n = \sum_1^n X_i$ and $\mathbb{E}[X_i] = 0$, $\mathbb{E}\left[X_i^4\right] < \infty$. First,

$$S_n^4 = \left(\sum_{i=1}^n X_i\right)^4 = \sum_{i=1}^n X_i^4 + \binom{4}{2}\sum_{i=1}^n X_i^2 X_j^2 + R$$

where R is a sum of terms of the form $X_i^2 X_j X_k$ or $X_i^3 X_j$ or $X_i X_j X_k X_\ell$ for i, j, k, l distinct. Once we take expectations, each term in R will have no contribution to the result, since they all contain an $\mathbb{E}[X_i] = 0$ term.

$$\begin{aligned}
\mathbb{E}\left[S_n^4\right] &= n\mathbb{E}\left[X_i^4\right] + \binom{4}{2}\frac{n(n-1)}{2}\mathbb{E}\left[X_i^2 X_j^2\right] + \mathbb{E}[R] \\
&= n\mathbb{E}\left[X_1^4\right] + 3n(n-1)\mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_1^2\right] \\
&\le n\mathbb{E}\left[X_1^4\right] + 3n(n-1)\mathbb{E}\left[X_1^4\right] \\
&= 3n^2\mathbb{E}\left[X_1^4\right]
\end{aligned}$$

by Jensen's inequality. Now,

$$\mathbb{E}\left[\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4\right] \le \sum_{n=1}^\infty \frac{3}{n^2}\mathbb{E}\left[X_1^4\right] < \infty$$

Hence,

$$\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4 < \infty \text{ with probability 1}$$

Then since the sum of infinitely many positive terms is finite, the terms must converge to zero.

$$\lim_{n\to\infty} \frac{S_n}{n} \to 0 \text{ a.s.}$$

□

15.5 Central limit theorem

Suppose, like before, that we have a sequence of independent and identically distributed random variables X_n, and suppose further that $\mathbb{E}[X_1] = \mu$, and $\operatorname{Var}(X_1) = \sigma^2 < \infty$.

$$\operatorname{Var}\left(\frac{S_n}{n} - \mu\right) = \frac{\sigma^2}{n}$$

We can normalise this new random variable $\frac{S_n}{n} - \mu$ by dividing by its standard deviation.

$$\frac{\frac{S_n}{n} - \mu}{\sqrt{\operatorname{Var}\left(\frac{S_n}{n} - \mu\right)}} = \frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{Sn - n\mu}{\sigma\sqrt{n}}$$

Theorem. For all $x \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x) = \int_{-\infty}^{x} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}y$$

In other words,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z$$

where Z is the standard normal distribution.

Less formally, we might say that the central limit theorem shows that, for a large n,

$$S_n \approx n\mu + \sigma\sqrt{n}Z \sim N(n\mu, n\sigma^2)$$

Proof. Consider $Y_i = \frac{X_i - \mu}{\sigma}$. Then the Y_i have zero expectation and unit variance. It then suffices to prove the central limit theorem when the X_i have zero expectation and unit variance. We assume further that there exists $\delta > 0$ such that

$$\mathbb{E}\left[e^{\delta X_1}\right] < \infty; \quad \mathbb{E}\left[e^{-\delta X_1}\right] < \infty$$

We will show that

$$\frac{S_n}{n} \xrightarrow{d} \mathrm{N}(0, 1)$$

By the continuity property of moment generating functions, it is sufficient to show that for all $\theta \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\left[e^{\frac{\theta S_n}{n}}\right] = \mathbb{E}\left[e^{\theta Z}\right] = e^{\frac{\theta^2}{2}}$$

Let $m(\theta) = \mathbb{E}\left[e^{\theta X_1}\right]$. Then

$$\mathbb{E}\left[e^{\frac{\theta S_n}{n}}\right] = \mathbb{E}\left[e^{\frac{\theta}{\sqrt{n}}X_1}\right]^n = \left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n$$

We now need to show that

$$\lim_{n\to\infty} \left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n = e^{\frac{\theta^2}{2}}$$

Now, let $|\theta| < \frac{\delta}{2}$. In this case,

$$\begin{aligned} m(\theta) &= \mathbb{E}\left[e^{\theta X_1}\right] \\ &= \mathbb{E}\left[1 + \theta X_1 + \frac{\theta^2}{2}X_1^2 + \sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right] \\ &= \mathbb{E}\left[1\right] + \mathbb{E}\left[\theta X_1\right] + \mathbb{E}\left[\frac{\theta^2}{2}X_1^2\right] + \mathbb{E}\left[\sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right] \\ &= 1 + \frac{\theta^2}{2} + \mathbb{E}\left[\sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right] \end{aligned}$$

Now, it suffices to prove that $\left|\mathbb{E}\left[\sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right]\right| = o(\theta^2)$ as $\theta \to 0$. Indeed, if we have this bound, then $m\left(\frac{\theta}{\sqrt{n}}\right) = 1 + \frac{\theta^2}{2n} + o\left(\frac{\theta^2}{n}\right)$, and hence $\lim_{n\to\infty}\left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n = e^{\frac{\theta^2}{2}}$. To find this bound, we know that

$$\begin{aligned} \left|\mathbb{E}\left[\sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right]\right| &\le \mathbb{E}\left[\sum_{k=3}^{\infty}\frac{|\theta|^k|X_1|^k}{k!}\right] \\ &= \mathbb{E}\left[|\theta X_1|^3\sum_{k=0}^{\infty}\frac{|\theta X_1|^k}{(k+3)!}\right] \\ &\le \mathbb{E}\left[|\theta X_1|^3\sum_{k=0}^{\infty}\frac{|\theta X_1|^k}{k!}\right] \end{aligned}$$

Since $|\theta| \le \frac{\delta}{2}$,

$$\mathbb{E}\left[|\theta X_1|^3\sum_{k=0}^{\infty}\frac{|\theta X_1|^k}{k!}\right] \le \mathbb{E}\left[|\theta X_1|^3 e^{\frac{\delta}{2}|X_1|}\right]$$

Now,

$$|\theta X_1|^3 e^{\frac{\delta}{2}|X_1|} = |\theta|^3\frac{\left(\frac{\delta}{2}|X_1|\right)^3}{3!}\cdot\frac{3!}{\left(\frac{\delta}{2}\right)^3}\cdot e^{\frac{\delta}{2}|X_1|}$$

Note that

$$\frac{\left(\frac{\delta}{2}|X_1|\right)^3}{3!} \le \sum_{k=0}^{\infty}\frac{\left(\frac{\delta}{2}|X_1|\right)^k}{k!} = e^{\frac{\delta}{2}|X_1|}$$

Hence,

$$|\theta X_1|^3 e^{\frac{\delta}{2}|X_1|} \le |\theta|^3 e^{\frac{\delta}{2}|X_1|}\cdot\frac{3!}{\left(\frac{\delta}{2}\right)^3}\cdot e^{\frac{\delta}{2}|X_1|} = \frac{3!|\theta|^3}{\left(\frac{\delta}{2}\right)^3}e^{\delta|X_1|} = 3!\left(\frac{2|\theta|}{\delta}\right)^3 e^{\delta|X_1|}$$

Therefore,

$$e^{\delta|X_1|} \le e^{\delta X_1} + e^{-\delta X_1}$$

So finally,

$$\mathbb{E}\left[|\theta X_1|^3\sum_{k=0}^{\infty}\frac{|\theta X_1|^k}{k!}\right] \le 3!\left(\frac{2|\theta|}{\delta}\right)^3\mathbb{E}\left[e^{\delta X_1} + e^{-\delta X_1}\right] = o(|\theta|^2)$$

as $\theta \to 0$. □

15.6 Applications of central limit theorem

We can use the central limit theorem to approximate the binomial distribution using the normal distribution. Suppose that $S_n \sim \text{Bin}(n, p)$. Then $S_n = \sum_{i=1}^{n} X_i$, where the X_i have the Bernoulli distribution with parameter p. We know that $\mathbb{E}[S_n] = np$, and $\text{Var}(S_n) = np(1-p)$. Therefore, in particular,

$$S_n \approx \text{N}(np, np(1-p))$$

for n large. Note that we showed before that

$$\text{Bin}\left(n, \frac{\lambda}{n}\right) \to \text{Poi}(\lambda)$$

Note that with this approximation to the binomial, we let the parameter p depend on n. Since this is the case, we can no longer apply the central limit theorem, and we get a Poisson distributed approximation.

We can, however, use the central limit theorem to find a normal approximation for a Poisson random variable $S_n \sim \text{Poi}(n)$, since S_n can be written as $\sum_{i=1}^{n} X_i$ where the $X_i \sim \text{Poi}(1)$. Then

$$S_n \approx \text{N}(n, n)$$

15.7 Sampling error via central limit theorem

Suppose individuals independently vote 'yes' (with probability p) or 'no' (with probability $1-p$). We can sample the population to find an approximation for p. Pick N individuals at random, and let $\hat{p}_N = \frac{S_N}{N}$, where S_n is the number of individuals who voted 'yes'. We would like to find the minimum N such that $|\hat{p}_N - p| \leq 4\%$ with probability at least 99%. We have

$$S_N \sim \text{Bin}(N, p) \approx Np + \sqrt{Np(1-p)}Z; \quad Z \sim \text{N}(0, 1)$$

Hence,

$$\frac{S_N}{N} \approx p + \sqrt{\frac{p(1-p)}{N}}Z \implies |\hat{p}_N - p| \approx \sqrt{\frac{p(1-p)}{N}}|Z|$$

We then want to find N such that

$$\mathbb{P}\left(\sqrt{\frac{p(1-p)}{N}}|Z| \leq 0.04\right) \geq 0.99$$

We can compute this from the tables of the standard normal distribution. If $z = 2.58$, then $\mathbb{P}(|Z| \geq 2.58) = 0.01$, hence we need an N such that

$$0.04\sqrt{\frac{N}{p(1-p)}} \geq 2.58$$

In the worst case scenario, $p = \frac{1}{2}$ would give the largest N. So we need $N \geq 1040$ to get a good result for all p.

15.8 Buffon's needle

Consider a set of parallel lines on a plane, all a distance L apart. Imagine dropping a needle of length $\ell \leq L$ onto this plane at random. What is the probability that it intersects at least one line?

We will interpret a random drop to be represented by independent values x and θ, where x is the perpendicular distance from the lower end of the needle to the nearest line above it, and θ is the angle between the horizontal and the needle, where a value of $\theta = 0$ means that the needle is horizontal, and higher values of θ mean that the needle has been rotated θ radians anticlockwise. We assume that $\Theta \sim \mathrm{U}[0, \pi]$, and $X \sim \mathrm{U}[0, L]$, and that they are independent. The needle intersects a line if and only if $\ell \sin\theta \geq x$. We have

$$\begin{aligned}\mathbb{P}(\text{intersection}) &= \mathbb{P}(X \leq \ell \sin\Theta) \\ &= \int_0^L \int_0^\pi \frac{1}{\pi L} 1(x \leq \ell \sin\theta)\, \mathrm{d}x\, \mathrm{d}\theta \\ &= \frac{2\ell}{\pi L}\end{aligned}$$

Let this probability be denoted by p. So we can compute an approximation to π by finding

$$\pi = \frac{2\ell}{pL}$$

We can use the sampling error calculation above to find the amount of needles required to get a good approximation to π (within 0.1%) with probability ast least 99%, so we want

$$\mathbb{P}(|\hat{\pi}_n - \pi| \leq 0.001) \geq 0.99$$

Let S_n be the number of needles intersecting a line. Then $S_n \sim \mathrm{Bin}(n, p)$. So by the central limit theorem,

$$S_n \approx np + \sqrt{np(1-p)}Z \implies \hat{p}_n = \frac{S_n}{n} = p + \sqrt{\frac{p(1-p)}{n}}Z$$

Hence,

$$\hat{p}_n - p \approx \sqrt{\frac{p(1-p)}{n}}Z$$

Now, let $f(x) = 2\ell/xL$. Then $f(p) = \pi$, $f'(p) = -\frac{\pi}{p}$, and $\hat{\pi}_n = f(\hat{p}_n)$. We can then use a Taylor expansion to find

$$\hat{\pi}_n = f(\hat{p}_n) \approx f(p) + (\hat{p}_n - p)f'(p) \implies \hat{\pi}_n \approx \pi - (\hat{p}_n - p)\frac{\pi}{p}$$

Hence,

$$\hat{\pi}_n - \pi \approx -\frac{\pi}{p}\sqrt{\frac{p(1-p)}{n}} = -\pi\sqrt{\frac{1-p}{pn}}Z$$

We want

$$\mathbb{P}\left(\pi\sqrt{\frac{1-p}{pn}}|Z| \leq 0.001\right) \geq 0.99$$

So using tables, we find in the worst case scenario that $n \approx 3.75 \times 10^7$. So this approximation becomes good very slowly.

15.9 Bertrand's paradox

Consider a circle of radius r, and draw a random chord on the circle. What is the probability that its length C is less than r? There are two interpretations of the words 'random chord', that give different results. This is Bertrand's paradox.

(i) First, let us interpret 'random chord' as follows. Let $X \sim \mathrm{U}[0, r]$, and then we draw a chord perpendicular to a radius, such that it intersects the radius at a distance of X from the origin. Then we have formed a triangle between this intersection point, one end of the chord, and the circle's centre. By Pythagoras' theorem, the length of the chord is then twice the height of this triangle, so $C = 2\sqrt{r^2 - X^2}$. Hence,

$$\begin{aligned}\mathbb{P}(C \le r) &= \mathbb{P}\left(2\sqrt{r^2 - X^2} \le r\right) \\ &= \mathbb{P}\left(4(r^2 - X^2) \le r^2\right) \\ &= \mathbb{P}\left(X \ge \frac{\sqrt{3}}{2} r\right) \\ &= 1 - \frac{\sqrt{3}}{2} \approx 0.134\end{aligned}$$

(ii) Instead, let us fix one end point of the chord A, and let $\Theta \sim \mathrm{U}[0, 2\pi]$. Let the other end point B be such that the angle between the radii OA and OB is Θ. Then if $\Theta \in [0, \pi]$, the length of the chord can be found by splitting this triangle in two by dropping a perpendicular from the centre, giving

$$C = 2r \sin \frac{\Theta}{2}$$

If $\Theta \in [\pi, 2\pi]$, then

$$C = 2r \sin \frac{2\pi - \Theta}{2} = 2r \sin \frac{\Theta}{2}$$

as before. Now,

$$\begin{aligned}\mathbb{P}(C \le r) &= \mathbb{P}\left(2r \sin \frac{\Theta}{2} \le r\right) \\ &= \mathbb{P}\left(\sin \frac{\Theta}{2} \le \frac{1}{2}\right) \\ &= \mathbb{P}\left(\Theta \le \frac{\pi}{3}\right) + \mathbb{P}\left(\Theta \ge \frac{5\pi}{3}\right) \\ &= \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{3} \approx 0.333\end{aligned}$$

Clearly, the two probabilities do not match.

16 Gaussian vectors

16.1 Multidimensional Gaussian random variables

Recall that a random variable X with values in $\mathbb{R}$ is called Gaussian (or normal) if

$$X = \mu + \sigma Z; \quad \mu \in \mathbb{R}, \sigma \ge 0, Z \sim \mathrm{N}(0, 1)$$

The density function of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Now, let $X = (X_1, \dots, X_n)^\intercal$ with values in $\mathbb{R}^n$. Then we define that X is a Gaussian vector (also called Gaussian) if

$$\forall u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \in \mathbb{R}^n,\ u^\intercal X = \sum_{i=1}^{n} u_i X_i = \mu + \sigma Z$$

so any linear combination of the X_i is Gaussian. This does not require that the X_i are independent, just that their sum is always Gaussian.

Let X be Gaussian in $\mathbb{R}^n$. Suppose that A is an $m \times n$ matrix, and $b \in \mathbb{R}^m$. Then $AX + b$ is also Gaussian. Indeed, let $u \in \mathbb{R}^m$, and let $v = A^\intercal u$. Then

$$u^\intercal(AX + b) = u^\intercal AX + u^\intercal b = v^\intercal X + u^\intercal b$$

Since X is Gaussian, $v^\intercal X$ is also Gaussian. An additive constant preserves this property, so the entire expression is Gaussian.

16.2 Expectation and variance

We define the mean of a Gaussian vector X as

$$\mu = \mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}; \quad \mu_i = \mathbb{E}[X_i]$$

We further define

$$\begin{aligned} V &= \operatorname{Var}(X) = \mathbb{E}[(X-\mu)(X-\mu)^\intercal] \\ &= \begin{pmatrix} \mathbb{E}\left[(X_1-\mu_1)^2\right] & \mathbb{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathbb{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathbb{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathbb{E}\left[(X_2-\mu_2)^2\right] & \cdots & \mathbb{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathbb{E}[(X_n-\mu_1)(X_n-\mu_2)] & \cdots & \mathbb{E}\left[(X_n-\mu_n)^2\right] \end{pmatrix} \end{aligned}$$

Hence the components of V are

$$V_{ij} = \operatorname{Cov}(X_i, X_j)$$

In particular, V is a symmetric matrix, and

$$\mathbb{E}[u^\intercal X] = \mathbb{E}\left[\sum_{i=1}^{n} u_i X_i\right] = \sum_{i=1}^{n} u_i \mu_i = u^\intercal \mu$$

and

$$\operatorname{Var}(u^\intercal X) = \operatorname{Var}\left(\sum_{i=1}^{n} u_i X_i\right) = \sum_{i,j=1}^{n} u_i \operatorname{Cov}(X_i, X_j)\, u_j = u^\intercal V u$$

Hence $u^\intercal X \sim \mathrm{N}(u^\intercal \mu, u^\intercal V u)$. Further, V is a non-negative definite matrix. Indeed, let $u \in \mathbb{R}^n$. Then $\operatorname{Var}(u^\intercal X) = u^\intercal V u$. Since $\operatorname{Var}(u^\intercal X) \geq 0$, we have $u^\intercal V u \geq 0$.

16.3 Moment generating function

We define the moment generating function of X by

$$m(\lambda) = \mathbb{E}\left[e^{\lambda^\intercal X}\right]$$

where $\lambda \in \mathbb{R}^n$. Then, we know that $\lambda^\intercal X \sim \mathrm{N}(\lambda^\intercal \mu, \lambda^\intercal V \lambda)$. Hence $m(\lambda)$ is the moment generating function of a normal random variable with the above mean and variance, applied to the parameter $\theta = 1$.

$$m(\lambda) = \exp\left(\lambda^\intercal \mu + \frac{\lambda^\intercal V \lambda}{2}\right)$$

Since the moment generating function uniquely characterises the distribution, it is clear that a Gaussian vector is uniquely characterised by its mean vector μ and variance matrix V. In this case, we write $X \sim \mathrm{N}(\mu, V)$.

16.4 Constructing Gaussian vectors

Given a μ and a V matrix, we might like to create a Gaussian vector that has this mean and variance. Let $Z_1, \dots, Z_n$ be a list of independent and identically distributed standard normal random variables. Let $Z = (Z_1, \dots, Z_n)^\intercal$. Then Z is a Gaussian vector.

Proof. For any vector $u \in \mathbb{R}^n$, we have

$$u^\intercal Z = \sum_{i=1}^{n} u_i Z_i$$

Because the Z_i are independent, it is easy to take the moment generating function to get

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} u_i z_i\right)\right] &= \mathbb{E}\left[\prod_{i=1}^{n} \exp(\lambda u_i Z_i)\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda u_i Z_i)\right] \\
&= \prod_{i=1}^{n} \exp\left(\frac{(\lambda u_i)^2}{2}\right) \\
&= \exp\left(\frac{\lambda^2 |u|^2}{2}\right)
\end{aligned}$$

So $u^\intercal Z \sim \mathrm{N}(0, |u|^2)$, which is normal as required. □

Now, $\mathbb{E}[Z] = 0$, and $\mathrm{Var}(Z) = I$, the identity matrix. We then write $Z \sim \mathrm{N}(0, I)$. Now, let $\mu \in \mathbb{R}^n$, and V be a non-negative definite matrix. We want to construct a Gaussian vector X such that its mean is μ and its expectation is V, by using Z. In the one-dimensional case, this is easy, since μ is a single value, and V contains only one element, σ^2. In this case therefore, $Z \sim \mathrm{N}(0, 1)$ so $\mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$. In the general case, since V is non-negative definite, we can write

$$V = U^\intercal D U$$

where $U^{-1} = U^\intercal$, and D is a diagonal matrix with diagonal entries $\lambda_i \geq 0$. We define the square root of the matrix V to be

$$\sigma = U^\intercal \sqrt{D} U$$

where $\sqrt{D}$ is the diagonal matrix with diagonal entries $\sqrt{\lambda_i}$. Then clearly,

$$\sigma^2 = U^\intercal \sqrt{D} U U^\intercal \sqrt{D} U = U^\intercal \sqrt{D} \sqrt{D} U = U^\intercal D U = V$$

Now, let $X = \mu + \sigma Z$. We now want to show that $X \sim \mathrm{N}(\mu, V)$.

Proof. X is certainly Gaussian, since it is generated by a linear multiple of the Gaussian vector Z, with an added constant. By linearity,

$$\mathbb{E}[X] = \mu$$

and

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[(X-\mu)(X-\mu)^\intercal] \\
&= \mathbb{E}[(\sigma Z)(\sigma Z)^\intercal] \\
&= \mathbb{E}[\sigma Z Z^\intercal \sigma^\intercal] \\
&= \sigma \mathbb{E}[Z Z^\intercal] \sigma^\intercal \\
&= \sigma \sigma^\intercal \\
&= \sigma \sigma \\
&= V
\end{aligned}$$

□

16.5 Density

We can calculate the density of such a Gaussian vector $X \sim \mathrm{N}(\mu, V)$. First, consider the case where V is positive definite. Recall that in the one-dimensional case,

$$f_X(x) = f_Z(z)|J|; \quad x = \mu + \sigma z$$

In general, since V is positive definite, σ is invertible. So $x = \mu + \sigma z$ gives $z = \sigma^{-1}(x - \mu)$. Hence,

$$\begin{aligned}
f_X(x) &= f_Z(z)|J| \\
&= \prod_{i=1}^{n} \frac{\exp\left(-\frac{z_i^2}{2}\right)}{\sqrt{2\pi}} |\det \sigma^{-1}| \\
&= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{|z|^2}{2}\right) \cdot \frac{1}{\sqrt{\det V}} \\
&= \frac{1}{\sqrt{(2\pi)^n \det V}} \exp\left(-\frac{z^\intercal z}{2}\right)
\end{aligned}$$

Now,

$$\begin{aligned}
z^\intercal z &= (\sigma^{-1}(x-\mu))^\intercal (\sigma^{-1}(x-\mu)) \\
&= (x-\mu)^\intercal (\sigma^{-1})^\intercal \sigma^{-1} (x-\mu) \\
&= (x-\mu)^\intercal \sigma^{-2} (x-\mu) \\
&= (x-\mu)^\intercal V^{-1} (x-\mu)
\end{aligned}$$

Hence,

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det V}} \exp\left(-\frac{(x-\mu)^\intercal V^{-1}(x-\mu)}{2}\right)$$

In the case where V is just non-negative definite (so it could have some zero eigenvalues), we can make an orthogonal change of basis, and assume that

$$V = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix}; \quad \mu = \begin{pmatrix} \lambda \\ \nu \end{pmatrix}$$

where U is an $m \times m$ positive definite matrix, where $m < n$, and where $\lambda \in \mathbb{R}^m$, $\nu \in \mathbb{R}^{n-m}$. For U, we can then apply the result above. We can write

$$X = \begin{pmatrix} Y \\ \nu \end{pmatrix}$$

where Y has density

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \det U}} \exp\left(-\frac{(y-\lambda)^\intercal U^{-1}(y-\lambda)}{2}\right)$$

16.6 Diagonal variance

Note that if a Gaussian vector $X = (X_1, \dots, X_n)$ is comprised of independent normal random variables, then V is a diagonal matrix. Indeed, since the X_i are independent then $\operatorname{Cov}(X_i, X_j) = 0$ for all $i \neq j$, so V is diagonal.

Lemma. If V is diagonal, then the X_i are independent.

Note that zero covariance does not in general imply independence, as we saw earlier in the course, but in this specific case with Gaussian variables, this is true.

Proof. Since V is diagonal with diagonal entries λ_i, we have

$$(x-\mu)^\intercal V^{-1}(x-\mu) = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\lambda_i}$$

Hence,

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det V}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{2\lambda_i}\right)$$

So f_X factorises into a product. Hence the X_i are independent. □

We can construct an alternative proof using moment generating functions.

Proof.

$$\begin{aligned} m(\theta) &= \mathbb{E}\left[e^{\theta^\intercal X}\right] \\ &= \exp\left(\theta^\intercal \mu + \frac{\theta^\intercal V \theta}{2}\right) \\ &= \exp\left(\sum_{i=1}^{n} \theta_i \mu_i + \frac{1}{2}\sum_{i=1}^{n} \theta_i^2 \lambda_i\right) \end{aligned}$$

Hence $m(\theta)$ factorises into the moment generating functions of Gaussian random variables in $\mathbb{R}$. $\square$

In summary, for Gaussian vectors, we have $(X_1, \dots, X_n)$ independent if and only if V is diagonal.

16.7 Bivariate Gaussian vectors

A bivariate Gaussian is a Gaussian vector of two variables ($n = 2$). Let $X = (X_1, X_2)$. Let $\mu_k = \mathbb{E}[X_k]$ and $\sigma_k^2 = \operatorname{Var}(X_k)$. We further define the *correlation*

$$\rho = \operatorname{Corr}(X_1, X_2) = \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}}$$

Note that due to the Cauchy–Schwarz inequality, we have $\rho \in [-1, 1]$. We can write the variance matrix as

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

This matrix V is non-negative definite. Indeed, let $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$, then

$$\begin{aligned} u^\intercal V u &= (1-\rho)(\sigma_1^2 u_1^2 + \sigma_2^2 u_2^2) + \rho(\sigma_1 u_1 + \sigma_2 u_2)^2 \\ &= (1+\rho)(\sigma_1^2 u_1^2 + \sigma_2^2 u_2^2) - \rho(\sigma_1 u_1 - \sigma_2 u_2)^2 \end{aligned}$$

Since $\rho \in [-1, 1]$, this is non-negative for all choices of ρ.

16.8 Density of bivariate Gaussian

When $\rho = 0$ and $\sigma_1, \sigma_2 > 0$, we have

$$f_{X_1,X_2}(x_1, x_2) = \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_k - \mu_k)^2}{2\sigma_k^2}\right)$$

So X_1 and X_2 are independent in this case.

16.9 Conditional expectation

Let (X_1, X_2) be a bivariate Gaussian vector. Then let $a \in \mathbb{R}$, and consider $X_2 - aX_1$. We have

$$\operatorname{Cov}(X_2 - aX_1, X_1) = \operatorname{Cov}(X_2, X_1) - a\operatorname{Cov}(X_1, X_1) = \operatorname{Cov}(X_2, X_1) - a\operatorname{Var}(X_1) = \rho\sigma_1\sigma_2 - a\sigma_1^2$$

Now, let $a = \frac{\rho\sigma_2}{\sigma_1}$, so $\operatorname{Cov}(X_2 - aX_1, X_1) = 0$. Since $Y = X_2 - aX_1$ is Gaussian, (X_1, Y) is a Gaussian vector, and so Y and X_1 are independent. Now, we can find

$$\begin{aligned} \mathbb{E}[X_2 \mid X_1] &= \mathbb{E}[Y + aX_1 \mid X_1] \\ &= \mathbb{E}[Y] + a\mathbb{E}[X_1 \mid X_1] \\ &= \mathbb{E}[X_2 - aX_1] + aX_1 \end{aligned}$$

In particular, since $X_2 = (X_2 - aX_1) + aX_1$, we can say that given X_1,

$$X_2 \sim \mathrm{N}(\mu_2 - a\mu_1 + aX_1, \operatorname{Var}(X_2 - aX_1))$$

and

$$\operatorname{Var}(X_2 - aX_1) = \operatorname{Var}(X_2) + a^2\operatorname{Var}(X_1) - 2a\operatorname{Cov}(X_1, X_2)$$

16.10 Multivariate central limit theorem

This subsection is non-examinable, but included for completeness. Let X be a random vector in $\mathbb{R}^k$ with $\mu = \mathbb{E}[X]$ and covariance matrix Σ. Let $X_1, X_2, \ldots$ be independent and identically distributed with the same distribution as X. Then

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \xrightarrow{d} \mathrm{N}(\mu, \Sigma)$$

Convergence in distribution here means that for all reasonable $B \subseteq \mathbb{R}^k$, we have

$$\mathbb{P}(S_n \in B) \to \mathbb{P}(\mathrm{N}(\mu, \Sigma) \in B)$$

17 Simulation of random variables

17.1 Sampling from uniform distribution

It is easy for a computer to generate a random number in the interval $[0, 1)$.

We can use this as a source of randomness to simulate a random variable with an arbitrary density. Let $U \sim \mathrm{U}[0, 1]$, then let $X = -\log U$. Then

$$\mathbb{P}(X \leq x) = \mathbb{P}(\log U \leq x) = \mathbb{P}(U \geq e^{-x}) = 1 - e^{-x}$$

So X is exponentially distributed with parameter 1. More generally, we have the following.

Theorem. Let X be a continuous random variable with distribution function F. Then, if $U \sim \mathrm{U}[0, 1]$, then $F^{-1}(U) \sim F$.

Proof. Set $Y = F^{-1}(U)$. Then

$$\begin{aligned} \mathbb{P}(Y \leq x) &= \mathbb{P}\left(F^{-1}(U) \leq x\right) \\ &= \mathbb{P}(U \leq F(x)) \\ &= F(x) \end{aligned}$$

□

One way of thinking of this function F^{-1} function is that it takes an input probability p, and outputs the x value such that $\mathbb{P}(X \leq x) = p$. Then, if U is uniformly distributed, we are essentially sampling a random p.

17.2 Rejection sampling

In certain cases, finding such an F^{-1} function is difficult, if not impossible, especially where this function has jumps or has a higher dimension. Here is an alternative sampling method. Suppose $A \subset [0, 1]^d$. We then define

$$f(x) = \frac{1(x \in A)}{|A|}$$

where $|A|$ is the size or volume of this set A. Let X have density function f. How can we simulate X? Let (U_n) be an independent and identically distributed sequence of d-dimensional uniform random variables, i.e.

$$U_n = (U_{k,n} : \ k \in \{1, \dots, d\}); \quad (U_{k,n}) \sim \mathrm{U}[0,1] \text{ i.i.d.}$$

Now, let

$$N = \min\{n \geq 1 : \ U_n \in A\}$$

So we keep generating random numbers until a U_n lies in A, and reject all other possibilities. We now show that $U_N \sim f$. In particular, we want to show that for all $B \subseteq [0,1]^d$,

$$\mathbb{P}(U_n \in B) = \int_B f(x)\,\mathrm{d}x$$

We have

$$\begin{aligned}
\mathbb{P}(U_n \in B) &= \sum_{n=1}^{\infty} \mathbb{P}(U_N \in B, N = n) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(U_n \in A \cap B, U_{n-1} \notin A, \dots, U_1 \notin A) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(U_n \in A \cap B)\,\mathbb{P}(U_{n-1} \notin A) \cdots \mathbb{P}(U_1 \notin A) \\
&= \sum_{n=1}^{\infty} |A \cap B|(1 - |A|)^{n-1} \\
&= \frac{|A \cap B|}{|A|} \\
&= \int_A \frac{1(x \in B)}{|A|}\,\mathrm{d}x \\
&= \int_B f(x)\,\mathrm{d}x
\end{aligned}$$

Now suppose that f is a density on $[0,1]^{d-1}$ which is bounded by $\lambda > 0$. We can use rejection sampling to sample a random variable X with this density. Consider the set

$$A = \left\{(x_1, \dots, x_d) \in [0,1]^d : \ x_d \leq \frac{f(x_1, \dots, x_{d-1})}{\lambda}\right\}$$

From the above, we can generate a uniform random variable $Y = (X_1, \dots, X_d)$ on A. Let $X = (X_1, \dots, X_{d-1})$, then we will show that $X \sim f$. In particular, we want to show that for all $B \subseteq [0,1]^{d-1}$,

$$\mathbb{P}(X \in B) = \int_B f(x)\,\mathrm{d}x$$

We find that

$$
\begin{aligned}
\mathbb{P}(X \in B) &= \mathbb{P}((X_1, \dots, X_{d-1}) \in B) \\
&= \mathbb{P}((X_1, \dots, X_d) \in (B \times [0,1]) \cap A) \\
&= \frac{|(B \times [0,1]) \cap A|}{|A|} \\
|(B \times [0,1]) \cap A| &= \int \cdots \int 1((X_1, \dots, X_d) \in (B \times [0,1]) \cap A) \, \mathrm{d}x_1 \dots \mathrm{d}x_d \\
&= \int \cdots \int 1((X_1, \dots, X_{d-1}) \in B) \cdot 1\left(x_d \leq \frac{f(x_1, \dots, x_{d-1})}{\lambda}\right) \mathrm{d}x_1 \dots \mathrm{d}x_d \\
&= \int \cdots \int 1((X_1, \dots, X_{d-1}) \in B) \cdot \frac{f(x_1, \dots, x_{d-1})}{\lambda} \, \mathrm{d}x_1 \dots \mathrm{d}x_{d-1} \\
&= \frac{1}{\lambda} \int \cdots \int 1((X_1, \dots, X_{d-1}) \in B) \cdot f(x_1, \dots, x_{d-1}) \, \mathrm{d}x_1 \dots \mathrm{d}x_{d-1} \\
&= \frac{1}{\lambda} \int_B f(x) \, \mathrm{d}x \\
|A| &= \frac{1}{\lambda} \int_{[0,1]^{d-1}} f(x) \, \mathrm{d}x \\
&= \frac{1}{\lambda} \\
\therefore \mathbb{P}(X \in B) &= \int_B f(x) \, \mathrm{d}x
\end{aligned}
$$