1 Introduction

2 Classical probability

2.1 Classical probability

NOTE:

– `probability.tex 1`

Classical probability

Definition (Classical probability)**.** *Classical probability* applies in a situation when there are a finite number of equally likely outcome.

– `1.1 Classical probability`

– `1 Classical probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 2`

Example. A and B play a game in which they keep throwing coins. If a head lands, then A gets a point. Otherwise, B gets a point. The first person to get 10 points wins a prize. Now suppose A has got 8 points and B has got 7, but the game has to end because an earthquake struck. How should they divide the prize? We answer this by finding the probability of A winning. Someone must have won by the end of 19 rounds, i.e. after 4 more rounds. If A wins at least 2 of them, then A wins. Otherwise, B wins. The number of ways this can happen is $\binom{4}{2}+\binom{4}{3}+\binom{4}{4}=11$, while there are 16 possible outcomes in total. So A should get $11/16$ of the prize.

Example. A and B play a game in which they keep throwing coins. If a head lands, then A gets a point. Otherwise, B gets a point. The first person to get 10 points wins a prize. Now suppose A has got 8 points and B has got 7, but the game has to end because an earthquake struck. How should they divide the prize? We answer this by finding the probability of A winning. Someone must have won by the end of 19 rounds, i.e. after 4 more rounds. If A wins at least 2 of them, then A wins. Otherwise, B wins. The number of ways this can happen is $\binom{4}{2}+\binom{4}{3}+\binom{4}{4}=11$, while there are 16 possible outcomes in total. So A should get $11/16$ of the prize.

– `1.1 Classical probability`

– `1 Classical probability`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 3`

Sample space

Definition (Sample space)**.** The set of all possible outcomes is the *sample space*, Ω. We can lists the outcomes as $\omega_1, \omega_2, \cdots \in \Omega$. Each $\omega \in \Omega$ is an *outcome*.

– `1.1 Classical probability`

– `1 Classical probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 4`

Event

Definition (Event)**.** A subset of Ω is called an *event*.

– `1.1 Classical probability`

– `1 Classical probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 5`

Example. When rolling a dice, the sample space is $\{1,2,3,4,5,6\}$, and each item is an outcome. "Getting an odd number" and "getting 3" are two possible events.

Example. When rolling a dice, the sample space is $\{1,2,3,4,5,6\}$, and each item is an outcome. "Getting an odd number" and "getting 3" are two possible events.

– `1.1 Classical probability`

– `1 Classical probability`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 6`

Set notations

Definition (Set notations)**.** Given any two events $A, B \subseteq \Omega$,

– The *complement* of A is $A^C = A' = \bar{A} = \Omega \setminus A$.

– "A or B" is the set $A \cup B$.

– "A and B" is the set $A \cap B$.

– A and B are *mutually exclusive* or *disjoint* if $A \cap B = \emptyset$.

– If $A \subseteq B$, then A occurring implies B occurring.

– `1.1 Classical probability`

– `1 Classical probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 7`

Probability

Definition (Probability)**.** Suppose $\Omega = \{\omega_1, \omega_2, \cdots, \omega_N\}$. Let $A \subseteq \Omega$ be an event. Then the *probability* of A is

$$\mathbb{P}(A) = \frac{\text{Number of outcomes in } A}{\text{Number of outcomes in } \Omega} = \frac{|A|}{N}.$$

– `1.1 Classical probability`

– `1 Classical probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 8`

Example. Suppose r digits are drawn at random from a table of random digits from 0 to 9. What is the probability that

(i) No digit exceeds k;

(ii) The largest digit drawn is k?

The sample space is $\Omega = \{(a_1, a_2, \cdots, a_r) : 0 \le a_i \le 9\}$. Then $|\Omega| = 10^r$. Let $A_k = [\text{no digit exceeds } k] = \{(a_1, \cdots, a_r) : 0 \le a_i \le k\}$. Then $|A_k| = (k+1)^r$. So

$$P(A_k) = \frac{(k+1)^r}{10^r}.$$

Now let $B_k = [\text{largest digit drawn is } k]$. We can find this by finding all outcomes in which no digits exceed k, and subtract it by the number of outcomes in which no digit exceeds $k-1$. So $|B_k| = |A_k| - |A_{k-1}|$ and

$$P(B_k) = \frac{(k+1)^r - k^r}{10^r}.$$

Example. Suppose r digits are drawn at random from a table of random digits from 0 to 9. What is the probability that

(i) No digit exceeds k;

(ii) The largest digit drawn is k?

The sample space is $\Omega = \{(a_1, a_2, \cdots, a_r) : 0 \le a_i \le 9\}$. Then $|\Omega| = 10^r$. Let $A_k = [\text{no digit exceeds } k] = \{(a_1, \cdots, a_r) : 0 \le a_i \le k\}$. Then $|A_k| = (k+1)^r$. So

$$P(A_k) = \frac{(k+1)^r}{10^r}.$$

Now let $B_k = [\text{largest digit drawn is } k]$. We can find this by finding all outcomes in which no digits exceed k, and subtract it by the number of outcomes in which no digit exceeds $k-1$. So $|B_k| = |A_k| - |A_{k-1}|$ and

$$P(B_k) = \frac{(k+1)^r - k^r}{10^r}.$$

– `1.1 Classical probability`

– `1 Classical probability`

–

– `GENERAL KNOWLEDGE`

2.2 Counting

NOTE:

- probability.tex 9

Example. A menu has 6 starters, 7 mains and 6 desserts. How many possible meals combinations are there? Clearly $6 \times 7 \times 6 = 252$.

Example. A menu has 6 starters, 7 mains and 6 desserts. How many possible meals combinations are there? Clearly $6 \times 7 \times 6 = 252$.

- 1.2 Counting
- 1 Classical probability
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 10

Fundamental rule of counting

Theorem (Fundamental rule of counting)**.** Suppose we have to make r multiple choices in sequence. There are m_1 possibilities for the first choice, m_2 possibilities for the second etc. Then the total number of choices is $m_1 \times m_2 \times \cdots m_r$.

- 1.2 Counting
- 1 Classical probability
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 11

Example. How many ways can $1, 2, \cdots, n$ be ordered? The first choice has n possibilities, the second has $n-1$ possibilities etc. So there are $n \times (n-1) \times \cdots \times 1 = n!$.

Example. How many ways can $1, 2, \cdots, n$ be ordered? The first choice has n possibilities, the second has $n-1$ possibilities etc. So there are $n \times (n-1) \times \cdots \times 1 = n!$.

- 1.2 Counting
- 1 Classical probability
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 12

Sampling with replacement

Definition (Sampling with replacement)**.** When we *sample with replacement*, after choosing at item, it is put back and can be chosen again. Then *any* sampling function f satisfies sampling with replacement.

- 1.2 Counting
- 1 Classical probability
-
- VOCABULARY

NOTE:

- probability.tex 13

Sampling without replacement

Definition (Sampling without replacement)**.** When we *sample without replacement*, after choosing an item, we kill it with fire and cannot choose it again. Then f must be an injective function, and clearly we must have $x \geq n$.

- 1.2 Counting
- 1 Classical probability
-
- VOCABULARY

NOTE:

- probability.tex 14

Example. Suppose $N = \{a, b, c\}$ and $X = \{p, q, r, s\}$. How many injective functions are there $N \to X$? When we choose $f(a)$, we have 4 options. When we choose $f(b)$, we have 3 left. When we choose $f(c)$, we have 2 choices left. So there are 24 possible choices.

Example. Suppose $N = \{a, b, c\}$ and $X = \{p, q, r, s\}$. How many injective functions are there $N \to X$? When we choose $f(a)$, we have 4 options. When we choose $f(b)$, we have 3 left. When we choose $f(c)$, we have 2 choices left. So there are 24 possible choices.

- 1.2 Counting
- 1 Classical probability
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 15

Example. I have n keys in my pocket. We select one at random once and try to unlock. What is the possibility that I succeed at the rth trial? Suppose we do it with replacement. We have to fail the first $r-1$ trials and succeed in the rth. So the probability is

$$\frac{(n-1)(n-1)\cdots(n-1)(1)}{n^r} = \frac{(n-1)^{r-1}}{n^r}.$$

Now suppose we are smarter and try without replacement. Then the probability is

$$\frac{(n-1)(n-2)\cdots(n-r+1)(1)}{n(n-1)\cdots(n-r+1)} = \frac{1}{n}.$$

Example. I have n keys in my pocket. We select one at random once and try to unlock. What is the possibility that I succeed at the rth trial? Suppose we do it with replacement. We have to fail the first $r-1$ trials and succeed in the rth. So the probability is

$$\frac{(n-1)(n-1)\cdots(n-1)(1)}{n^r} = \frac{(n-1)^{r-1}}{n^r}.$$

Now suppose we are smarter and try without replacement. Then the probability is

$$\frac{(n-1)(n-2)\cdots(n-r+1)(1)}{n(n-1)\cdots(n-r+1)} = \frac{1}{n}.$$

- 1.2 Counting
- 1 Classical probability
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 16

Birthday problem

Example (Birthday problem)**.** How many people are needed in a room for there to be a probability that two people have the same birthday to be at least a half? Suppose $f(r)$ is the probability that, in a room of r people, there is a birthday match. We solve this by finding the probability of no match, $1 - f(r)$. The total number of possibilities of birthday combinations is 365^r. For nobody to have the same birthday, the first person can have any birthday. The second has 364 else to choose, etc. So

$$\mathbb{P}(\text{no match}) = \frac{365 \cdot 364 \cdot 363 \cdots (366-r)}{365 \cdot 365 \cdot 365 \cdots 365}.$$

If we calculate this with a computer, we find that $f(22) = 0.475695$ and $f(23) = 0.507297$. While this might sound odd since 23 is small, this is because we are thinking about the wrong thing. The probability of match is related more to the number of *pairs* of people, not the number of people. With 23 people, we have $23 \times 22/2 = 253$ pairs, which is quite large compared to 365.

- 1.2 Counting
- 1 Classical probability
-
- GENERAL KNOWLEDGE

NOTE:

– `probability.tex 17`

Multinomial coefficient

Definition (Multinomial coefficient)**.** A *multinomial coefficient* is

$$\binom{n}{n_1, n_2, \cdots, n_k} = \binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1\cdots-n_{k-1}}{n_k} = \frac{n!}{n_1!n_2!\cdots n_k!}.$$

It is the number of ways to distribute n items into k positions, in which the ith position has n_i items.

– `1.2 Counting`

– `1 Classical probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 18`

Example. We know that

$$(x+y)^n = x^n + \binom{n}{1}x^{n-1}y + \cdots + y^n.$$

If we have a trinomial, then

$$(x+y+z)^n = \sum_{n_1,n_2,n_3}\binom{n}{n_1,n_2,n_3}x^{n_1}y^{n_2}z^{n_3}.$$

Example. We know that

$$(x+y)^n = x^n + \binom{n}{1}x^{n-1}y + \cdots + y^n.$$

If we have a trinomial, then

$$(x+y+z)^n = \sum_{n_1,n_2,n_3}\binom{n}{n_1,n_2,n_3}x^{n_1}y^{n_2}z^{n_3}.$$

– `1.2 Counting`

– `1 Classical probability`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 19`

Example. How many ways can we deal 52 cards to 4 player, each with a hand of 13? The total number of ways is

$$\binom{52}{13,13,13,13} = \frac{52!}{(13!)^4} = 53644737765488792839237440000 = 5.36\times 10^{28}.$$

Example. How many ways can we deal 52 cards to 4 player, each with a hand of 13? The total number of ways is

$$\binom{52}{13,13,13,13} = \frac{52!}{(13!)^4} = 53644737765488792839237440000 = 5.36\times 10^{28}.$$

– `1.2 Counting`

– `1 Classical probability`

–

– `GENERAL KNOWLEDGE`

2.3 Stirling's formula

NOTE:

– `probability.tex 20`

Proposition. $\log n! \sim n\log n$

Proposition. $\log n! \sim n\log n$

– `1.3 Stirling's formula`

– `1 Classical probability`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 21`

Proposition. $\log n! \sim n\log n$

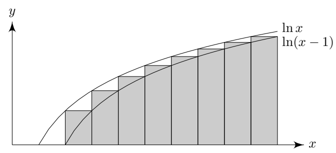
Proof. Note that

$$\log n! = \sum_{k=1}^{n}\log k.$$

Now we claim that

$$\int_1^n \log x\,\mathrm{d}x \le \sum_1^n \log k \le \int_1^{n+1}\log x\,\mathrm{d}x.$$

This is true by considering the diagram:



We actually evaluate the integral to obtain

$$n\log n - n + 1 \le \log n! \le (n+1)\log(n+1) - n;$$

Divide both sides by $n\log n$ and let $n\to\infty$. Both sides tend to 1. So

$$\frac{\log n!}{n\log n}\to 1.$$ □

– `1.3 Stirling's formula`

– `1 Classical probability`

–

– `PROOF EXERCISE`

NOTE:

– `probability.tex 22`

Stirling's formula

Theorem (Stirling's formula)**.** As $n\to\infty$,

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

– `1.3 Stirling's formula`

– `1 Classical probability`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 23`

Corollary.

$$n! \sim \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}$$

Corollary.

$$n! \sim \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}$$

– `1.3 Stirling's formula`

– `1 Classical probability`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 24`

Corollary.

$$n! \sim \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}$$

Proof. (non-examinable) Define

$$d_n = \log\left(\frac{n!e^n}{n^{n+1/2}}\right) = \log n! - (n+1/2)\log n + n$$

Then

$$d_n - d_{n+1} = (n+1/2)\log\left(\frac{n+1}{n}\right) - 1.$$

Write $t = 1/(2n+1)$. Then

$$d_n - d_{n+1} = \frac{1}{2t}\log\left(\frac{1+t}{1-t}\right) - 1.$$

We can simplifying by noting that

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

Then if we subtract the equations and divide by $2t$, we obtain

$$\begin{aligned}d_n - d_{n+1} &= \frac{1}{3}t^2 + \frac{1}{5}t^4 + \frac{1}{7}t^6 + \cdots\\ &< \frac{1}{3}t^2 + \frac{1}{3}t^4 + \frac{1}{3}t^6 + \cdots\\ &= \frac{1}{3}\frac{t^2}{1-t^2}\\ &= \frac{1}{3}\frac{1}{(2n+1)^2-1}\\ &= \frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right)\end{aligned}$$

By summing these bounds, we know that

$$d_1 - d_n < \frac{1}{12}\left(1 - \frac{1}{n}\right)$$

Then we know that d_n is bounded below by d_1+ something, and is decreasing since $d_n - d_{n+1}$ is positive. So it converges to a limit A. We know A is a lower bound for d_n since (d_n) is decreasing. Suppose $m > n$. Then $d_n - d_m < \left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{12}$. So taking the limit as $m\to\infty$, we obtain an upper bound for d_n: $d_n < A + 1/(12n)$. Hence we know that

$$A < d_n < A + \frac{1}{12n}.$$

However, all these results are useless if we don't know what A is. To find A, we have a small detour to prove a formula: Take $I_n = \int_0^{\pi/2} \sin^n \theta\ \mathrm{d}\theta$. This is decreasing for increasing n as $\sin^n\theta$ gets smaller. We also know that

$$\begin{aligned}
I_n &= \int_0^{\pi/2} \sin^n \theta\ \mathrm{d}\theta\\
&= \left[-\cos\theta\sin^{n-1}\theta\right]_0^{\pi/2} + \int_0^{\pi/2}(n-1)\cos^2\theta\sin^{n-2}\theta\ \mathrm{d}\theta\\
&= 0 + \int_0^{\pi/2}(n-1)(1-\sin^2\theta)\sin^{n-2}\theta\ \mathrm{d}\theta\\
&= (n-1)(I_{n-2} - I_n)
\end{aligned}$$

So

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

We can directly evaluate the integral to obtain $I_0 = \pi/2$, $I_1 = 1$. Then

$$\begin{aligned}
I_{2n} &= \frac{1}{2}\cdot\frac{3}{4}\cdots\frac{2n-1}{2n}\pi/2 = \frac{(2n)!}{(2^n n!)^2}\frac{\pi}{2}\\
I_{2n+1} &= \frac{2}{3}\cdot\frac{4}{5}\cdots\frac{2n}{2n+1} = \frac{(2^n n!)^2}{(2n+1)!}
\end{aligned}$$

So using the fact that I_n is decreasing, we know that

$$1 \le \frac{I_{2n}}{I_{2n+1}} \le \frac{I_{2n-1}}{I_{2n+1}} = 1 + \frac{1}{2n} \to 1.$$

Using the approximation $n! \sim n^{n+1/2}e^{-n+A}$, where A is the limit we want to find, we can approximate

$$\frac{I_{2n}}{I_{2n+1}} = \pi(2n+1)\left[\frac{((2n)!)^2}{2^{4n+1}(n!)^4}\right] \sim \pi(2n+1)\frac{1}{ne^{2A}} \to \frac{2\pi}{e^{2A}}.$$

Since the last expression is equal to 1, we know that $A = \log\sqrt{2\pi}$. Hooray for magic! □

– 1.3 Stirling's formula

– 1 Classical probability

–

– PROOF EXERCISE

NOTE:

– probability.tex 25

non-examinable

Proposition (non-examinable)**.** We use the $1/12n$ term from the proof above to get a better approximation:

$$\sqrt{2\pi}n^{n+1/2}e^{-n+\frac{1}{12n+1}} \le n! \le \sqrt{2\pi}n^{n+1/2}e^{-n+\frac{1}{12n}}.$$

– 1.3 Stirling's formula

– 1 Classical probability

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 26

Example. Suppose we toss a coin $2n$ times. What is the probability of equal number of heads and tails? The probability is

$$\frac{\binom{2n}{n}}{2^{2n}} = \frac{(2n)!}{(n!)^2 2^{2n}} \sim \frac{1}{\sqrt{n\pi}}$$

Example. Suppose we toss a coin $2n$ times. What is the probability of equal number of heads and tails? The probability is

$$\frac{\binom{2n}{n}}{2^{2n}} = \frac{(2n)!}{(n!)^2 2^{2n}} \sim \frac{1}{\sqrt{n\pi}}$$

– 1.3 Stirling's formula

– 1 Classical probability

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 27

Example. Suppose we draw 26 cards from 52. What is the probability of getting 13 reds and 13 blacks? The probability is

$$\frac{\binom{26}{13}\binom{26}{13}}{\binom{52}{26}} = 0.2181.$$

Example. Suppose we draw 26 cards from 52. What is the probability of getting 13 reds and 13 blacks? The probability is

$$\frac{\binom{26}{13}\binom{26}{13}}{\binom{52}{26}} = 0.2181.$$

– 1.3 Stirling's formula

– 1 Classical probability

–

– GENERAL KNOWLEDGE

3 Axioms of probability

3.1 Axioms and definitions

NOTE:

– probability.tex 28

Definition (Probability space)**.** A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$. Ω is a set called the *sample space*, $\mathcal{F}$ is a collection of subsets of Ω, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is the *probability measure*. $\mathcal{F}$ has to satisfy the following axioms:

(i) $\emptyset, \Omega \in \mathcal{F}$.

(ii) $A \in \mathcal{F} \Rightarrow A^C \in \mathcal{F}$.

(iii) $A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^\infty A_i \in \mathcal{F}$.

And $\mathbb{P}$ has to satisfy the following *Kolmogorov axioms*:

(i) $0 \le \mathbb{P}(A) \le 1$ for all $A \in \mathcal{F}$

(ii) $\mathbb{P}(\Omega) = 1$

(iii) For any countable collection of events $A_1, A_2, \cdots$ which are disjoint, i.e. $A_i \cap A_j = \emptyset$ for all i, j, we have

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

Items in Ω are known as the *outcomes*, items in $\mathcal{F}$ are known as the *events*, and $\mathbb{P}(A)$ is the *probability* of the event A.

Definition (Probability space)**.** A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$. Ω is a set called the *sample space*, $\mathcal{F}$ is a collection of subsets of Ω, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is the *probability measure*. $\mathcal{F}$ has to satisfy the following axioms:

(i) $\emptyset, \Omega \in \mathcal{F}$.

(ii) $A \in \mathcal{F} \Rightarrow A^C \in \mathcal{F}$.

(iii) $A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^\infty A_i \in \mathcal{F}$.

And $\mathbb{P}$ has to satisfy the following *Kolmogorov axioms*:

(i) $0 \le \mathbb{P}(A) \le 1$ for all $A \in \mathcal{F}$

(ii) $\mathbb{P}(\Omega) = 1$

(iii) For any countable collection of events $A_1, A_2, \cdots$ which are disjoint, i.e. $A_i \cap A_j = \emptyset$ for all i, j, we have

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

Items in Ω are known as the *outcomes*, items in $\mathcal{F}$ are known as the *events*, and $\mathbb{P}(A)$ is the *probability* of the event A.

– 2.1 Axioms and definitions

– 2 Axioms of probability

–

– VOCABULARY

NOTE:

– probability.tex 29

Probability distribution

Definition (Probability distribution)**.** Let $\Omega = \{\omega_1, \omega_2, \cdots\}$. Choose numbers $p_1, p_2, \cdots$ such that $\sum_{i=1}^\infty p_i = 1$. Let $p(\omega_i) = p_i$. Then define

$$\mathbb{P}(A) = \sum_{\omega_i \in A} p(\omega_i).$$

This $\mathbb{P}(A)$ satisfies the above axioms, and $p_1, p_2, \cdots$ is the *probability distribution*

– 2.1 Axioms and definitions

– 2 Axioms of probability

–

– VOCABULARY

NOTE:

– probability.tex 30

Theorem.

(i) $\mathbb{P}(\emptyset) = 0$

(ii) $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$

(iii) $A \subseteq B \Rightarrow \mathbb{P}(A) \le \mathbb{P}(B)$

(iv) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Theorem.

(i) $\mathbb{P}(\emptyset) = 0$

(ii) $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$

(iii) $A \subseteq B \Rightarrow \mathbb{P}(A) \le \mathbb{P}(B)$

(iv) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

– 2.1 Axioms and definitions

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

NOTE:

- probability.tex 31

Theorem.

(i) $\mathbb{P}(\emptyset) = 0$

(ii) $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$

(iii) $A \subseteq B \Rightarrow \mathbb{P}(A) \le \mathbb{P}(B)$

(iv) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Proof.

(i) Ω and $\emptyset$ are disjoint. So $\mathbb{P}(\Omega) + \mathbb{P}(\emptyset) = \mathbb{P}(\Omega \cup \emptyset) = \mathbb{P}(\Omega)$. So $\mathbb{P}(\emptyset) = 0$.

(ii) $\mathbb{P}(A) + \mathbb{P}(A^C) = \mathbb{P}(\Omega) = 1$ since A and A^C are disjoint.

(iii) Write $B = A \cup (B \cap A^C)$. Then $P(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^C) \ge \mathbb{P}(A)$.

(iv) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^C)$. We also know that $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \cap A^C)$. Then the result follows. □

- 2.1 Axioms and definitions
- 2 Axioms of probability
-
- PROOF EXERCISE

NOTE:

- probability.tex 32

Limit of events

Definition (Limit of events)**.** A sequence of events $A_1, A_2, \cdots$ is *increasing* if $A_1 \subseteq A_2 \cdots$. Then we define the *limit* as

$$\lim_{n\to\infty} A_n = \bigcup_1^\infty A_n.$$

Similarly, if they are *decreasing*, i.e. $A_1 \supseteq A_2 \cdots$, then

$$\lim_{n\to\infty} A_n = \bigcap_1^\infty A_n.$$

- 2.1 Axioms and definitions
- 2 Axioms of probability
-
- VOCABULARY

NOTE:

- probability.tex 33

Theorem. If $A_1, A_2, \cdots$ is increasing or decreasing, then

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

Theorem. If $A_1, A_2, \cdots$ is increasing or decreasing, then

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

- 2.1 Axioms and definitions
- 2 Axioms of probability
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 34

Theorem. If $A_1, A_2, \cdots$ is increasing or decreasing, then

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

Proof. Take $B_1 = A_1$, $B_2 = A_2 \setminus A_1$. In general,

$$B_n = A_n \setminus \bigcup_1^{n-1} A_i.$$

Then

$$\bigcup_1^n B_i = \bigcup_1^n A_i, \quad \bigcup_1^\infty B_i = \bigcup_1^\infty A_i.$$

Then

$$\begin{aligned}\mathbb{P}(\lim A_n) &= \mathbb{P}\left(\bigcup_1^\infty A_i\right)\\ &= \mathbb{P}\left(\bigcup_1^\infty B_i\right)\\ &= \sum_1^\infty \mathbb{P}(B_i) \text{ (Axiom III)}\\ &= \lim_{n\to\infty}\sum_{i=1}^n \mathbb{P}(B_i)\\ &= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_1^n A_i\right)\\ &= \lim_{n\to\infty} \mathbb{P}(A_n).\end{aligned}$$

and the decreasing case is proven similarly (or we can simply apply the above to A_i^C). □

- 2.1 Axioms and definitions
- 2 Axioms of probability
-
- PROOF EXERCISE

3.2 Inequalities and formulae

NOTE:

- probability.tex 35

Boole's inequality

Theorem (Boole's inequality)**.** For any $A_1, A_2, \cdots$,

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

- 2.2 Inequalities and formulae
- 2 Axioms of probability
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 36

Theorem (Boole's inequality)**.** For any $A_1, A_2, \cdots$,

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

Proof. Our third axiom states a similar formula that only holds for disjoint sets. So we need a (not so) clever trick to make them disjoint. We define

$$\begin{aligned}B_1 &= A_1\\ B_2 &= A_2 \setminus A_1\\ B_i &= A_i \setminus \bigcup_{k=1}^{i-1} A_k.\end{aligned}$$

So we know that

$$\bigcup B_i = \bigcup A_i.$$

But the B_i are disjoint. So our Axiom (iii) gives

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

Where the last inequality follows from (iii) of the theorem above. □

- 2.2 Inequalities and formulae
- 2 Axioms of probability
-
- PROOF EXERCISE

NOTE:

- probability.tex 37

Example. Suppose we have countably infinite number of biased coins. Let $A_k = [k\text{th toss head}]$ and $\mathbb{P}(A_k) = p_k$. Suppose $\sum_1^\infty p_k < \infty$. What is the probability that there are infinitely many heads? The event "there is at least one more head after the ith coin toss" is $\bigcup_{k=i}^\infty A_k$. There are infinitely many heads if and only if there are unboundedly many coin tosses, i.e. no matter how high i is, there is still at least more more head after the ith toss. So the probability required is

$$\mathbb{P}\left(\bigcap_{i=1}^\infty \bigcup_{k=i}^\infty A_k\right) = \lim_{i\to\infty} \mathbb{P}\left(\bigcup_{k=i}^\infty A_k\right) \le \lim_{i\to\infty} \sum_{k=i}^\infty p_k = 0$$

Therefore $\mathbb{P}(\text{infinite number of heads}) = 0$.

Example. Suppose we have countably infinite number of biased coins. Let $A_k = [k\text{th toss head}]$ and $\mathbb{P}(A_k) = p_k$. Suppose $\sum_1^\infty p_k < \infty$. What is the probability that there are infinitely many heads? The event "there is at least one more head after the ith coin toss" is $\bigcup_{k=i}^\infty A_k$. There are infinitely many heads if and only if there are unboundedly many coin tosses, i.e. no matter how high i is, there is still at least more more head after the ith toss. So the probability required is

$$\mathbb{P}\left(\bigcap_{i=1}^\infty \bigcup_{k=i}^\infty A_k\right) = \lim_{i\to\infty} \mathbb{P}\left(\bigcup_{k=i}^\infty A_k\right) \le \lim_{i\to\infty} \sum_{k=i}^\infty p_k = 0$$

Therefore $\mathbb{P}(\text{infinite number of heads}) = 0$.

- 2.2 Inequalities and formulae
- 2 Axioms of probability
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 38

Example (Erdös 1947)**.** Is it possible to colour a complete n-graph (i.e. a graph of n vertices with edges between every pair of vertices) red and black such that there is no k-vertex complete subgraph with monochrome edges? Erdös said this is possible if

$$\binom{n}{k} 2^{1-\binom{k}{2}} < 1.$$

We colour edges randomly, and let $A_i = [i\text{th subgraph has monochrome edges}]$. Then the probability that at least one subgraph has monochrome edges is

$$\mathbb{P}\left(\bigcup A_i\right) \le \sum \mathbb{P}(A_i) = \binom{n}{k} 2 \cdot 2^{-\binom{k}{2}}.$$

The last expression is obtained since there are $\binom{n}{k}$ ways to choose a subgraph; a monochrome subgraph can be either red or black, thus the multiple of 2; and the probability of getting all red (or black) is $2^{-\binom{k}{2}}$. If this probability is less than 1, then there must be a way to colour them in which it is impossible to find a monochrome subgraph, or else the probability is 1. So if $\binom{n}{k} 2^{1-\binom{k}{2}} < 1$, the colouring is possible.

Example (Erdös 1947)**.** Is it possible to colour a complete n-graph (i.e. a graph of n vertices with edges between every pair of vertices) red and black such that there is no k-vertex complete subgraph with monochrome edges? Erdös said this is possible if

$$\binom{n}{k} 2^{1-\binom{k}{2}} < 1.$$

We colour edges randomly, and let $A_i = [i\text{th subgraph has monochrome edges}]$. Then the probability that at least one subgraph has monochrome edges is

$$\mathbb{P}\left(\bigcup A_i\right) \le \sum \mathbb{P}(A_i) = \binom{n}{k} 2 \cdot 2^{-\binom{k}{2}}.$$

The last expression is obtained since there are $\binom{n}{k}$ ways to choose a subgraph; a monochrome subgraph can be either red or black, thus the multiple of 2; and the probability of getting all red (or black) is $2^{-\binom{k}{2}}$. If this probability is less than 1, then there must be a way to colour them in which it is impossible to find a monochrome subgraph, or else the probability is 1. So if $\binom{n}{k} 2^{1-\binom{k}{2}} < 1$, the colouring is possible.

- 2.2 Inequalities and formulae
- 2 Axioms of probability
-
- GENERAL KNOWLEDGE

NOTE:

– probability.tex 39

Inclusion-exclusion formula

Theorem (Inclusion-exclusion formula)**.**

$$\mathbb{P}\left(\bigcup_i^n A_i\right) = \sum_1^n \mathbb{P}(A_i) - \sum_{i_1<i_2}\mathbb{P}(A_{i_1}\cap A_{j_2}) + \sum_{i_1<i_2<i_3}\mathbb{P}(A_{i_1}\cap A_{i_2}\cap A_{i_3}) - \cdots + (-1)^{n-1}\mathbb{P}(A_1\cap\cdots\cap A_n).$$

– 2.2 Inequalities and formulae

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 40

Theorem (Inclusion-exclusion formula)**.**

$$\mathbb{P}\left(\bigcup_i^n A_i\right) = \sum_1^n \mathbb{P}(A_i) - \sum_{i_1<i_2}\mathbb{P}(A_{i_1}\cap A_{j_2}) + \sum_{i_1<i_2<i_3}\mathbb{P}(A_{i_1}\cap A_{i_2}\cap A_{i_3}) - \cdots + (-1)^{n-1}\mathbb{P}(A_1\cap\cdots\cap A_n).$$

Proof. Perform induction on n. $n = 2$ is proven above. Then

$$\mathbb{P}(A_1\cup A_2\cup\cdots\cup A_n) = \mathbb{P}(A_1)+\mathbb{P}(A_2\cup\cdots\cup A_n)-\mathbb{P}\left(\bigcup_{i=2}^n (A_1\cap A_i)\right).$$

Then we can apply the induction hypothesis for $n-1$, and expand the mess. The details are very similar to that in IA Numbers and Sets. □

– 2.2 Inequalities and formulae

– 2 Axioms of probability

–

– PROOF EXERCISE

NOTE:

– probability.tex 41

Example. Let $1, 2, \cdots, n$ be randomly permuted to $\pi(1), \pi(2), \cdots, \pi(n)$. If $i \neq \pi(i)$ for all i, we say we have a *derangement*. Let $A_i = [i = \pi(i)]$. Then

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_k \mathbb{P}(A_k) - \sum_{k_1<k_2}\mathbb{P}(A_{k_1}\cap A_{k_2}) + \cdots\\ &= n\cdot\frac{1}{n} - \binom{n}{2}\frac{1}{n}\frac{1}{n-1} + \binom{n}{3}\frac{1}{n}\frac{1}{n-1}\frac{1}{n-2} + \cdots\\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!}\\ &\to e^{-1}\end{aligned}$$

So the probability of derangement is $1 - \mathbb{P}(\bigcup A_k) \approx 1 - e^{-1} \approx 0.632$.

Example. Let $1, 2, \cdots, n$ be randomly permuted to $\pi(1), \pi(2), \cdots, \pi(n)$. If $i \neq \pi(i)$ for all i, we say we have a *derangement*. Let $A_i = [i = \pi(i)]$. Then

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_k \mathbb{P}(A_k) - \sum_{k_1<k_2}\mathbb{P}(A_{k_1}\cap A_{k_2}) + \cdots\\ &= n\cdot\frac{1}{n} - \binom{n}{2}\frac{1}{n}\frac{1}{n-1} + \binom{n}{3}\frac{1}{n}\frac{1}{n-1}\frac{1}{n-2} + \cdots\\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!}\\ &\to e^{-1}\end{aligned}$$

So the probability of derangement is $1 - \mathbb{P}(\bigcup A_k) \approx 1 - e^{-1} \approx 0.632$.

– 2.2 Inequalities and formulae

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 42

Bonferroni's inequalities

Theorem (Bonferroni's inequalities)**.** For any events $A_1, A_2, \cdots, A_n$ and $1 \leq r \leq n$, if r is odd, then

$$\mathbb{P}\left(\bigcup_1^n A_i\right) \leq \sum_{i_1}\mathbb{P}(A_{i_1}) - \sum_{i_1<i_2}\mathbb{P}(A_{i_1}A_{i_2}) + \sum_{i_1<i_2<i_3}\mathbb{P}(A_{i_1}A_{i_2}A_{i_3}) + \cdots + \sum_{i_1<i_2<\cdots<i_r}\mathbb{P}(A_{i_1}A_{i_2}A_{i_3}\cdots A_{i_r}).$$

If r is even, then

$$\mathbb{P}\left(\bigcup_1^n A_i\right) \geq \sum_{i_1}\mathbb{P}(A_{i_1}) - \sum_{i_1<i_2}\mathbb{P}(A_{i_1}A_{i_2}) + \sum_{i_1<i_2<i_3}\mathbb{P}(A_{i_1}A_{i_2}A_{i_3}) + \cdots - \sum_{i_1<i_2<\cdots<i_r}\mathbb{P}(A_{i_1}A_{i_2}A_{i_3}\cdots A_{i_r}).$$

– 2.2 Inequalities and formulae

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 43

Theorem (Bonferroni's inequalities)**.** For any events $A_1, A_2, \cdots, A_n$ and $1 \leq r \leq n$, if r is odd, then

$$\mathbb{P}\left(\bigcup_1^n A_i\right) \leq \sum_{i_1}\mathbb{P}(A_{i_1}) - \sum_{i_1<i_2}\mathbb{P}(A_{i_1}A_{i_2}) + \sum_{i_1<i_2<i_3}\mathbb{P}(A_{i_1}A_{i_2}A_{i_3}) + \cdots + \sum_{i_1<i_2<\cdots<i_r}\mathbb{P}(A_{i_1}A_{i_2}A_{i_3}\cdots A_{i_r}).$$

If r is even, then

$$\mathbb{P}\left(\bigcup_1^n A_i\right) \geq \sum_{i_1}\mathbb{P}(A_{i_1}) - \sum_{i_1<i_2}\mathbb{P}(A_{i_1}A_{i_2}) + \sum_{i_1<i_2<i_3}\mathbb{P}(A_{i_1}A_{i_2}A_{i_3}) + \cdots - \sum_{i_1<i_2<\cdots<i_r}\mathbb{P}(A_{i_1}A_{i_2}A_{i_3}\cdots A_{i_r}).$$

Proof. Easy induction on n. □

– 2.2 Inequalities and formulae

– 2 Axioms of probability

–

– PROOF EXERCISE

NOTE:

– probability.tex 44

Example. Let $\Omega = \{1, 2, \cdots, m\}$ and $1 \leq j, k \leq m$. Write $A_k = \{1, 2, \cdots, k\}$. Then

$$A_k \cap A_j = \{1, 2, \cdots, \min(j, k)\} = A_{\min(j,k)}$$

and

$$A_k \cup A_j = \{1, 2, \cdots, \max(j, k)\} = A_{\max(j,k)}.$$

We also have $\mathbb{P}(A_k) = k/m$. Now let $1 \leq x_1, \cdots, x_n \leq m$ be some numbers. Then Bonferroni's inequality says

$$\mathbb{P}\left(\bigcup A_{x_i}\right) \geq \sum \mathbb{P}(A_{x_i}) - \sum_{i<j}\mathbb{P}(A_{x_i}\cap A_{x_j}).$$

So

$$\max\{x_1, x_2, \cdots, x_n\} \geq \sum x_i - \sum_{i_1<i_2}\min\{x_1, x_2\}.$$

Example. Let $\Omega = \{1, 2, \cdots, m\}$ and $1 \leq j, k \leq m$. Write $A_k = \{1, 2, \cdots, k\}$. Then

$$A_k \cap A_j = \{1, 2, \cdots, \min(j, k)\} = A_{\min(j,k)}$$

and

$$A_k \cup A_j = \{1, 2, \cdots, \max(j, k)\} = A_{\max(j,k)}.$$

We also have $\mathbb{P}(A_k) = k/m$. Now let $1 \leq x_1, \cdots, x_n \leq m$ be some numbers. Then Bonferroni's inequality says

$$\mathbb{P}\left(\bigcup A_{x_i}\right) \geq \sum \mathbb{P}(A_{x_i}) - \sum_{i<j}\mathbb{P}(A_{x_i}\cap A_{x_j}).$$

So

$$\max\{x_1, x_2, \cdots, x_n\} \geq \sum x_i - \sum_{i_1<i_2}\min\{x_1, x_2\}.$$

– 2.2 Inequalities and formulae

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

3.3 Independence

NOTE:

– probability.tex 45

Independent events

Definition (Independent events)**.** Two events A and B are *independent* if

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

Otherwise, they are said to be *dependent*.

– 2.3 Independence

– 2 Axioms of probability

–

– VOCABULARY

NOTE:

– probability.tex 46

Proposition. If A and B are independent, then A and B^C are independent.

Proposition. If A and B are independent, then A and B^C are independent.

– 2.3 Independence

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

NOTE:

– `probability.tex 47`

Proposition. If A and B are independent, then A and B^C are independent.

Proof.

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

□

– `2.3 Independence`

– `2 Axioms of probability`

–

– `PROOF EXERCISE`

NOTE:

– `probability.tex 48`

Example. Roll two fair dice. Let A_1 and A_2 be the event that the first and second die is odd respectively. Let $A_3 =$ [sum is odd]. The event probabilities are as follows:

Event	Probability
A_1	1/2
A_2	1/2
A_3	1/2
$A_1 \cap A_2$	1/4
$A_1 \cap A_3$	1/4
$A_2 \cap A_3$	1/4
$A_1 \cap A_2 \cap A_3$	0

We see that A_1 and A_2 are independent, A_1 and A_3 are independent, and A_2 and A_3 are independent. However, the collection of all three are *not* independent, since if A_1 and A_2 are true, then A_3 cannot possibly be true.

Example. Roll two fair dice. Let A_1 and A_2 be the event that the first and second die is odd respectively. Let $A_3 =$ [sum is odd]. The event probabilities are as follows:

Event	Probability
A_1	1/2
A_2	1/2
A_3	1/2
$A_1 \cap A_2$	1/4
$A_1 \cap A_3$	1/4
$A_2 \cap A_3$	1/4
$A_1 \cap A_2 \cap A_3$	0

We see that A_1 and A_2 are independent, A_1 and A_3 are independent, and A_2 and A_3 are independent. However, the collection of all three are *not* independent, since if A_1 and A_2 are true, then A_3 cannot possibly be true.

– `2.3 Independence`

– `2 Axioms of probability`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 49`

Independence of multiple events

Definition (Independence of multiple events)**.** Events $A_1, A_2, \cdots$ are said to be *mutually independent* if

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_r})$$

for any $i_1, i_2, \cdots i_r$ and $r \geq 2$.

– `2.3 Independence`

– `2 Axioms of probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 50`

Example. Let A_{ij} be the event that i and j roll the same. We roll 4 dice. Then

$$\mathbb{P}(A_{12} \cap A_{13}) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = \mathbb{P}(A_{12})\mathbb{P}(A_{13}).$$

But

$$\mathbb{P}(A_{12} \cap A_{13} \cap A_{23}) = \frac{1}{36} \neq \mathbb{P}(A_{12})\mathbb{P}(A_{13})\mathbb{P}(A_{23}).$$

So they are not mutually independent.

Example. Let A_{ij} be the event that i and j roll the same. We roll 4 dice. Then

$$\mathbb{P}(A_{12} \cap A_{13}) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = \mathbb{P}(A_{12})\mathbb{P}(A_{13}).$$

But

$$\mathbb{P}(A_{12} \cap A_{13} \cap A_{23}) = \frac{1}{36} \neq \mathbb{P}(A_{12})\mathbb{P}(A_{13})\mathbb{P}(A_{23}).$$

So they are not mutually independent.

– `2.3 Independence`

– `2 Axioms of probability`

–

– `GENERAL KNOWLEDGE`

3.4 Important discrete distributions

NOTE:

– `probability.tex 51`

Bernoulli distribution

Definition (Bernoulli distribution)**.** Suppose we toss a coin. $\Omega = \{H, T\}$ and $p \in [0, 1]$. The *Bernoulli distribution*, denoted $B(1, p)$ has

$$\mathbb{P}(H) = p; \quad \mathbb{P}(T) = 1 - p.$$

– `2.4 Important discrete distributions`

– `2 Axioms of probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 52`

Definition (Binomial distribution)**.** Suppose we toss a coin n times, each with probability p of getting heads. Then

$$\mathbb{P}(HHTT \cdots T) = pp(1-p) \cdots (1-p).$$

So

$$\mathbb{P}(\text{two heads}) = \binom{n}{2} p^2 (1-p)^{n-2}.$$

In general,

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

We call this the *binomial distribution* and write it as $B(n, p)$.

Definition (Binomial distribution)**.** Suppose we toss a coin n times, each with probability p of getting heads. Then

$$\mathbb{P}(HHTT \cdots T) = pp(1-p) \cdots (1-p).$$

So

$$\mathbb{P}(\text{two heads}) = \binom{n}{2} p^2 (1-p)^{n-2}.$$

In general,

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

We call this the *binomial distribution* and write it as $B(n, p)$.

– `2.4 Important discrete distributions`

– `2 Axioms of probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 53`

Geometric distribution

Definition (Geometric distribution)**.** Suppose we toss a coin with probability p of getting heads. The probability of having a head after k consecutive tails is

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

This is *geometric distribution*. We say it is *memoryless* because how many tails we've got in the past does not give us any information to how long I'll have to wait until I get a head.

– `2.4 Important discrete distributions`

– `2 Axioms of probability`

–

– `VOCABULARY`

NOTE:

– `probability.tex 54`

Hypergeometric distribution

Definition (Hypergeometric distribution)**.** Suppose we have an urn with n_1 red balls and n_2 black balls. We choose n balls. The probability that there are k red balls is

$$\mathbb{P}(k \text{ red}) = \frac{\binom{n_1}{k}\binom{n_2}{n-k}}{\binom{n_1+n_2}{n}}.$$

– `2.4 Important discrete distributions`

– `2 Axioms of probability`

–

– `VOCABULARY`

NOTE:

- `probability.tex 55`

Poisson distribution

Definition (Poisson distribution)**.** The *Poisson distribution* denoted $P(\lambda)$ is

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

for $k \in \mathbb{N}$.

- `2.4 Important discrete distributions`
- `2 Axioms of probability`
-
- `VOCABULARY`

NOTE:

- `probability.tex 56`

Poisson approximation to binomial

Theorem (Poisson approximation to binomial)**.** Suppose $n \to \infty$ and $p \to 0$ such that $np = \lambda$. Then

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

- `2.4 Important discrete distributions`
- `2 Axioms of probability`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 57`

Theorem (Poisson approximation to binomial)**.** Suppose $n \to \infty$ and $p \to 0$ such that $np = \lambda$. Then

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

Proof.

$$\begin{aligned} q_k &= \binom{n}{k}p^k(1-p)^{n-k} \\ &= \frac{1}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}(np)^k\left(1-\frac{np}{n}\right)^{n-k} \\ &\to \frac{1}{k!}\lambda^k e^{-\lambda} \end{aligned}$$

since $(1-a/n)^n \to e^{-a}$. □

- `2.4 Important discrete distributions`
- `2 Axioms of probability`
-
- `PROOF EXERCISE`

3.5 Conditional probability

NOTE:

- `probability.tex 58`

Conditional probability

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

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

We interpret as the probability of A happening given that B has happened.

- `2.5 Conditional probability`
- `2 Axioms of probability`
-
- `VOCABULARY`

NOTE:

- `probability.tex 59`

Example. In a game of poker, let A_i = [player i gets royal flush]. Then

$$\mathbb{P}(A_1) = 1.539 \times 10^{-6}.$$

and

$$\mathbb{P}(A_2 \mid A_1) = 1.969 \times 10^{-6}.$$

It is significantly bigger, albeit still incredibly tiny. So we say "good hands attract". If $\mathbb{P}(A \mid B) > \mathbb{P}(A)$, then we say that B attracts A. Since

$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} > \mathbb{P}(A) \Leftrightarrow \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} > \mathbb{P}(B),$$

A attracts B if and only if B attracts A. We can also say A repels B if A attracts B^C.

Example. In a game of poker, let A_i = [player i gets royal flush]. Then

$$\mathbb{P}(A_1) = 1.539 \times 10^{-6}.$$

and

$$\mathbb{P}(A_2 \mid A_1) = 1.969 \times 10^{-6}.$$

It is significantly bigger, albeit still incredibly tiny. So we say "good hands attract". If $\mathbb{P}(A \mid B) > \mathbb{P}(A)$, then we say that B attracts A. Since

$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} > \mathbb{P}(A) \Leftrightarrow \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} > \mathbb{P}(B),$$

A attracts B if and only if B attracts A. We can also say A repels B if A attracts B^C.

- `2.5 Conditional probability`
- `2 Axioms of probability`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 60`

Theorem.

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

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

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

(iv) The function $\mathbb{P}(\cdot \mid B)$ restricted to subsets of B is a probability function (or measure).

Theorem.

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

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

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

(iv) The function $\mathbb{P}(\cdot \mid B)$ restricted to subsets of B is a probability function (or measure).

- `2.5 Conditional probability`
- `2 Axioms of probability`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 61`

Theorem.

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

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

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

(iv) The function $\mathbb{P}(\cdot \mid B)$ restricted to subsets of B is a probability function (or measure).

Proof. Proofs of (i), (ii) and (iii) are trivial. So we only prove (iv). To prove this, we have to check the axioms.

(i) Let $A \subseteq B$. Then $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \leq 1$.

(ii) $\mathbb{P}(B \mid B) = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$.

(iii) Let A_i be disjoint events that are subsets of B. Then

$$\begin{aligned} \mathbb{P}\left(\bigcup_i A_i \middle| B\right) &= \frac{\mathbb{P}(\bigcup_i A_i \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(\bigcup_i A_i)}{\mathbb{P}(B)} \\ &= \sum \frac{\mathbb{P}(A_i)}{\mathbb{P}(B)} \\ &= \sum \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \\ &= \sum \mathbb{P}(A_i \mid B). \end{aligned}$$

□

- `2.5 Conditional probability`
- `2 Axioms of probability`
-
- `PROOF EXERCISE`

NOTE:

- `probability.tex 62`

Partition

Definition (Partition)**.** A *partition of the sample space* is a collection of disjoint events $\{B_i\}_{i=0}^\infty$ such that $\bigcup_i B_i = \Omega$.

- `2.5 Conditional probability`
- `2 Axioms of probability`
-
- `VOCABULARY`

NOTE:

– probability.tex 63

Proposition. If B_i is a partition of the sample space, and A is any event, then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Proposition. If B_i is a partition of the sample space, and A is any event, then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

– 2.5 Conditional probability

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 64

Example. A fair coin is tossed repeatedly. The gambler gets $+1$ for head, and -1 for tail. Continue until he is broke or achieves \$$a$. Let

$$p_x = \mathbb{P}(\text{goes broke} \mid \text{starts with } \$x),$$

and B_1 be the event that he gets head on the first toss. Then

$$p_x = \mathbb{P}(B_1)p_{x+1} + \mathbb{P}(B_1^C)p_{x-1}$$
$$p_x = \frac{1}{2}p_{x+1} + \frac{1}{2}p_{x-1}$$

We have two boundary conditions $p_0 = 1$, $p_a = 0$. Then solving the recurrence relation, we have

$$p_x = 1 - \frac{x}{a}.$$

Example. A fair coin is tossed repeatedly. The gambler gets $+1$ for head, and -1 for tail. Continue until he is broke or achieves \$$a$. Let

$$p_x = \mathbb{P}(\text{goes broke} \mid \text{starts with } \$x),$$

and B_1 be the event that he gets head on the first toss. Then

$$p_x = \mathbb{P}(B_1)p_{x+1} + \mathbb{P}(B_1^C)p_{x-1}$$
$$p_x = \frac{1}{2}p_{x+1} + \frac{1}{2}p_{x-1}$$

We have two boundary conditions $p_0 = 1$, $p_a = 0$. Then solving the recurrence relation, we have

$$p_x = 1 - \frac{x}{a}.$$

– 2.5 Conditional probability

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 65

Bayes' formula

Theorem (Bayes' formula)**.** Suppose B_i is a partition of the sample space, and A and B_i all have non-zero probability. Then for any B_i,

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}.$$

Note that the denominator is simply $\mathbb{P}(A)$ written in a fancy way.

– 2.5 Conditional probability

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 66

Example (Screen test)**.** Suppose we have a screening test that tests whether a patient has a particular disease. We denote positive and negative results as $+$ and $-$ respectively, and D denotes the person having disease. Suppose that the test is not absolutely accurate, and

$$\begin{aligned} \mathbb{P}(+ \mid D) &= 0.98 \\ \mathbb{P}(+ \mid D^C) &= 0.01 \\ \mathbb{P}(D) &= 0.001. \end{aligned}$$

So what is the probability that a person has the disease given that he received a positive result?

$$\begin{aligned} \mathbb{P}(D \mid +) &= \frac{\mathbb{P}(+ \mid D)\mathbb{P}(D)}{\mathbb{P}(+ \mid D)\mathbb{P}(D) + \mathbb{P}(+ \mid D^C)\mathbb{P}(D^C)} \\ &= \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.01 \cdot 0.999} \\ &= 0.09 \end{aligned}$$

So this test is pretty useless. Even if you get a positive result, since the disease is so rare, it is more likely that you don't have the disease and get a false positive.

Example (Screen test)**.** Suppose we have a screening test that tests whether a patient has a particular disease. We denote positive and negative results as $+$ and $-$ respectively, and D denotes the person having disease. Suppose that the test is not absolutely accurate, and

$$\begin{aligned} \mathbb{P}(+ \mid D) &= 0.98 \\ \mathbb{P}(+ \mid D^C) &= 0.01 \\ \mathbb{P}(D) &= 0.001. \end{aligned}$$

So what is the probability that a person has the disease given that he received a positive result?

$$\begin{aligned} \mathbb{P}(D \mid +) &= \frac{\mathbb{P}(+ \mid D)\mathbb{P}(D)}{\mathbb{P}(+ \mid D)\mathbb{P}(D) + \mathbb{P}(+ \mid D^C)\mathbb{P}(D^C)} \\ &= \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.01 \cdot 0.999} \\ &= 0.09 \end{aligned}$$

So this test is pretty useless. Even if you get a positive result, since the disease is so rare, it is more likely that you don't have the disease and get a false positive.

– 2.5 Conditional probability

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 67

Example. Consider the two following cases:

(i) I have 2 children, one of whom is a boy.

(ii) I have two children, one of whom is a son born on a Tuesday.

What is the probability that both of them are boys?

(i) $\mathbb{P}(BB \mid BB \cup BG) = \frac{1/4}{1/4 + 2/4} = \frac{1}{3}$.

(ii) Let B^* denote a boy born on a Tuesday, and B a boy not born on a Tuesday. Then

$$\begin{aligned} \mathbb{P}(B^*B^* \cup B^*B \mid BB^* \cup B^*B^* \cup B^*G) &= \frac{\frac{1}{14}\cdot\frac{1}{14} + 2\cdot\frac{1}{14}\cdot\frac{6}{14}}{\frac{1}{14}\cdot\frac{1}{14} + 2\cdot\frac{1}{14}\cdot\frac{6}{14} + 2\cdot\frac{1}{14}\cdot\frac{1}{2}} \\ &= \frac{13}{27}. \end{aligned}$$

How can we understand this? It is much easier to have a boy born on a Tuesday if you have two boys than one boy. So if we have the information that a boy is born on a Tuesday, it is now less likely that there is just one boy. In other words, it is more likely that there are two boys.

Example. Consider the two following cases:

(i) I have 2 children, one of whom is a boy.

(ii) I have two children, one of whom is a son born on a Tuesday.

What is the probability that both of them are boys?

(i) $\mathbb{P}(BB \mid BB \cup BG) = \frac{1/4}{1/4 + 2/4} = \frac{1}{3}$.

(ii) Let B^* denote a boy born on a Tuesday, and B a boy not born on a Tuesday. Then

$$\begin{aligned} \mathbb{P}(B^*B^* \cup B^*B \mid BB^* \cup B^*B^* \cup B^*G) &= \frac{\frac{1}{14}\cdot\frac{1}{14} + 2\cdot\frac{1}{14}\cdot\frac{6}{14}}{\frac{1}{14}\cdot\frac{1}{14} + 2\cdot\frac{1}{14}\cdot\frac{6}{14} + 2\cdot\frac{1}{14}\cdot\frac{1}{2}} \\ &= \frac{13}{27}. \end{aligned}$$

How can we understand this? It is much easier to have a boy born on a Tuesday if you have two boys than one boy. So if we have the information that a boy is born on a Tuesday, it is now less likely that there is just one boy. In other words, it is more likely that there are two boys.

– 2.5 Conditional probability

– 2 Axioms of probability

–

– GENERAL KNOWLEDGE

4 Discrete random variables

4.1 Discrete random variables

NOTE:

– probability.tex 68

Random variable

Definition (Random variable)**.** A *random variable* X taking values in a set Ω_X is a function $X : \Omega \to \Omega_X$. Ω_X is usually a set of numbers, e.g. $\mathbb{R}$ or $\mathbb{N}$.

– 3.1 Discrete random variables

– 3 Discrete random variables

–

– VOCABULARY

NOTE:

– probability.tex 69

Discrete random variables

Definition (Discrete random variables)**.** A random variable is *discrete* if Ω_X is finite or countably infinite.

– 3.1 Discrete random variables

– 3 Discrete random variables

–

– VOCABULARY

NOTE:

– probability.tex 70

Example. Let X be the value shown by rolling a fair die. Then $\Omega_X = \{1, 2, 3, 4, 5, 6\}$. We know that

$$\mathbb{P}(X = i) = \frac{1}{6}.$$

We call this the discrete uniform distribution.

Example. Let X be the value shown by rolling a fair die. Then $\Omega_X = \{1, 2, 3, 4, 5, 6\}$. We know that

$$\mathbb{P}(X = i) = \frac{1}{6}.$$

We call this the discrete uniform distribution.

– 3.1 Discrete random variables

– 3 Discrete random variables

–

– GENERAL KNOWLEDGE

NOTE:

- probability.tex 71

Discrete uniform distribution

Definition (Discrete uniform distribution)**.** A *discrete uniform distribution* is a discrete distribution with finitely many possible outcomes, in which each outcome is equally likely.

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- VOCABULARY

NOTE:

- probability.tex 72

Example. Suppose we roll two dice, and let the values obtained by X and Y. Then the sum can be represented by $X+Y$, with

$$\Omega_{X+Y} = \{2, 3, \cdots, 12\}.$$

Example. Suppose we roll two dice, and let the values obtained by X and Y. Then the sum can be represented by $X+Y$, with

$$\Omega_{X+Y} = \{2, 3, \cdots, 12\}.$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 73

Definition (Expectation)**.** The *expectation* (or *mean*) of a real-valued X is equal to

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} p_\omega X(\omega).$$

provided this is *absolutely convergent*. Otherwise, we say the expectation doesn't exist. Alternatively,

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

We are sometimes lazy and just write $\mathbb{E}X$.

Definition (Expectation)**.** The *expectation* (or *mean*) of a real-valued X is equal to

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} p_\omega X(\omega).$$

provided this is *absolutely convergent*. Otherwise, we say the expectation doesn't exist. Alternatively,

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

We are sometimes lazy and just write $\mathbb{E}X$.

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- VOCABULARY

NOTE:

- probability.tex 74

Example. Let X be the sum of the outcomes of two dice. Then

$$\mathbb{E}[X] = 2\cdot\frac{1}{36} + 3\cdot\frac{2}{36} + \cdots + 12\cdot\frac{1}{36} = 7.$$

Example. Let X be the sum of the outcomes of two dice. Then

$$\mathbb{E}[X] = 2\cdot\frac{1}{36} + 3\cdot\frac{2}{36} + \cdots + 12\cdot\frac{1}{36} = 7.$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 75

St. Petersburg paradox

Example (St. Petersburg paradox)**.** Suppose we play a game in which we keep tossing a coin until you get a tail. If you get a tail on the ith round, then I pay you $\$2^i$. The expected value is

$$\mathbb{E}[X] = \frac{1}{2}\cdot 2 + \frac{1}{4}\cdot 4 + \frac{1}{8}\cdot 8 + \cdots = \infty.$$

This means that on average, you can expect to get an infinite amount of money! In real life, though, people would hardly be willing to pay $20 to play this game. There are many ways to resolve this paradox, such as taking into account the fact that the host of the game has only finitely many money and thus your real expected gain is much smaller.

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 76

Example. We calculate the expected values of different distributions:

(i) Poisson $P(\lambda)$. Let $X\sim P(\lambda)$. Then

$$P_X(r) = \frac{\lambda^r e^{-\lambda}}{r!}.$$

So

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

(ii) Let $X\sim B(n,p)$. Then

$$\begin{aligned}\mathbb{E}[X] &= \sum_0^n rP(x=r)\\ &= \sum_0^n r\binom{n}{r}p^r(1-p)^{n-r}\\ &= \sum_0^n r\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}\\ &= np\sum_{r=1}^n\frac{(n-1)!}{(r-1)![(n-1)-(r-1)]!}p^{r-1}(1-p)^{(n-1)-(r-1)}\\ &= np\sum_0^{n-1}\binom{n-1}{r}p^r(1-p)^{n-1-r}\\ &= np.\end{aligned}$$

Example. We calculate the expected values of different distributions:

(i) Poisson $P(\lambda)$. Let $X\sim P(\lambda)$. Then

$$P_X(r) = \frac{\lambda^r e^{-\lambda}}{r!}.$$

So

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

(ii) Let $X\sim B(n,p)$. Then

$$\begin{aligned}\mathbb{E}[X] &= \sum_0^n rP(x=r)\\ &= \sum_0^n r\binom{n}{r}p^r(1-p)^{n-r}\\ &= \sum_0^n r\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}\\ &= np\sum_{r=1}^n\frac{(n-1)!}{(r-1)![(n-1)-(r-1)]!}p^{r-1}(1-p)^{(n-1)-(r-1)}\\ &= np\sum_0^{n-1}\binom{n-1}{r}p^r(1-p)^{n-1-r}\\ &= np.\end{aligned}$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 77

Example. if a, b, c are constants, then $a+bX$ and $(X-c)^2$ are random variables, defined as

$$\begin{aligned}(a+bX)(\omega) &= a+bX(\omega)\\ (X-c)^2(\omega) &= (X(\omega)-c)^2.\end{aligned}$$

Example. if a, b, c are constants, then $a+bX$ and $(X-c)^2$ are random variables, defined as

$$\begin{aligned}(a+bX)(\omega) &= a+bX(\omega)\\ (X-c)^2(\omega) &= (X(\omega)-c)^2.\end{aligned}$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- `probability.tex 78`

Theorem.

(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 a and b are constants, then $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$.

(iv) If X and Y are random variables, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. This is true even if X and Y are not independent.

(v) $\mathbb{E}[X]$ is a constant that minimizes $\mathbb{E}[(X - c)^2]$ over c.

Theorem.

(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 a and b are constants, then $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$.

(iv) If X and Y are random variables, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. This is true even if X and Y are not independent.

(v) $\mathbb{E}[X]$ is a constant that minimizes $\mathbb{E}[(X - c)^2]$ over c.

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 79`

Theorem.

(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 a and b are constants, then $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$.

(iv) If X and Y are random variables, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. This is true even if X and Y are not independent.

(v) $\mathbb{E}[X]$ is a constant that minimizes $\mathbb{E}[(X - c)^2]$ over c.

Proof.

(i) $X \geq 0$ means that $X(\omega) \geq 0$ for all ω. Then

$$\mathbb{E}[X] = \sum_{\omega} p_{\omega} X(\omega) \geq 0.$$

(ii) If there exists ω such that $X(\omega) > 0$ and $p_{\omega} > 0$, then $\mathbb{E}[X] > 0$. So $X(\omega) = 0$ for all ω.

(iii)

$$\mathbb{E}[a+bX] = \sum_{\omega}(a+bX(\omega))p_{\omega} = a+b\sum_{\omega} p_{\omega} = a+b\,\mathbb{E}[X].$$

(iv)

$$\mathbb{E}[X+Y] = \sum_{\omega} p_{\omega}[X(\omega)+Y(\omega)] = \sum_{\omega} p_{\omega}X(\omega)+\sum_{\omega} p_{\omega}Y(\omega) = \mathbb{E}[X]+\mathbb{E}[Y].$$

(v)

$$\begin{aligned}\mathbb{E}[(X-c)^2] &= \mathbb{E}[(X - \mathbb{E}[X] + \mathbb{E}[X] - c)^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2 + 2(\mathbb{E}[X] - c)(X - \mathbb{E}[X]) + (\mathbb{E}[X] - c)^2] \\ &= \mathbb{E}(X - \mathbb{E}[X])^2 + 0 + (\mathbb{E}[X] - c)^2.\end{aligned}$$

This is clearly minimized when $c = \mathbb{E}[X]$. Note that we obtained the zero in the middle because $\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$. □

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `PROOF EXERCISE`

NOTE:

- `probability.tex 80`

Theorem. For any random variables $X_1, X_2, \cdots X_n$, for which the following expectations exist,

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Theorem. For any random variables $X_1, X_2, \cdots X_n$, for which the following expectations exist,

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 81`

Theorem. For any random variables $X_1, X_2, \cdots X_n$, for which the following expectations exist,

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Proof.

$$\sum_{\omega} p(\omega)[X_1(\omega)+\cdots+X_n(\omega)] = \sum_{\omega} p(\omega)X_1(\omega)+\cdots+\sum_{\omega} p(\omega)X_n($$

□

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `PROOF EXERCISE`

NOTE:

- `probability.tex 82`

Variance and standard deviation

Definition (Variance and standard deviation)**.** The *variance* of a random variable X is defined as

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

The *standard deviation* is the square root of the variance, $\sqrt{\mathrm{var}(X)}$.

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `VOCABULARY`

NOTE:

- `probability.tex 83`

Theorem.

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

(ii) $\mathrm{var}(a + bX) = b^2 \mathrm{var}(X)$. This can be proved by expanding the definition and using the linearity of the expected value.

(iii) $\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, also proven by expanding the definition.

Theorem.

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

(ii) $\mathrm{var}(a + bX) = b^2 \mathrm{var}(X)$. This can be proved by expanding the definition and using the linearity of the expected value.

(iii) $\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, also proven by expanding the definition.

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 84`

Example (Binomial distribution)**.** Let $X \sim B(n, p)$ be a binomial distribution. Then $\mathbb{E}[X] = np$. We also have

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

The sum goes to 1 since it is the sum of all probabilities of a binomial $N(n-2, p)$ So $\mathbb{E}[X^2] = n(n-1)p^2 + \mathbb{E}[X] = n(n-1)p^2 + np$. So

$$\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = np(1-p) = npq.$$

Example (Binomial distribution)**.** Let $X \sim B(n, p)$ be a binomial distribution. Then $\mathbb{E}[X] = np$. We also have

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

The sum goes to 1 since it is the sum of all probabilities of a binomial $N(n-2, p)$ So $\mathbb{E}[X^2] = n(n-1)p^2 + \mathbb{E}[X] = n(n-1)p^2 + np$. So

$$\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = np(1-p) = npq.$$

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 85`

Poisson distribution

Example (Poisson distribution)**.** If $X \sim P(\lambda)$, then $\mathbb{E}[X] = \lambda$, and $\mathrm{var}(X) = \lambda$, since $P(\lambda)$ is $B(n, p)$ with $n \to \infty, p \to 0, np \to \lambda$.

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- probability.tex 86

Example (Geometric distribution)**.** Suppose $\mathbb{P}(X = r) = q^r p$ for $r = 0, 1, 2, \cdots$. Then

$$\begin{aligned}
\mathbb{E}[X] &= \sum_0^\infty rpq^r \\
&= pq\sum_0^\infty rq^{r-1} \\
&= pq\sum_0^\infty \frac{\mathrm{d}}{\mathrm{d}q}q^r \\
&= pq\frac{\mathrm{d}}{\mathrm{d}q}\sum_0^\infty q^r \\
&= pq\frac{\mathrm{d}}{\mathrm{d}q}\frac{1}{1-q} \\
&= \frac{pq}{(1-q)^2} \\
&= \frac{q}{p}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_0^\infty r(r-1)pq^r \\
&= pq^2\sum_0^\infty r(r-1)q^{r-2} \\
&= pq^2\frac{\mathrm{d}^2}{\mathrm{d}q^2}\frac{1}{1-q} \\
&= \frac{2pq^2}{(1-q)^3}
\end{aligned}$$

So the variance is

$$\mathrm{var}(X) = \frac{2pq^2}{(1-q)^3} + \frac{q}{p} - \frac{q^2}{p^2} = \frac{q}{p^2}.$$

Example (Geometric distribution)**.** Suppose $\mathbb{P}(X = r) = q^r p$ for $r = 0, 1, 2, \cdots$. Then

$$\begin{aligned}
\mathbb{E}[X] &= \sum_0^\infty rpq^r \\
&= pq\sum_0^\infty rq^{r-1} \\
&= pq\sum_0^\infty \frac{\mathrm{d}}{\mathrm{d}q}q^r \\
&= pq\frac{\mathrm{d}}{\mathrm{d}q}\sum_0^\infty q^r \\
&= pq\frac{\mathrm{d}}{\mathrm{d}q}\frac{1}{1-q} \\
&= \frac{pq}{(1-q)^2} \\
&= \frac{q}{p}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_0^\infty r(r-1)pq^r \\
&= pq^2\sum_0^\infty r(r-1)q^{r-2} \\
&= pq^2\frac{\mathrm{d}^2}{\mathrm{d}q^2}\frac{1}{1-q} \\
&= \frac{2pq^2}{(1-q)^3}
\end{aligned}$$

So the variance is

$$\mathrm{var}(X) = \frac{2pq^2}{(1-q)^3} + \frac{q}{p} - \frac{q^2}{p^2} = \frac{q}{p^2}.$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 87

Indicator function

Definition (Indicator function)**.** The *indicator function* or *indicator variable* $I[A]$ (or I_A) of an event $A \subseteq \Omega$ is

$$I[A](\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- VOCABULARY

NOTE:

- probability.tex 88

Proposition.

- $\mathbb{E}[I[A]] = \sum_\omega p(\omega)I[A](\omega) = \mathbb{P}(A)$.
- $I[A^C] = 1 - I[A]$.
- $I[A \cap B] = I[A]I[B]$.
- $I[A \cup B] = I[A] + I[B] - I[A]I[B]$.
- $I[A]^2 = I[A]$.

Proposition.

- $\mathbb{E}[I[A]] = \sum_\omega p(\omega)I[A](\omega) = \mathbb{P}(A)$.
- $I[A^C] = 1 - I[A]$.
- $I[A \cap B] = I[A]I[B]$.
- $I[A \cup B] = I[A] + I[B] - I[A]I[B]$.
- $I[A]^2 = I[A]$.

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 89

Example. Let $2n$ people (n husbands and n wives, with $n > 2$) sit alternate man-woman around the table randomly. Let N be the number of couples sitting next to each other. Let $A_i = [i\text{th couple sits together}]$. Then

$$N = \sum_{i=1}^n I[A_i].$$

Then

$$\mathbb{E}[N] = \mathbb{E}\left[\sum I[A_i]\right] = \sum_1^n \mathbb{E}\big[I[A_i]\big] = n\mathbb{E}\big[I[A_1]\big] = n\mathbb{P}(A_i) = n\frac{2}{n} = 2.$$

We also have

$$\begin{aligned}
\mathbb{E}[N^2] &= \mathbb{E}\left[\left(\sum I[A_i]\right)^2\right] \\
&= \mathbb{E}\left[\sum_i I[A_i]^2 + 2\sum_{i<j} I[A_i]I[A_j]\right] \\
&= n\mathbb{E}\big[I[A_i]\big] + n(n-1)\mathbb{E}\big[I[A_1]I[A_2]\big]
\end{aligned}$$

We have $\mathbb{E}[I[A_1]I[A_2]] = \mathbb{P}(A_1 \cap A_2) = \frac{2}{n}\left(\frac{1}{n-1}\frac{1}{n-1} + \frac{n-2}{n-1}\frac{2}{n-1}\right)$. Plugging in, we ultimately obtain $\mathrm{var}(N) = \frac{2(n-2)}{n-1}$. In fact, as $n \to \infty$, $N \sim P(2)$.

Example. Let $2n$ people (n husbands and n wives, with $n > 2$) sit alternate man-woman around the table randomly. Let N be the number of couples sitting next to each other. Let $A_i = [i\text{th couple sits together}]$. Then

$$N = \sum_{i=1}^n I[A_i].$$

Then

$$\mathbb{E}[N] = \mathbb{E}\left[\sum I[A_i]\right] = \sum_1^n \mathbb{E}\big[I[A_i]\big] = n\mathbb{E}\big[I[A_1]\big] = n\mathbb{P}(A_i) = n\frac{2}{n} = 2.$$

We also have

$$\begin{aligned}
\mathbb{E}[N^2] &= \mathbb{E}\left[\left(\sum I[A_i]\right)^2\right] \\
&= \mathbb{E}\left[\sum_i I[A_i]^2 + 2\sum_{i<j} I[A_i]I[A_j]\right] \\
&= n\mathbb{E}\big[I[A_i]\big] + n(n-1)\mathbb{E}\big[I[A_1]I[A_2]\big]
\end{aligned}$$

We have $\mathbb{E}[I[A_1]I[A_2]] = \mathbb{P}(A_1 \cap A_2) = \frac{2}{n}\left(\frac{1}{n-1}\frac{1}{n-1} + \frac{n-2}{n-1}\frac{2}{n-1}\right)$. Plugging in, we ultimately obtain $\mathrm{var}(N) = \frac{2(n-2)}{n-1}$. In fact, as $n \to \infty$, $N \sim P(2)$.

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 90

Inclusion-exclusion formula

Theorem (Inclusion-exclusion formula)**.**

$$\mathbb{P}\left(\bigcup_i^n A_i\right) = \sum_1^n \mathbb{P}(A_i) - \sum_{i_1<i_2} \mathbb{P}(A_{i_1} \cap A_{j_2}) + \sum_{i_1<i_2<i_3} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n-1}\mathbb{P}(A_1 \cap \cdots \cap A_n).$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 91

Theorem (Inclusion-exclusion formula)**.**

$$\mathbb{P}\left(\bigcup_i^n A_i\right) = \sum_1^n \mathbb{P}(A_i) - \sum_{i_1<i_2} \mathbb{P}(A_{i_1} \cap A_{j_2}) + \sum_{i_1<i_2<i_3} \mathbb{P}(A_{i_1} \cap A$$

Proof. Let I_j be the indicator function for A_j. Write

$$S_r = \sum_{i_1<i_2<\cdots<i_r} I_{i_1}I_{i_2}\cdots I_{i_r},$$

and

$$s_r = \mathbb{E}[S_r] = \sum_{i_1<\cdots<i_r} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_r}).$$

Then

$$1 - \prod_{j=1}^n (1 - I_j) = S_1 - S_2 + S_3 \cdots + (-1)^{n-1}S_n.$$

So

$$\mathbb{P}\left(\bigcup_1^n A_j\right) = \mathbb{E}\left[1 - \prod_1^n (1 - I_j)\right] = s_1 - s_2 + s_3 - \cdots + (-1)^{n-1}s_n.$$

□

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 92

Independent random variables

Definition (Independent random variables)**.** Let $X_1, X_2, \cdots, X_n$ be discrete random variables. They are *independent* iff for any $x_1, x_2, \cdots, x_n$,

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

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- VOCABULARY

NOTE:

- probability.tex 93

Theorem. If $X_1, \cdots, X_n$ are independent random variables, and $f_1, \cdots, f_n$ are functions $\mathbb{R} \to \mathbb{R}$, then $f_1(X_1), \cdots, f_n(X_n)$ are independent random variables.

Theorem. If $X_1, \cdots, X_n$ are independent random variables, and $f_1, \cdots, f_n$ are functions $\mathbb{R} \to \mathbb{R}$, then $f_1(X_1), \cdots, f_n(X_n)$ are independent random variables.

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 94

Theorem. If $X_1, \cdots, X_n$ are independent random variables, and $f_1, \cdots, f_n$ are functions $\mathbb{R} \to \mathbb{R}$, then $f_1(X_1), \cdots, f_n(X_n)$ are independent random variables.

Proof. Note that given a particular y_i, there can be many different x_i for which $f_i(x_i) = y_i$. When finding $\mathbb{P}(f_i(x_i) = y_i)$, we need to sum over all x_i such that $f_i(x_i) = f_i$. Then

$$\begin{aligned}
\mathbb{P}(f_1(X_1) = y_1, \cdots f_n(X_n) = y_n) &= \sum_{\substack{x_1 : f_1(x_1) = y_1 \\ \vdots \\ x_n : f_n(x_n) = y_n}} \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) \\
&= \sum_{\substack{x_1 : f_1(x_1) = y_1 \\ \vdots \\ x_n : f_n(x_n) = y_n}} \prod_{i=1}^n \mathbb{P}(X_i = x_i) \\
&= \prod_{i=1}^n \sum_{x_i : f_i(x_i) = y_i} \mathbb{P}(X_i = x_i) \\
&= \prod_{i=1}^n \mathbb{P}(f_i(x_i) = y_i).
\end{aligned}$$

Note that the switch from the second to third line is valid since they both expand to the same mess. □

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 95

Theorem. If $X_1, \cdots, X_n$ are independent random variables and all the following expectations exists, then

$$\mathbb{E}\left[\prod X_i\right] = \prod \mathbb{E}[X_i].$$

Theorem. If $X_1, \cdots, X_n$ are independent random variables and all the following expectations exists, then

$$\mathbb{E}\left[\prod X_i\right] = \prod \mathbb{E}[X_i].$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 96

Theorem. If $X_1, \cdots, X_n$ are independent random variables and all the following expectations exists, then

$$\mathbb{E}\left[\prod X_i\right] = \prod \mathbb{E}[X_i].$$

Proof. Write R_i for the range of X_i. Then

$$\begin{aligned}
\mathbb{E}\left[\prod_1^n X_i\right] &= \sum_{x_1 \in R_1} \cdots \sum_{x_n \in R_n} x_1 x_2 \cdots x_n \times \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) \\
&= \prod_{i=1}^n \sum_{x_i \in R_i} x_i \mathbb{P}(X_i = x_i) \\
&= \prod_{i=1}^n \mathbb{E}[X_i].
\end{aligned}$$

□

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 97

Corollary. Let $X_1, \cdots X_n$ be independent random variables, and $f_1, f_2, \cdots f_n$ are functions $\mathbb{R} \to \mathbb{R}$. Then

$$\mathbb{E}\left[\prod f_i(x_i)\right] = \prod \mathbb{E}[f_i(x_i)].$$

Corollary. Let $X_1, \cdots X_n$ be independent random variables, and $f_1, f_2, \cdots f_n$ are functions $\mathbb{R} \to \mathbb{R}$. Then

$$\mathbb{E}\left[\prod f_i(x_i)\right] = \prod \mathbb{E}[f_i(x_i)].$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 98

Theorem. If $X_1, X_2, \cdots X_n$ are independent random variables, then

$$\operatorname{var}\left(\sum X_i\right) = \sum \operatorname{var}(X_i).$$

Theorem. If $X_1, X_2, \cdots X_n$ are independent random variables, then

$$\operatorname{var}\left(\sum X_i\right) = \sum \operatorname{var}(X_i).$$

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 99

Theorem. If $X_1, X_2, \cdots X_n$ are independent random variables, then

$$\operatorname{var}\left(\sum X_i\right) = \sum \operatorname{var}(X_i).$$

Proof.

$$\begin{aligned}
\operatorname{var}\left(\sum X_i\right) &= \mathbb{E}\left[\left(\sum X_i\right)^2\right] - \left(\mathbb{E}\left[\sum X_i\right]\right)^2 \\
&= \mathbb{E}\left[\sum X_i^2 + \sum_{i \neq j} X_i X_j\right] - \left(\sum \mathbb{E}[X_i]\right)^2 \\
&= \sum \mathbb{E}[X_i^2] + \sum_{i \neq j} \mathbb{E}[X_i]\mathbb{E}[X_j] - \sum (\mathbb{E}[X_i])^2 - \sum_{i \neq j} \\
&= \sum \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2.
\end{aligned}$$

□

- 3.1 Discrete random variables
- 3 Discrete random variables
-
- PROOF EXERCISE

NOTE:

- `probability.tex 100`

Corollary. Let $X_1, X_2, \cdots X_n$ be independent identically distributed random variables (iid rvs). Then

$$\operatorname{var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\operatorname{var}(X_1).$$

Corollary. Let $X_1, X_2, \cdots X_n$ be independent identically distributed random variables (iid rvs). Then

$$\operatorname{var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\operatorname{var}(X_1).$$

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 101`

Corollary. Let $X_1, X_2, \cdots X_n$ be independent identically distributed random variables (iid rvs). Then

$$\operatorname{var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\operatorname{var}(X_1).$$

Proof.

$$\begin{aligned}\operatorname{var}\left(\frac{1}{n}\sum X_i\right) &= \frac{1}{n^2}\operatorname{var}\left(\sum X_i\right)\\ &= \frac{1}{n^2}\sum \operatorname{var}(X_i)\\ &= \frac{1}{n^2} n \operatorname{var}(X_1)\\ &= \frac{1}{n}\operatorname{var}(X_1)\end{aligned}$$

□

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `PROOF EXERCISE`

NOTE:

- `probability.tex 102`

Example. Let X_i be iid $B(1,p)$, i.e. $\mathbb{P}(1) = p$ and $\mathbb{P}(0) = 1-p$. Then $Y = X_1 + X_2 + \cdots + X_n \sim B(n,p)$. Since $\operatorname{var}(X_i) = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = p - p^2 = p(1-p)$, we have $\operatorname{var}(Y) = np(1-p)$.

Example. Let X_i be iid $B(1,p)$, i.e. $\mathbb{P}(1) = p$ and $\mathbb{P}(0) = 1-p$. Then $Y = X_1 + X_2 + \cdots + X_n \sim B(n,p)$. Since $\operatorname{var}(X_i) = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = p - p^2 = p(1-p)$, we have $\operatorname{var}(Y) = np(1-p)$.

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 103`

Example. Suppose we have two rods of unknown lengths a, b. We can measure the lengths, but is not accurate. Let A and B be the measured value. Suppose

$$\mathbb{E}[A] = a, \quad \operatorname{var}(A) = \sigma^2$$
$$\mathbb{E}[B] = b, \quad \operatorname{var}(B) = \sigma^2.$$

We can measure it more accurately by measuring $X = A + B$ and $Y = A - B$. Then we estimate a and b by

$$\hat{a} = \frac{X+Y}{2},\ \hat{b} = \frac{X-Y}{2}.$$

Then $\mathbb{E}[\hat{a}] = a$ and $\mathbb{E}[\hat{b}] = b$, i.e. they are unbiased. Also

$$\operatorname{var}(\hat{a}) = \frac{1}{4}\operatorname{var}(X+Y) = \frac{1}{4}2\sigma^2 = \frac{1}{2}\sigma^2,$$

and similarly for b. So we can measure it more accurately by measuring the sticks together instead of separately.

Example. Suppose we have two rods of unknown lengths a, b. We can measure the lengths, but is not accurate. Let A and B be the measured value. Suppose

$$\mathbb{E}[A] = a, \quad \operatorname{var}(A) = \sigma^2$$
$$\mathbb{E}[B] = b, \quad \operatorname{var}(B) = \sigma^2.$$

We can measure it more accurately by measuring $X = A + B$ and $Y = A - B$. Then we estimate a and b by

$$\hat{a} = \frac{X+Y}{2},\ \hat{b} = \frac{X-Y}{2}.$$

Then $\mathbb{E}[\hat{a}] = a$ and $\mathbb{E}[\hat{b}] = b$, i.e. they are unbiased. Also

$$\operatorname{var}(\hat{a}) = \frac{1}{4}\operatorname{var}(X+Y) = \frac{1}{4}2\sigma^2 = \frac{1}{2}\sigma^2,$$

and similarly for b. So we can measure it more accurately by measuring the sticks together instead of separately.

- `3.1 Discrete random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

4.2 Inequalities

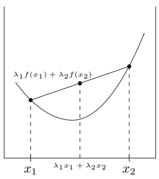
NOTE:

- `probability.tex 104`

Definition (Convex function)**.** A function $f : (a,b) \to \mathbb{R}$ is *convex* if for all $x_1, x_2 \in (a,b)$ and $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1 + \lambda_2 = 1$,

$$\lambda_1 f(x_1) + \lambda_2 f(x_2) \geq f(\lambda_1 x_1 + \lambda_2 x_2).$$

It is *strictly convex* if the inequality above is strict (except when $x_1 = x_2$ or λ_1 or $\lambda_2 = 0$).

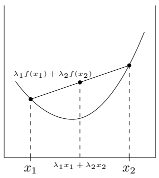


A function is *concave* if $-f$ is convex.

Definition (Convex function)**.** A function $f : (a,b) \to \mathbb{R}$ is *convex* if for all $x_1, x_2 \in (a,b)$ and $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1 + \lambda_2 = 1$,

$$\lambda_1 f(x_1) + \lambda_2 f(x_2) \geq f(\lambda_1 x_1 + \lambda_2 x_2).$$

It is *strictly convex* if the inequality above is strict (except when $x_1 = x_2$ or λ_1 or $\lambda_2 = 0$).



A function is *concave* if $-f$ is convex.

- `3.2 Inequalities`
- `3 Discrete random variables`
-
- `VOCABULARY`

NOTE:

- `probability.tex 105`

Proposition. If f is differentiable and $f''(x) \geq 0$ for all $x \in (a,b)$, then it is convex. It is strictly convex if $f''(x) > 0$.

Proposition. If f is differentiable and $f''(x) \geq 0$ for all $x \in (a,b)$, then it is convex. It is strictly convex if $f''(x) > 0$.

- `3.2 Inequalities`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 106`

Jensen's inequality

Theorem (Jensen's inequality)**.** If $f : (a,b) \to \mathbb{R}$ is convex, then

$$\sum_{i=1}^n p_i f(x_i) \geq f\left(\sum_{i=1}^n p_i x_i\right)$$

for all $p_1, p_2, \cdots, p_n$ such that $p_i \geq 0$ and $\sum p_i = 1$, and $x_i \in (a,b)$. This says that $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ (where $\mathbb{P}(X = x_i) = p_i$). If f is strictly convex, then equalities hold only if all x_i are equal, i.e. X takes only one possible value.

- `3.2 Inequalities`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 107`

Theorem (Jensen's inequality)**.** If $f : (a,b) \to \mathbb{R}$ is convex, then

$$\sum_{i=1}^n p_i f(x_i) \geq f\left(\sum_{i=1}^n p_i x_i\right)$$

for all $p_1, p_2, \cdots, p_n$ such that $p_i \geq 0$ and $\sum p_i = 1$, and $x_i \in (a,b)$. This says that $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ (where $\mathbb{P}(X = x_i) = p_i$). If f is strictly convex, then equalities hold only if all x_i are equal, i.e. X takes only one possible value.

Proof. Induct on n. It is true for $n = 2$ by the definition of convexity. Then

$$\begin{aligned}f(p_1x_1 + \cdots + p_nx_n) &= f\left(p_1x_1 + (p_2 + \cdots + p_n)\frac{p_2x_2 + \cdots + l_n}{p_2 + \cdots + p}\right.\\ &\leq p_1 f(x_1) + (p_2 + \cdots p_n) f\left(\frac{p_2x_2 + \cdots + p}{p_2 + \cdots + p}\right.\\ &\leq p_1 f(x_1) + (p_2 + \cdots + p_n)\left[\frac{p_2}{(\ )}f(x_2) + \cdots\right.\\ &= p_1 f(x_1) + \cdots + p_n(x_n).\end{aligned}$$

where the ( ) is $p_2 + \cdots + p_n$. Strictly convex case is proved with $\leq$ replaced by $<$ by definition of strict convexity. □

- `3.2 Inequalities`
- `3 Discrete random variables`
-
- `PROOF EXERCISE`

NOTE:

- probability.tex 108

AM-GM inequality

Corollary (AM-GM inequality)**.** Given $x_1, \cdots, x_n$ positive reals, then

$$\left(\prod x_i\right)^{1/n} \leq \frac{1}{n}\sum x_i.$$

- 3.2 Inequalities
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 109

Corollary (AM-GM inequality)**.** Given $x_1, \cdots, x_n$ positive reals, then

$$\left(\prod x_i\right)^{1/n} \leq \frac{1}{n}\sum x_i.$$

Proof. Take $f(x) = -\log x$. This is convex since its second derivative is $x^{-2} > 0$. Take $\mathbb{P}(x = x_i) = 1/n$. Then

$$\mathbb{E}[f(x)] = \frac{1}{n}\sum -\log x_i = -\log \mathrm{GM}$$

and

$$f(\mathbb{E}[x]) = -\log \frac{1}{n}\sum x_i = -\log \mathrm{AM}$$

Since $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$, AM $\geq$ GM. Since $-\log x$ is strictly convex, AM = GM only if all x_i are equal. □

- 3.2 Inequalities
- 3 Discrete random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 110

Cauchy-Schwarz inequality

Theorem (Cauchy-Schwarz inequality)**.** For any two random variables X, Y,

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

- 3.2 Inequalities
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 111

Theorem (Cauchy-Schwarz inequality)**.** For any two random variables X, Y,

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Proof. If $Y = 0$, then both sides are 0. Otherwise, $\mathbb{E}[Y^2] > 0$. Let

$$w = X - Y\cdot\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}.$$

Then

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

Since $\mathbb{E}[w^2] \geq 0$, the Cauchy-Schwarz inequality follows. □

- 3.2 Inequalities
- 3 Discrete random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 112

Markov inequality

Theorem (Markov inequality)**.** If X is a random variable with $\mathbb{E}|X| < \infty$ and $\varepsilon > 0$, then

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}|X|}{\varepsilon}.$$

- 3.2 Inequalities
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 113

Theorem (Markov inequality)**.** If X is a random variable with $\mathbb{E}|X| < \infty$ and $\varepsilon > 0$, then

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}|X|}{\varepsilon}.$$

Proof. We make use of the indicator function. We have

$$I[|X| \geq \varepsilon] \leq \frac{|X|}{\varepsilon}.$$

This is proved by exhaustion: if $|X| \geq \varepsilon$, then LHS = 1 and RHS ≥ 1; If $|X| < \varepsilon$, then LHS = 0 and RHS is non-negative. Take the expected value to obtain

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}|X|}{\varepsilon}.$$

□

- 3.2 Inequalities
- 3 Discrete random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 114

Chebyshev inequality

Theorem (Chebyshev inequality)**.** If X is a random variable with $\mathbb{E}[X^2] < \infty$ and $\varepsilon > 0$, then

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[X^2]}{\varepsilon^2}.$$

- 3.2 Inequalities
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 115

Theorem (Chebyshev inequality)**.** If X is a random variable with $\mathbb{E}[X^2] < \infty$ and $\varepsilon > 0$, then

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[X^2]}{\varepsilon^2}.$$

Proof. Again, we have

$$I[\{|X| \geq \varepsilon\}] \leq \frac{x^2}{\varepsilon^2}.$$

Then take the expected value and the result follows. □

- 3.2 Inequalities
- 3 Discrete random variables
-
- PROOF EXERCISE

4.3 Weak law of large numbers

NOTE:

- `probability.tex 116`

Theorem (Weak law of large numbers)**.** Let $X_1, X_2, \cdots$ be iid random variables, with mean μ and var σ^2. Let $S_n = \sum_{i=1}^n X_i$. Then for all $\varepsilon > 0$,

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

as $n \to \infty$. We say, $\frac{S_n}{n}$ tends to μ (in probability), or

$$\frac{S_n}{n} \to_p \mu.$$

Theorem (Weak law of large numbers)**.** Let $X_1, X_2, \cdots$ be iid random variables, with mean μ and var σ^2. Let $S_n = \sum_{i=1}^n X_i$. Then for all $\varepsilon > 0$,

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

as $n \to \infty$. We say, $\frac{S_n}{n}$ tends to μ (in probability), or

$$\frac{S_n}{n} \to_p \mu.$$

- `3.3 Weak law of large numbers`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 117`

Theorem (Weak law of large numbers)**.** Let $X_1, X_2, \cdots$ be iid random variables, with mean μ and var σ^2. Let $S_n = \sum_{i=1}^n X_i$. Then for all $\varepsilon > 0$,

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

as $n \to \infty$. We say, $\frac{S_n}{n}$ tends to μ (in probability), or

$$\frac{S_n}{n} \to_p \mu.$$

Proof. By Chebyshev,

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) &\leq \frac{\mathbb{E}\left(\frac{S_n}{n} - \mu\right)^2}{\varepsilon^2} \\
&= \frac{1}{n^2}\frac{\mathbb{E}(S_n - n\mu)^2}{\varepsilon^2} \\
&= \frac{1}{n^2\varepsilon^2}\operatorname{var}(S_n) \\
&= \frac{n}{n^2\varepsilon^2}\operatorname{var}(X_1) \\
&= \frac{\sigma^2}{n\varepsilon^2} \to 0
\end{aligned}$$

□

- `3.3 Weak law of large numbers`
- `3 Discrete random variables`
-
- `PROOF EXERCISE`

NOTE:

- `probability.tex 118`

Example. Suppose we toss a coin with probability p of heads. Then

$$\frac{S_n}{n} = \frac{\text{number of heads}}{\text{number of tosses}}.$$

Since $\mathbb{E}[X_i] = p$, then the weak law of large number tells us that

$$\frac{S_n}{n} \to_p p.$$

This means that as we toss more and more coins, the proportion of heads will tend towards p.

Example. Suppose we toss a coin with probability p of heads. Then

$$\frac{S_n}{n} = \frac{\text{number of heads}}{\text{number of tosses}}.$$

Since $\mathbb{E}[X_i] = p$, then the weak law of large number tells us that

$$\frac{S_n}{n} \to_p p.$$

This means that as we toss more and more coins, the proportion of heads will tend towards p.

- `3.3 Weak law of large numbers`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 119`

Strong law of large numbers

Theorem (Strong law of large numbers)**.**

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

We say

$$\frac{S_n}{n} \to_{\text{as}} \mu,$$

where "as" means "almost surely".

- `3.3 Weak law of large numbers`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

4.4 Multiple random variables

NOTE:

- `probability.tex 120`

Covariance

Definition (Covariance)**.** Given two random variables X, Y, the *covariance* is

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `VOCABULARY`

NOTE:

- `probability.tex 121`

Proposition.

(i) $\operatorname{cov}(X, c) = 0$ for constant c.
(ii) $\operatorname{cov}(X + c, Y) = \operatorname{cov}(X, Y)$.
(iii) $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$.
(iv) $\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
(v) $\operatorname{cov}(X, X) = \operatorname{var}(X)$.
(vi) $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y)$.
(vii) If X, Y are independent, $\operatorname{cov}(X, Y) = 0$.

Proposition.

(i) $\operatorname{cov}(X, c) = 0$ for constant c.
(ii) $\operatorname{cov}(X + c, Y) = \operatorname{cov}(X, Y)$.
(iii) $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$.
(iv) $\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
(v) $\operatorname{cov}(X, X) = \operatorname{var}(X)$.
(vi) $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y)$.
(vii) If X, Y are independent, $\operatorname{cov}(X, Y) = 0$.

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 122`

Example.

- Let $(X, Y) = (2, 0), (-1, -1)$ or $(-1, 1)$ with equal probabilities of $1/3$. These are not independent since $Y = 0 \Rightarrow X = 2$. However, $\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 \cdot 0 = 0$.
- If we randomly pick a point on the unit circle, and let the coordinates be (X, Y), then $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[XY] = 0$ by symmetry. So $\operatorname{cov}(X, Y) = 0$ but X and Y are clearly not independent (they have to satisfy $x^2 + y^2 = 1$).

Example.

- Let $(X, Y) = (2, 0), (-1, -1)$ or $(-1, 1)$ with equal probabilities of $1/3$. These are not independent since $Y = 0 \Rightarrow X = 2$. However, $\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 \cdot 0 = 0$.
- If we randomly pick a point on the unit circle, and let the coordinates be (X, Y), then $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[XY] = 0$ by symmetry. So $\operatorname{cov}(X, Y) = 0$ but X and Y are clearly not independent (they have to satisfy $x^2 + y^2 = 1$).

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 123`

Correlation coefficient

Definition (Correlation coefficient)**.** The *correlation coefficient* of X and Y is

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}.$$

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `VOCABULARY`

NOTE:

- `probability.tex 124`

Proposition. $|\operatorname{corr}(X,Y)| \leq 1$.

Proposition. $|\operatorname{corr}(X,Y)| \leq 1$.

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 125`

Proposition. $|\operatorname{corr}(X,Y)| \leq 1$.

Proof. Apply Cauchy-Schwarz to $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$. □

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `PROOF EXERCISE`

NOTE:

- `probability.tex 126`

Definition (Conditional distribution)**.** Let X and Y be random variables (in general not independent) with joint distribution $\mathbb{P}(X = x, Y = y)$. Then the *marginal distribution* (or simply *distribution*) of X is

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

The *conditional distribution* of X given Y is

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

The *conditional expectation* of X given Y is

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

We can view $\mathbb{E}[X \mid Y]$ as a random variable in Y: given a value of Y, we return the expectation of X.

Definition (Conditional distribution)**.** Let X and Y be random variables (in general not independent) with joint distribution $\mathbb{P}(X = x, Y = y)$. Then the *marginal distribution* (or simply *distribution*) of X is

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

The *conditional distribution* of X given Y is

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

The *conditional expectation* of X given Y is

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

We can view $\mathbb{E}[X \mid Y]$ as a random variable in Y: given a value of Y, we return the expectation of X.

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `VOCABULARY`

NOTE:

- `probability.tex 127`

Example. Consider a dice roll. Let $Y = 1$ denote an even roll and $Y = 0$ denote an odd roll. Let X be the value of the roll. Then $\mathbb{E}[X \mid Y] = 3 + Y$, ie 4 if even, 3 if odd.

Example. Consider a dice roll. Let $Y = 1$ denote an even roll and $Y = 0$ denote an odd roll. Let X be the value of the roll. Then $\mathbb{E}[X \mid Y] = 3 + Y$, ie 4 if even, 3 if odd.

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 128`

Example. Let $X_1, \cdots, X_n$ be iid $B(1, p)$. Let $Y = X_1 + \cdots + X_n$. Then

$$\begin{aligned}\mathbb{P}(X_1 = 1 \mid Y = r) &= \frac{\mathbb{P}(X_1 = 1, \sum_2^n X_i = r - 1)}{\mathbb{P}(Y = r)}\\ &= \frac{p\binom{n-1}{r-1}p^{r-1}(1-p)^{(n-1)-(r-1)}}{\binom{n}{r}p^r(1-p)^{n-1}} = \frac{r}{n}.\end{aligned}$$

So

$$\mathbb{E}[X_1 \mid Y] = 1 \cdot \frac{r}{n} + 0\left(1 - \frac{r}{n}\right) = \frac{r}{n} = \frac{Y}{n}.$$

Note that this is a random variable!

Example. Let $X_1, \cdots, X_n$ be iid $B(1, p)$. Let $Y = X_1 + \cdots + X_n$. Then

$$\begin{aligned}\mathbb{P}(X_1 = 1 \mid Y = r) &= \frac{\mathbb{P}(X_1 = 1, \sum_2^n X_i = r - 1)}{\mathbb{P}(Y = r)}\\ &= \frac{p\binom{n-1}{r-1}p^{r-1}(1-p)^{(n-1)-(r-1)}}{\binom{n}{r}p^r(1-p)^{n-1}} = \frac{r}{n}.\end{aligned}$$

So

$$\mathbb{E}[X_1 \mid Y] = 1 \cdot \frac{r}{n} + 0\left(1 - \frac{r}{n}\right) = \frac{r}{n} = \frac{Y}{n}.$$

Note that this is a random variable!

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 129`

Theorem. If X and Y are independent, then

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

Theorem. If X and Y are independent, then

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

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

- `probability.tex 130`

Theorem. If X and Y are independent, then

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

Proof.

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

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `PROOF EXERCISE`

NOTE:

- `probability.tex 131`

Tower property of conditional expectation

Theorem (Tower property of conditional expectation)**.**

$$\mathbb{E}_Y[\mathbb{E}_X[X \mid Y]] = \mathbb{E}_X[X],$$

where the subscripts indicate what variable the expectation is taken over.

- `3.4 Multiple random variables`
- `3 Discrete random variables`
-
- `GENERAL KNOWLEDGE`

NOTE:

– probability.tex 132

Theorem (Tower property of conditional expectation)**.**

$$\mathbb{E}_Y[\mathbb{E}_X[X \mid Y]] = \mathbb{E}_X[X],$$

where the subscripts indicate what variable the expectation is taken over.

Proof.

$$\begin{aligned}\mathbb{E}_Y[\mathbb{E}_X[X \mid Y]] &= \sum_y \mathbb{P}(Y = y)\mathbb{E}[X \mid Y = y] \\ &= \sum_y \mathbb{P}(Y = y) \sum_x x\mathbb{P}(X = x \mid Y = y) \\ &= \sum_x \sum_y x\mathbb{P}(X = x, Y = y) \\ &= \sum_x x \sum_y \mathbb{P}(X = x, Y = y) \\ &= \sum_x x\mathbb{P}(X = x) \\ &= \mathbb{E}[X].\end{aligned}$$

□

– 3.4 Multiple random variables

– 3 Discrete random variables

–

– PROOF EXERCISE

4.5 Probability generating functions

NOTE:

– probability.tex 133

Definition (Probability generating function (pgf))**.** The *probability generating function (pgf)* of X is

$$p(z) = \mathbb{E}[z^X] = \sum_{r=0}^\infty \mathbb{P}(X = r)z^r = p_0 + p_1 z + p_2 z^2 \cdots = \sum_0^\infty p_r z^r.$$

This is a power series (or polynomial), and converges if $|z| \leq 1$, since

$$|p(z)| \leq \sum_r p_r |z^r| \leq \sum_r p_r = 1.$$

We sometimes write as $p_X(z)$ to indicate what the random variable.

Definition (Probability generating function (pgf))**.** The *probability generating function (pgf)* of X is

$$p(z) = \mathbb{E}[z^X] = \sum_{r=0}^\infty \mathbb{P}(X = r)z^r = p_0 + p_1 z + p_2 z^2 \cdots = \sum_0^\infty p_r z^r.$$

This is a power series (or polynomial), and converges if $|z| \leq 1$, since

$$|p(z)| \leq \sum_r p_r |z^r| \leq \sum_r p_r = 1.$$

We sometimes write as $p_X(z)$ to indicate what the random variable.

– 3.5 Probability generating functions

– 3 Discrete random variables

–

– VOCABULARY

NOTE:

– probability.tex 134

Example. Consider a fair di.e. Then $p_r = 1/6$ for $r = 1, \cdots, 6$. So

$$p(z) = \mathbb{E}[z^X] = \frac{1}{6}(z + z^2 + \cdots + z^6) = \frac{1}{6}z\left(\frac{1 - z^6}{1 - z}\right).$$

Example. Consider a fair di.e. Then $p_r = 1/6$ for $r = 1, \cdots, 6$. So

$$p(z) = \mathbb{E}[z^X] = \frac{1}{6}(z + z^2 + \cdots + z^6) = \frac{1}{6}z\left(\frac{1 - z^6}{1 - z}\right).$$

– 3.5 Probability generating functions

– 3 Discrete random variables

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 135

Theorem. The distribution of X is uniquely determined by its probability generating function.

Theorem. The distribution of X is uniquely determined by its probability generating function.

– 3.5 Probability generating functions

– 3 Discrete random variables

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 136

Theorem. The distribution of X is uniquely determined by its probability generating function.

Proof. By definition, $p_0 = p(0)$, $p_1 = p'(0)$ etc. (where p' is the derivative of p). In general,

$$\left.\frac{\mathrm{d}^i}{\mathrm{d}z^i}p(z)\right|_{z=0} = i!p_i.$$

So we can recover $(p_0, p_1, \cdots)$ from $p(z)$. □

– 3.5 Probability generating functions

– 3 Discrete random variables

–

– PROOF EXERCISE

NOTE:

– probability.tex 137

Abel's lemma

Theorem (Abel's lemma)**.**

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

If $p'(z)$ is continuous, then simply $\mathbb{E}[X] = p'(1)$.

– 3.5 Probability generating functions

– 3 Discrete random variables

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 138

Theorem (Abel's lemma)**.**

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

If $p'(z)$ is continuous, then simply $\mathbb{E}[X] = p'(1)$.

Proof. For $z < 1$, we have

$$p'(z) = \sum_1^\infty r p_r z^{r-1} \leq \sum_1^\infty r p_r = \mathbb{E}[X].$$

So we must have

$$\lim_{z \to 1} p'(z) \leq \mathbb{E}[X].$$

On the other hand, for any ε, if we pick N large, then

$$\sum_1^N r p_r \geq \mathbb{E}[X] - \varepsilon.$$

So

$$\mathbb{E}[X] - \varepsilon \leq \sum_1^N r p_r = \lim_{z \to 1} \sum_1^N r p_r z^{r-1} \leq \lim_{z \to 1} \sum_1^\infty r p_r z^{r-1} = \lim_{z \to 1} p'(z).$$

So $\mathbb{E}[X] \leq \lim_{z \to 1} p'(z)$. So the result follows □

– 3.5 Probability generating functions

– 3 Discrete random variables

–

– PROOF EXERCISE

NOTE:

– probability.tex 139

Theorem.

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

Theorem.

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

– 3.5 Probability generating functions

– 3 Discrete random variables

–

– GENERAL KNOWLEDGE

NOTE:

- probability.tex 140

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

Proof. Same as above. $\square$

- 3.5 Probability generating functions
- 3 Discrete random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 141

Example. Consider the Poisson distribution. Then
$$p_r = \mathbb{P}(X=r) = \frac{1}{r!}\lambda^r e^{-\lambda}.$$
Then
$$p(z) = \mathbb{E}[z^X] = \sum_0^\infty z^r \frac{1}{r!}\lambda^r e^{-\lambda} = e^{\lambda z}e^{-\lambda} = e^{\lambda(z-1)}.$$
We can have a sanity check: $p(1) = 1$, which makes sense, since $p(1)$ is the sum of probabilities. We have
$$\mathbb{E}[X] = \frac{\mathrm{d}}{\mathrm{d}z}e^{\lambda(z-1)}\bigg|_{z=1} = \lambda,$$
and
$$\mathbb{E}[X(X-1)] = \frac{\mathrm{d}^2}{\mathrm{d}x^2}e^{\lambda(z-1)}\bigg|_{z=1} = \lambda^2$$
So
$$\operatorname{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Example. Consider the Poisson distribution. Then
$$p_r = \mathbb{P}(X=r) = \frac{1}{r!}\lambda^r e^{-\lambda}.$$
Then
$$p(z) = \mathbb{E}[z^X] = \sum_0^\infty z^r \frac{1}{r!}\lambda^r e^{-\lambda} = e^{\lambda z}e^{-\lambda} = e^{\lambda(z-1)}.$$
We can have a sanity check: $p(1) = 1$, which makes sense, since $p(1)$ is the sum of probabilities. We have
$$\mathbb{E}[X] = \frac{\mathrm{d}}{\mathrm{d}z}e^{\lambda(z-1)}\bigg|_{z=1} = \lambda,$$
and
$$\mathbb{E}[X(X-1)] = \frac{\mathrm{d}^2}{\mathrm{d}x^2}e^{\lambda(z-1)}\bigg|_{z=1} = \lambda^2$$
So
$$\operatorname{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

- 3.5 Probability generating functions
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 142

Theorem. Suppose $X_1, X_2, \cdots, X_n$ are independent random variables with pgfs $p_1, p_2, \cdots, p_n$. Then the pgf of $X_1 + X_2 + \cdots + X_n$ is $p_1(z)p_2(z)\cdots p_n(z)$.

Theorem. Suppose $X_1, X_2, \cdots, X_n$ are independent random variables with pgfs $p_1, p_2, \cdots, p_n$. Then the pgf of $X_1 + X_2 + \cdots + X_n$ is $p_1(z)p_2(z)\cdots p_n(z)$.

- 3.5 Probability generating functions
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 143

Theorem. Suppose $X_1, X_2, \cdots, X_n$ are independent random variables with pgfs $p_1, p_2, \cdots, p_n$. Then the pgf of $X_1 + X_2 + \cdots + X_n$ is $p_1(z)p_2(z)\cdots p_n(z)$.

Proof.
$\mathbb{E}[z^{X_1+\cdots+X_n}] = \mathbb{E}[z^{X_1}\cdots z^{X_n}] = \mathbb{E}[z^{X_1}]\cdots\mathbb{E}[z^{X_n}] = p_1(z)\cdots p_n$
$\square$

- 3.5 Probability generating functions
- 3 Discrete random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 144

Example. Let $X \sim B(n,p)$. Then
$$p(z) = \sum_{r=0}^n \mathbb{P}(X=r)z^r = \sum \binom{n}{r}p^r(1-p)^{n-r}z^r = (pz+(1-p))^n = (pz+q)^n.$$
So $p(z)$ is the product of n copies of $pz+q$. But $pz+q$ is the pgf of $Y \sim B(1,p)$. This shows that $X = Y_1 + Y_2 + \cdots + Y_n$ (which we already knew), i.e. a binomial distribution is the sum of Bernoulli trials.

Example. Let $X \sim B(n,p)$. Then
$$p(z) = \sum_{r=0}^n \mathbb{P}(X=r)z^r = \sum \binom{n}{r}p^r(1-p)^{n-r}z^r = (pz+(1-p))^n = (pz+q)^n.$$
So $p(z)$ is the product of n copies of $pz+q$. But $pz+q$ is the pgf of $Y \sim B(1,p)$. This shows that $X = Y_1 + Y_2 + \cdots + Y_n$ (which we already knew), i.e. a binomial distribution is the sum of Bernoulli trials.

- 3.5 Probability generating functions
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 145

Example. If X and Y are independent Poisson random variables with parameters λ, μ respectively, then
$$\mathbb{E}[t^{X+Y}] = \mathbb{E}[t^X]\mathbb{E}[t^Y] = e^{\lambda(t-1)}e^{\mu(t-1)} = e^{(\lambda+\mu)(t-1)}$$
So $X + Y \sim \mathbb{P}(\lambda+\mu)$. We can also do it directly:
$$\mathbb{P}(X+Y=r) = \sum_{i=0}^r \mathbb{P}(X=i, Y=r-i) = \sum_{i=0}^r \mathbb{P}(X=i)\mathbb{P}(X=r-i),$$
but is much more complicated.

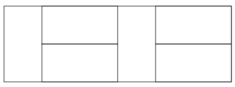
Example. If X and Y are independent Poisson random variables with parameters λ, μ respectively, then
$$\mathbb{E}[t^{X+Y}] = \mathbb{E}[t^X]\mathbb{E}[t^Y] = e^{\lambda(t-1)}e^{\mu(t-1)} = e^{(\lambda+\mu)(t-1)}$$
So $X + Y \sim \mathbb{P}(\lambda+\mu)$. We can also do it directly:
$$\mathbb{P}(X+Y=r) = \sum_{i=0}^r \mathbb{P}(X=i, Y=r-i) = \sum_{i=0}^r \mathbb{P}(X=i)\mathbb{P}(X=r-i),$$
but is much more complicated.

- 3.5 Probability generating functions
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

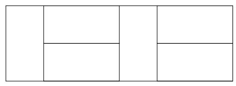
- probability.tex 146

Example. Suppose we want to tile a $2 \times n$ bathroom by 2×1 tiles. One way to do it is



We can do it recursively: suppose there are f_n ways to tile a $2 \times n$ grid. Then if we start tiling, the first tile is either vertical, in which we have f_{n-1} ways to tile the remaining ones; or the first tile is horizontal, in which we have f_{n-2} ways to tile the remaining. So
$$f_n = f_{n-1} + f_{n-2},$$
which is simply the Fibonacci sequence, with $f_0 = f_1 = 1$. Let
$$F(z) = \sum_{n=0}^\infty f_n z^n.$$
Then from our recurrence relation, we obtain
$$f_n z^n = f_{n-1}z^n + f_{n-2}z^n.$$
So
$$\sum_{n=2}^\infty f_n z^n = \sum_{n=2}^\infty f_{n-1}z^n + \sum_{n=2}^\infty f_{n-2}z^n.$$
Since $f_0 = f_1 = 1$, we have
$$F(z) - f_0 - zf_1 = z(F(z) - f_0) + z^2F(z).$$
Thus $F(z) = (1 - z - z^2)^{-1}$. If we write
$$\alpha_1 = \frac{1}{2}(1+\sqrt{5}), \quad \alpha_2 = \frac{1}{2}(1-\sqrt{5}).$$
then we have
$$\begin{aligned} F(z) &= (1-z-z^2)^{-1} \\ &= \frac{1}{(1-\alpha_1 z)(1-\alpha_2 z)} \\ &= \frac{1}{\alpha_1 - \alpha_2}\left(\frac{\alpha_1}{1-\alpha_1 z} - \frac{\alpha_2}{1-\alpha_2 z}\right) \\ &= \frac{1}{\alpha_1-\alpha_2}\left(\alpha_1\sum_{n=0}^\infty \alpha_1^n z^n - \alpha_2 \sum_{n=0}^\infty \alpha_2^n z^n\right). \end{aligned}$$
So
$$f_n = \frac{\alpha_1^{n+1} - \alpha_2^{n+1}}{\alpha_1 - \alpha_2}.$$

Example. Suppose we want to tile a $2 \times n$ bathroom by 2×1 tiles. One way to do it is



We can do it recursively: suppose there are f_n ways to tile a $2 \times n$ grid. Then if we start tiling, the first tile is either vertical, in which we have f_{n-1} ways to tile the remaining ones; or the first tile is horizontal, in which we have f_{n-2} ways to tile the remaining. So
$$f_n = f_{n-1} + f_{n-2},$$
which is simply the Fibonacci sequence, with $f_0 = f_1 = 1$. Let
$$F(z) = \sum_{n=0}^\infty f_n z^n.$$
Then from our recurrence relation, we obtain
$$f_n z^n = f_{n-1}z^n + f_{n-2}z^n.$$
So
$$\sum_{n=2}^\infty f_n z^n = \sum_{n=2}^\infty f_{n-1}z^n + \sum_{n=2}^\infty f_{n-2}z^n.$$
Since $f_0 = f_1 = 1$, we have
$$F(z) - f_0 - zf_1 = z(F(z) - f_0) + z^2F(z).$$
Thus $F(z) = (1 - z - z^2)^{-1}$. If we write
$$\alpha_1 = \frac{1}{2}(1+\sqrt{5}), \quad \alpha_2 = \frac{1}{2}(1-\sqrt{5}).$$
then we have
$$\begin{aligned} F(z) &= (1-z-z^2)^{-1} \\ &= \frac{1}{(1-\alpha_1 z)(1-\alpha_2 z)} \\ &= \frac{1}{\alpha_1 - \alpha_2}\left(\frac{\alpha_1}{1-\alpha_1 z} - \frac{\alpha_2}{1-\alpha_2 z}\right) \\ &= \frac{1}{\alpha_1-\alpha_2}\left(\alpha_1\sum_{n=0}^\infty \alpha_1^n z^n - \alpha_2 \sum_{n=0}^\infty \alpha_2^n z^n\right). \end{aligned}$$
So
$$f_n = \frac{\alpha_1^{n+1} - \alpha_2^{n+1}}{\alpha_1 - \alpha_2}.$$

- 3.5 Probability generating functions
- 3 Discrete random variables
-
- GENERAL KNOWLEDGE

NOTE:

– `probability.tex 147`

Example. A *Dyck word* is a string of brackets that match, such as (), or ((())()). There is only one Dyck word of length 2, (). There are 2 of length 4, (()) and ()(). Similarly, there are 5 Dyck words of length 5. Let C_n be the number of Dyck words of length $2n$. We can split each Dyck word into $(w_1)w_2$, where w_1 and w_2 are Dyck words. Since the lengths of w_1 and w_2 must sum up to $2(n-1)$,

$$C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}. \qquad (*)$$

We again use pgf-like functions: let

$$c(x) = \sum_{n=0}^{\infty} C_n x^n.$$

From $(*)$, we can show that

$$c(x) = 1 + xc(x)^2.$$

We can solve to show that

$$c(x) = \frac{1-\sqrt{1-4x}}{2x} = \sum_{0}^{\infty} \binom{2n}{n} \frac{x^n}{n+1},$$

noting that $C_0 = 1$. Then

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

Example. A *Dyck word* is a string of brackets that match, such as (), or ((())()). There is only one Dyck word of length 2, (). There are 2 of length 4, (()) and ()(). Similarly, there are 5 Dyck words of length 5. Let C_n be the number of Dyck words of length $2n$. We can split each Dyck word into $(w_1)w_2$, where w_1 and w_2 are Dyck words. Since the lengths of w_1 and w_2 must sum up to $2(n-1)$,

$$C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}. \qquad (*)$$

We again use pgf-like functions: let

$$c(x) = \sum_{n=0}^{\infty} C_n x^n.$$

From $(*)$, we can show that

$$c(x) = 1 + xc(x)^2.$$

We can solve to show that

$$c(x) = \frac{1-\sqrt{1-4x}}{2x} = \sum_{0}^{\infty} \binom{2n}{n} \frac{x^n}{n+1},$$

noting that $C_0 = 1$. Then

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

– `3.5 Probability generating functions`

– `3 Discrete random variables`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 148`

Example. Let $X_1, X_2, \cdots, X_n$ be iid with pgf $p(z) = \mathbb{E}[z^X]$. Let N be a random variable independent of X_i with pgf $h(z)$. What is the pgf of $S = X_1 + \cdots + X_N$?

$$\begin{aligned}
\mathbb{E}[z^S] &= \mathbb{E}[z^{X_1+\cdots+X_N}]\\
&= \mathbb{E}_N[\underbrace{\mathbb{E}_{X_i}[z^{X_1+\ldots+X_N} \mid N]}_{\text{assuming fixed } N}]\\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n)\mathbb{E}[z^{X_1+X_2+\cdots+X_n}]\\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n)\mathbb{E}[z^{X_1}]\mathbb{E}[z^{X_2}]\cdots\mathbb{E}[z^{X_n}]\\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n)(\mathbb{E}[z^{X_1}])^n\\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n)p(z)^n\\
&= h(p(z))
\end{aligned}$$

since $h(x) = \sum_{n=0}^{\infty} \mathbb{P}(N=n)x^n$. So

$$\begin{aligned}
\mathbb{E}[S] &= \left.\frac{\mathrm{d}}{\mathrm{d}z}h(p(z))\right|_{z=1}\\
&= h'(p(1))p'(1)\\
&= \mathbb{E}[N]\mathbb{E}[X_1]
\end{aligned}$$

To calculate the variance, use the fact that

$$\mathbb{E}[S(S-1)] = \left.\frac{\mathrm{d}^2}{\mathrm{d}z^2}h(p(z))\right|_{z=1}.$$

Then we can find that

$$\operatorname{var}(S) = \mathbb{E}[N]\operatorname{var}(X_1) + \mathbb{E}[X_1^2]\operatorname{var}(N).$$

Example. Let $X_1, X_2, \cdots, X_n$ be iid with pgf $p(z) = \mathbb{E}[z^X]$. Let N be a random variable independent of X_i with pgf $h(z)$. What is the pgf of $S = X_1 + \cdots + X_N$?

$$\begin{aligned}
\mathbb{E}[z^S] &= \mathbb{E}[z^{X_1+\cdots+X_N}]\\
&= \mathbb{E}_N[\underbrace{\mathbb{E}_{X_i}[z^{X_1+\ldots+X_N} \mid N]}_{\text{assuming fixed } N}]\\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n)\mathbb{E}[z^{X_1+X_2+\cdots+X_n}]\\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n)\mathbb{E}[z^{X_1}]\mathbb{E}[z^{X_2}]\cdots\mathbb{E}[z^{X_n}]\\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n)(\mathbb{E}[z^{X_1}])^n\\
&= \sum_{n=0}^{\infty} \mathbb{P}(N=n)p(z)^n\\
&= h(p(z))
\end{aligned}$$

since $h(x) = \sum_{n=0}^{\infty} \mathbb{P}(N=n)x^n$. So

$$\begin{aligned}
\mathbb{E}[S] &= \left.\frac{\mathrm{d}}{\mathrm{d}z}h(p(z))\right|_{z=1}\\
&= h'(p(1))p'(1)\\
&= \mathbb{E}[N]\mathbb{E}[X_1]
\end{aligned}$$

To calculate the variance, use the fact that

$$\mathbb{E}[S(S-1)] = \left.\frac{\mathrm{d}^2}{\mathrm{d}z^2}h(p(z))\right|_{z=1}.$$

Then we can find that

$$\operatorname{var}(S) = \mathbb{E}[N]\operatorname{var}(X_1) + \mathbb{E}[X_1^2]\operatorname{var}(N).$$

– `3.5 Probability generating functions`

– `3 Discrete random variables`

–

– `GENERAL KNOWLEDGE`

5 Interesting problems

5.1 Branching processes

NOTE:

– `probability.tex 149`

Theorem.

$$F_{n+1}(z) = F_n(F(z)) = F(F(F(\cdots F(z)\cdots)))) = F(F_n(z)).$$

Theorem.

$$F_{n+1}(z) = F_n(F(z)) = F(F(F(\cdots F(z)\cdots)))) = F(F_n(z)).$$

– `4.1 Branching processes`

– `4 Interesting problems`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 150`

Theorem.

$$F_{n+1}(z) = F_n(F(z)) = F(F(F(\cdots F(z)\cdots)))) = F(F_n(z)).$$

Proof.

$$\begin{aligned}
F_{n+1}(z) &= \mathbb{E}[z^{X_{n+1}}]\\
&= \mathbb{E}[\mathbb{E}[z^{X_{n+1}} \mid X_n]]\\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_n=k)\mathbb{E}[z^{X_{n+1}} \mid X_n = k]\\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_n=k)\mathbb{E}[z^{Y_1^n+\cdots+Y_k^n} \mid X_n = k]\\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_n=k)\mathbb{E}[z^{Y_1}]\mathbb{E}[z^{Y_2}]\cdots\mathbb{E}[z^{Y_n}]\\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_n=k)(\mathbb{E}[z^{X_1}])^k\\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_n=k)F(z)^k\\
&= F_n(F(z))
\end{aligned}$$

□

– `4.1 Branching processes`

– `4 Interesting problems`

–

– `PROOF EXERCISE`

NOTE:

– `probability.tex 151`

Theorem. Suppose

$$\mathbb{E}[X_1] = \sum kp_k = \mu$$

and

$$\operatorname{var}(X_1) = \mathbb{E}[(X-\mu)^2] = \sum(k-\mu)^2 p_k < \infty.$$

Then

$$\mathbb{E}[X_n] = \mu^n, \quad \operatorname{var} X_n = \sigma^2\mu^{n-1}(1+\mu+\mu^2+\cdots+\mu^{n-1}).$$

Theorem. Suppose

$$\mathbb{E}[X_1] = \sum kp_k = \mu$$

and

$$\operatorname{var}(X_1) = \mathbb{E}[(X-\mu)^2] = \sum(k-\mu)^2 p_k < \infty.$$

Then

$$\mathbb{E}[X_n] = \mu^n, \quad \operatorname{var} X_n = \sigma^2\mu^{n-1}(1+\mu+\mu^2+\cdots+\mu^{n-1}).$$

– `4.1 Branching processes`

– `4 Interesting problems`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 152`

Theorem. Suppose

$$\mathbb{E}[X_1] = \sum kp_k = \mu$$

and

$$\operatorname{var}(X_1) = \mathbb{E}[(X-\mu)^2] = \sum(k-\mu)^2 p_k < \infty.$$

Then

$$\mathbb{E}[X_n] = \mu^n, \quad \operatorname{var} X_n = \sigma^2\mu^{n-1}(1+\mu+\mu^2+\cdots+\mu^{n-1}).$$

Proof.

$$\begin{aligned}
\mathbb{E}[X_n] &= \mathbb{E}[\mathbb{E}[X_n \mid X_{n-1}]]\\
&= \mathbb{E}[\mu X_{n-1}]\\
&= \mu\mathbb{E}[X_{n-1}]
\end{aligned}$$

Then by induction, $\mathbb{E}[X_n] = \mu^n$ (since $X_0 = 1$). To calculate the variance, note that

$$\operatorname{var}(X_n) = \mathbb{E}[X_n^2] - (\mathbb{E}[X_n])^2$$

and hence

$$\mathbb{E}[X_n^2] = \operatorname{var}(X_n) + (\mathbb{E}[X])^2$$

We then calculate

$$\begin{aligned}
\mathbb{E}[X_n^2] &= \mathbb{E}[\mathbb{E}[X_n^2 \mid X_{n-1}]]\\
&= \mathbb{E}[\operatorname{var}(X_n) + (\mathbb{E}[X_n])^2 \mid X_{n-1}]\\
&= \mathbb{E}[X_{n-1}\operatorname{var}(X_1) + (\mu X_{n-1})^2]\\
&= \mathbb{E}[X_{n-1}\sigma^2 + (\mu X_{n-1})^2]\\
&= \sigma^2\mu^{n-1} + \mu^2\mathbb{E}[X_{n-1}^2].
\end{aligned}$$

So

$$\begin{aligned}
\operatorname{var} X_n &= \mathbb{E}[X_n^2] - (\mathbb{E}[X_n])^2\\
&= \mu^2\mathbb{E}[X_{n-1}^2] + \sigma^2\mu^{n-1} - \mu^2(\mathbb{E}[X_{n-1}])^2\\
&= \mu^2(\mathbb{E}[X_{n-1}^2] - \mathbb{E}[X_{n-1}]^2) + \sigma^2\mu^{n-1}\\
&= \mu^2\operatorname{var}(X_{n-1}) + \sigma^2\mu^{n-1}\\
&= \mu^4\operatorname{var}(X_{n-2}) + \sigma^2(\mu^{n-1}+\mu^n)\\
&= \cdots\\
&= \mu^{2(n-1)}\operatorname{var}(X_1) + \sigma^2(\mu^{n-1}+\mu^n+\cdots+\mu^{2n-3})\\
&= \sigma^2\mu^{n-1}(1+\mu+\cdots+\mu^{n-1}).
\end{aligned}$$

Of course, we can also obtain this using the probability generating function as well. □

– `4.1 Branching processes`

– `4 Interesting problems`

–

– `PROOF EXERCISE`

NOTE:

- `probability.tex 153`

Theorem. The probability of extinction q is the smallest root to the equation $q = F(q)$. Write $\mu = \mathbb{E}[X_1]$. Then if $\mu \leq 1$, then $q = 1$; if $\mu > 1$, then $q < 1$.

Theorem. The probability of extinction q is the smallest root to the equation $q = F(q)$. Write $\mu = \mathbb{E}[X_1]$. Then if $\mu \leq 1$, then $q = 1$; if $\mu > 1$, then $q < 1$.

- `4.1 Branching processes`
- `4 Interesting problems`
-
- `GENERAL KNOWLEDGE`

NOTE:

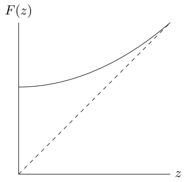
- `probability.tex 154`

Theorem. The probability of extinction q is the smallest root to the equation $q = F(q)$. Write $\mu = \mathbb{E}[X_1]$. Then if $\mu \leq 1$, then $q = 1$; if $\mu > 1$, then $q < 1$.

Proof. To show that it is the smallest root, let α be the smallest root. Then note that $0 \leq \alpha \Rightarrow F(0) \leq F(\alpha) = \alpha$ since F is increasing (proof: write the function out!). Hence $F(F(0)) \leq \alpha$. Continuing inductively, $F_n(0) \leq \alpha$ for all n. So

$$q = \lim_{n\to\infty} F_n(0) \leq \alpha.$$

So $q = \alpha$. To show that $q = 1$ when $\mu \leq 1$, we show that $q = 1$ is the only root. We know that $F'(z), F''(z) \geq 0$ for $z \in (0, 1)$ (proof: write it out again!). So F is increasing and convex. Since $F'(1) = \mu \leq 1$, it must approach $(1, 1)$ from above the $F = z$ line. So it must look like this:



So $z = 1$ is the only root. □

- `4.1 Branching processes`
- `4 Interesting problems`
-
- `PROOF EXERCISE`

5.2 Random walk and gambler's ruin

NOTE:

- `probability.tex 155`

Random walk

Definition (Random walk)**.** Let $X_1, \cdots, X_n$ be iid random variables such that $X_n = +1$ with probability p, and -1 with probability $1 - p$. Let $S_n = S_0 + X_1 + \cdots + X_n$. Then $(S_0, S_1, \cdots, S_n)$ is a *1-dimensional random walk*. If $p = q = \frac{1}{2}$, we say it is a *symmetric random walk*.

- `4.2 Random walk and gambler's ruin`
- `4 Interesting problems`
-
- `VOCABULARY`

NOTE:

- `probability.tex 156`

Example. A gambler starts with \$$z$, with $z < a$, and plays a game in which he wins \$1 or loses \$1 at each turn with probabilities p and q respectively. What are

$$p_z = \mathbb{P}(\text{random walk hits } a \text{ before } 0 \mid \text{starts at } z),$$

and

$$q_z = \mathbb{P}(\text{random walk hits } 0 \text{ before } a \mid \text{starts at } z)?$$

He either wins his first game, with probability p, or loses with probability q. So

$$p_z = qp_{z-1} + pp_{z+1},$$

for $0 < z < a$, and $p_0 = 0, p_a = 1$. Try $p_z = t^z$. Then

$$pt^2 - t + q = (pt - q)(t - 1) = 0,$$

noting that $p = 1 - q$. If $p \neq q$, then

$$p_z = A1^z + B\left(\frac{q}{p}\right)^z.$$

Since $p_0 = 0$, we get $A = -B$. Since $p_a = 1$, we obtain

$$p_z = \frac{1 - (q/p)^z}{1 - (q/p)^a}.$$

If $p = q$, then $p_z = A + Bz = z/a$. Similarly, (or perform the substitutions $p \mapsto q$, $q \mapsto p$ and $z \mapsto a - z$)

$$q_z = \frac{(q/p)^a - (q/p)^z}{(q/p)^a - 1}$$

if $p \neq q$, and

$$q_z = \frac{a - z}{a}$$

if $p = q$. Since $p_z + q_z = 1$, we know that the game will eventually end. What if $a \to \infty$? What is the probability of going bankrupt?

$$\begin{aligned}
\mathbb{P}(\text{path hits 0 ever}) &= \mathbb{P}\left(\bigcup_{a=z+1}^{\infty} [\text{path hits 0 before } a]\right)\\
&= \lim_{a\to\infty} \mathbb{P}(\text{path hits 0 before } a)\\
&= \lim_{a\to\infty} q_z\\
&= \begin{cases} (q/p)^z & p > q \\ 1 & p \leq q. \end{cases}
\end{aligned}$$

So if the odds are against you (i.e. the probability of losing is greater than the probability of winning), then no matter how small the difference is, you are bound to going bankrupt eventually.

Example. A gambler starts with \$$z$, with $z < a$, and plays a game in which he wins \$1 or loses \$1 at each turn with probabilities p and q respectively. What are

$$p_z = \mathbb{P}(\text{random walk hits } a \text{ before } 0 \mid \text{starts at } z),$$

and

$$q_z = \mathbb{P}(\text{random walk hits } 0 \text{ before } a \mid \text{starts at } z)?$$

He either wins his first game, with probability p, or loses with probability q. So

$$p_z = qp_{z-1} + pp_{z+1},$$

for $0 < z < a$, and $p_0 = 0, p_a = 1$. Try $p_z = t^z$. Then

$$pt^2 - t + q = (pt - q)(t - 1) = 0,$$

noting that $p = 1 - q$. If $p \neq q$, then

$$p_z = A1^z + B\left(\frac{q}{p}\right)^z.$$

Since $p_0 = 0$, we get $A = -B$. Since $p_a = 1$, we obtain

$$p_z = \frac{1 - (q/p)^z}{1 - (q/p)^a}.$$

If $p = q$, then $p_z = A + Bz = z/a$. Similarly, (or perform the substitutions $p \mapsto q$, $q \mapsto p$ and $z \mapsto a - z$)

$$q_z = \frac{(q/p)^a - (q/p)^z}{(q/p)^a - 1}$$

if $p \neq q$, and

$$q_z = \frac{a - z}{a}$$

if $p = q$. Since $p_z + q_z = 1$, we know that the game will eventually end. What if $a \to \infty$? What is the probability of going bankrupt?

$$\begin{aligned}
\mathbb{P}(\text{path hits 0 ever}) &= \mathbb{P}\left(\bigcup_{a=z+1}^{\infty} [\text{path hits 0 before } a]\right)\\
&= \lim_{a\to\infty} \mathbb{P}(\text{path hits 0 before } a)\\
&= \lim_{a\to\infty} q_z\\
&= \begin{cases} (q/p)^z & p > q \\ 1 & p \leq q. \end{cases}
\end{aligned}$$

So if the odds are against you (i.e. the probability of losing is greater than the probability of winning), then no matter how small the difference is, you are bound to going bankrupt eventually.

- `4.2 Random walk and gambler's ruin`
- `4 Interesting problems`
-
- `GENERAL KNOWLEDGE`

6 Continuous random variables

6.1 Continuous random variables

NOTE:

- `probability.tex 157`

Definition (Continuous random variable)**.** A random variable $X : \Omega \to \mathbb{R}$ is *continuous* if there is a function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that

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

We call f the *probability density function*, which satisfies

- $f \geq 0$
- $\int_{-\infty}^{\infty} f(x) = 1$.

Definition (Continuous random variable)**.** A random variable $X : \Omega \to \mathbb{R}$ is *continuous* if there is a function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that

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

We call f the *probability density function*, which satisfies

- $f \geq 0$
- $\int_{-\infty}^{\infty} f(x) = 1$.

- `5.1 Continuous random variables`
- `5 Continuous random variables`
-
- `VOCABULARY`

NOTE:

- `probability.tex 158`

Cumulative distribution function

Definition (Cumulative distribution function)**.** The *cumulative distribution function* (or simply *distribution function*) of a random variable X (discrete, continuous, or neither) is

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

- `5.1 Continuous random variables`
- `5 Continuous random variables`
-
- `VOCABULARY`

NOTE:

- `probability.tex 159`

Definition (Uniform distribution)**.** The *uniform distribution* on $[a, b]$ has pdf

$$f(x) = \frac{1}{b - a}.$$

Then

$$F(x) = \int_a^x f(z)\,\mathrm{d}z = \frac{x - a}{b - a}$$

for $a \leq x \leq b$. If X follows a uniform distribution on $[a, b]$, we write $X \sim U[a, b]$.

Definition (Uniform distribution)**.** The *uniform distribution* on $[a, b]$ has pdf

$$f(x) = \frac{1}{b - a}.$$

Then

$$F(x) = \int_a^x f(z)\,\mathrm{d}z = \frac{x - a}{b - a}$$

for $a \leq x \leq b$. If X follows a uniform distribution on $[a, b]$, we write $X \sim U[a, b]$.

- `5.1 Continuous random variables`
- `5 Continuous random variables`
-
- `VOCABULARY`

NOTE:

- probability.tex 160

Definition (Exponential random variable)**.** The *exponential random variable with parameter* λ has pdf

$$f(x) = \lambda e^{-\lambda x}$$

and

$$F(x) = 1 - e^{-\lambda x}$$

for $x \geq 0$. We write $X \sim \mathcal{E}(\lambda)$.

Definition (Exponential random variable)**.** The *exponential random variable with parameter* λ has pdf

$$f(x) = \lambda e^{-\lambda x}$$

and

$$F(x) = 1 - e^{-\lambda x}$$

for $x \geq 0$. We write $X \sim \mathcal{E}(\lambda)$.

- 5.1 Continuous random variables
- 5 Continuous random variables
-
- VOCABULARY

NOTE:

- probability.tex 161

Proposition. The exponential random variable is *memoryless*, i.e.

$$\mathbb{P}(X \geq x + z \mid X \geq x) = \mathbb{P}(X \geq z).$$

This means that, say if X measures the lifetime of a light bulb, knowing it has already lasted for 3 hours does not give any information about how much longer it will last.

Proposition. The exponential random variable is *memoryless*, i.e.

$$\mathbb{P}(X \geq x + z \mid X \geq x) = \mathbb{P}(X \geq z).$$

This means that, say if X measures the lifetime of a light bulb, knowing it has already lasted for 3 hours does not give any information about how much longer it will last.

- 5.1 Continuous random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 162

Proposition. The exponential random variable is *memoryless*, i.e.

$$\mathbb{P}(X \geq x + z \mid X \geq x) = \mathbb{P}(X \geq z).$$

This means that, say if X measures the lifetime of a light bulb, knowing it has already lasted for 3 hours does not give any information about how much longer it will last.

Proof.

$$\begin{aligned}\mathbb{P}(X \geq x + z \mid X \geq x) &= \frac{\mathbb{P}(X \geq x+z)}{\mathbb{P}(X \geq x)}\\ &= \frac{\int_{x+z}^\infty f(u)\;\mathrm{d}u}{\int_x^\infty f(u)\;\mathrm{d}u}\\ &= \frac{e^{-\lambda(x+z)}}{e^{-\lambda x}}\\ &= e^{-\lambda z}\\ &= \mathbb{P}(X \geq z).\end{aligned}$$ □

- 5.1 Continuous random variables
- 5 Continuous random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 163

Expectation

Definition (Expectation)**.** The *expectation* (or *mean*) of a continuous random variable is

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

provided not both $\int_0^\infty xf(x)\;\mathrm{d}x$ and $\int_{-\infty}^0 xf(x)\;\mathrm{d}x$ are infinite.

- 5.1 Continuous random variables
- 5 Continuous random variables
-
- VOCABULARY

NOTE:

- probability.tex 164

Theorem. If X is a continuous random variable, then

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

Theorem. If X is a continuous random variable, then

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

- 5.1 Continuous random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 165

Theorem. If X is a continuous random variable, then

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

Proof.

$$\begin{aligned}\int_0^\infty \mathbb{P}(X \geq x)\;\mathrm{d}x &= \int_0^\infty \int_x^\infty f(y)\;\mathrm{d}y\;\mathrm{d}x\\ &= \int_0^\infty \int_0^\infty I[y \geq x] f(y)\;\mathrm{d}y\;\mathrm{d}x\\ &= \int_0^\infty \left(\int_0^\infty I[x \leq y]\;\mathrm{d}x\right) f(y)\;\mathrm{d}y\\ &= \int_0^\infty yf(y)\;\mathrm{d}y.\end{aligned}$$

We can similarly show that $\int_0^\infty \mathbb{P}(X \leq -x)\;\mathrm{d}x = -\int_{-\infty}^0 yf(y)\;\mathrm{d}y$. □

- 5.1 Continuous random variables
- 5 Continuous random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 166

Example. Suppose $X \sim \mathcal{E}(\lambda)$. Then

$$\mathbb{P}(X \geq x) = \int_x^\infty \lambda e^{-\lambda t}\;\mathrm{d}t = e^{-\lambda x}.$$

So

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

Example. Suppose $X \sim \mathcal{E}(\lambda)$. Then

$$\mathbb{P}(X \geq x) = \int_x^\infty \lambda e^{-\lambda t}\;\mathrm{d}t = e^{-\lambda x}.$$

So

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

- 5.1 Continuous random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 167

Variance

Definition (Variance)**.** The *variance* of a continuous random variable is

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

- 5.1 Continuous random variables
- 5 Continuous random variables
-
- VOCABULARY

NOTE:

– `probability.tex 168`

Example. Let $X \sim U[a, b]$. Then
$$\mathbb{E}[X] = \int_a^b x \frac{1}{b-a}\, dx = \frac{a+b}{2}.$$
So
$$\begin{aligned}\operatorname{var}(X) &= \int_a^b x^2 \frac{1}{b-a}\, \mathrm{d}x - (\mathbb{E}[X])^2\\ &= \frac{1}{12}(b-a)^2.\end{aligned}$$

Example. Let $X \sim U[a, b]$. Then
$$\mathbb{E}[X] = \int_a^b x \frac{1}{b-a}\, dx = \frac{a+b}{2}.$$
So
$$\begin{aligned}\operatorname{var}(X) &= \int_a^b x^2 \frac{1}{b-a}\, \mathrm{d}x - (\mathbb{E}[X])^2\\ &= \frac{1}{12}(b-a)^2.\end{aligned}$$

– `5.1 Continuous random variables`

– `5 Continuous random variables`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 169`

Definition (Mode and median)**.** Given a pdf $f(x)$, we call $\hat{x}$ a *mode* if
$$f(\hat{x}) \geq f(x)$$
for all x. Note that a distribution can have many modes. For example, in the uniform distribution, all x are modes. We say it is a median if
$$\int_{-\infty}^{\hat{x}} f(x)\, \mathrm{d}x = \frac{1}{2} = \int_{\hat{x}}^{\infty} f(x)\, \mathrm{d}x.$$
For a discrete random variable, the median is $\hat{x}$ such that
$$\mathbb{P}(X \leq \hat{x}) \geq \frac{1}{2},\quad \mathbb{P}(X \geq \hat{x}) \geq \frac{1}{2}.$$
Here we have a non-strict inequality since if the random variable, say, always takes value 0, then both probabilities will be 1.

Definition (Mode and median)**.** Given a pdf $f(x)$, we call $\hat{x}$ a *mode* if
$$f(\hat{x}) \geq f(x)$$
for all x. Note that a distribution can have many modes. For example, in the uniform distribution, all x are modes. We say it is a median if
$$\int_{-\infty}^{\hat{x}} f(x)\, \mathrm{d}x = \frac{1}{2} = \int_{\hat{x}}^{\infty} f(x)\, \mathrm{d}x.$$
For a discrete random variable, the median is $\hat{x}$ such that
$$\mathbb{P}(X \leq \hat{x}) \geq \frac{1}{2},\quad \mathbb{P}(X \geq \hat{x}) \geq \frac{1}{2}.$$
Here we have a non-strict inequality since if the random variable, say, always takes value 0, then both probabilities will be 1.

– `5.1 Continuous random variables`

– `5 Continuous random variables`

–

– `VOCABULARY`

NOTE:

– `probability.tex 170`

Sample mean

Definition (Sample mean)**.** If $X_1, \cdots, X_n$ is a random sample from some distribution, then the *sample mean* is
$$\bar{X} = \frac{1}{n}\sum_1^n X_i.$$

– `5.1 Continuous random variables`

– `5 Continuous random variables`

–

– `VOCABULARY`

6.2 Stochastic ordering and inspection paradox

NOTE:

– `probability.tex 171`

Stochastic order

Definition (Stochastic order)**.** The *stochastic order* is defined as: $X \geq_{\text{st}} Y$ if $\mathbb{P}(X > t) \geq \mathbb{P}(Y > t)$ for all t.

– `5.2 Stochastic ordering and inspection paradox`

– `5 Continuous random variables`

–

– `VOCABULARY`

NOTE:

– `probability.tex 172`

Example (Inspection paradox)**.** Suppose that n families have children attending a school. Family i has X_i children at the school, where $X_1, \cdots, X_n$ are iid random variables, with $P(X_i = k) = p_k$. Suppose that the average family size is μ. Now pick a child at random. What is the probability distribution of his family size? Let J be the index of the family from which she comes (which is a random variable). Then
$$\mathbb{P}(X_J = k \mid J = j) = \frac{\mathbb{P}(J = j, X_j = k)}{\mathbb{P}(J = j)}.$$
The denominator is $1/n$. The numerator is more complex. This would require the jth family to have k members, which happens with probability p_k; and that we picked a member from the jth family, which happens with probability $\mathbb{E}\left[\frac{k}{k + \sum_{i \neq j} X_i}\right]$. So
$$\mathbb{P}(X_J = k \mid J = j) = \mathbb{E}\left[\frac{nkp_k}{k + \sum_{i\neq j} X_i}\right].$$
Note that this is independent of j. So
$$\mathbb{P}(X_J = k) = \mathbb{E}\left[\frac{nkp_k}{k + \sum_{i\neq j} X_i}\right].$$
Also, $\mathbb{P}(X_1 = k) = p_k$. So
$$\frac{\mathbb{P}(X_J = k)}{\mathbb{P}(X_1 = k)} = \mathbb{E}\left[\frac{nk}{k + \sum_{i\neq j} X_i}\right].$$
This is increasing in k, and greater than 1 for $k > \mu$. So the average value of the family size of the child we picked is greater than the average family size. It can be shown that X_J is stochastically greater than X_1. This means that if we pick the children randomly, the sample mean of the family size will be greater than the actual mean. This is since for the larger a family is, the more likely it is for us to pick a child from the family.

Example (Inspection paradox)**.** Suppose that n families have children attending a school. Family i has X_i children at the school, where $X_1, \cdots, X_n$ are iid random variables, with $P(X_i = k) = p_k$. Suppose that the average family size is μ. Now pick a child at random. What is the probability distribution of his family size? Let J be the index of the family from which she comes (which is a random variable). Then
$$\mathbb{P}(X_J = k \mid J = j) = \frac{\mathbb{P}(J = j, X_j = k)}{\mathbb{P}(J = j)}.$$
The denominator is $1/n$. The numerator is more complex. This would require the jth family to have k members, which happens with probability p_k; and that we picked a member from the jth family, which happens with probability $\mathbb{E}\left[\frac{k}{k + \sum_{i \neq j} X_i}\right]$. So
$$\mathbb{P}(X_J = k \mid J = j) = \mathbb{E}\left[\frac{nkp_k}{k + \sum_{i\neq j} X_i}\right].$$
Note that this is independent of j. So
$$\mathbb{P}(X_J = k) = \mathbb{E}\left[\frac{nkp_k}{k + \sum_{i\neq j} X_i}\right].$$
Also, $\mathbb{P}(X_1 = k) = p_k$. So
$$\frac{\mathbb{P}(X_J = k)}{\mathbb{P}(X_1 = k)} = \mathbb{E}\left[\frac{nk}{k + \sum_{i\neq j} X_i}\right].$$
This is increasing in k, and greater than 1 for $k > \mu$. So the average value of the family size of the child we picked is greater than the average family size. It can be shown that X_J is stochastically greater than X_1. This means that if we pick the children randomly, the sample mean of the family size will be greater than the actual mean. This is since for the larger a family is, the more likely it is for us to pick a child from the family.

– `5.2 Stochastic ordering and inspection paradox`

– `5 Continuous random variables`

–

– `GENERAL KNOWLEDGE`

6.3 Jointly distributed random variables

NOTE:

– `probability.tex 173`

Joint distribution

Definition (Joint distribution)**.** Two random variables X, Y have *joint distribution* $F : \mathbb{R}^2 \mapsto [0, 1]$ defined by
$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$
The *marginal distribution* of X is
$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq x, Y < \infty) = F(x, \infty) = \lim_{y\to\infty} F(x, y)$$

– `5.3 Jointly distributed random variables`

– `5 Continuous random variables`

–

– `VOCABULARY`

NOTE:

– `probability.tex 174`

Definition (Jointly distributed random variables)**.** We say $X_1, \cdots, X_n$ are *jointly distributed continuous random variables* and have *joint pdf* f if for any set $A \subseteq \mathbb{R}^n$
$$\mathbb{P}((X_1, \cdots, X_n) \in A) = \int_{(x_1,\cdots x_n)\in A} f(x_1, \cdots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$
where
$$f(x_1, \cdots, x_n) \geq 0$$
and
$$\int_{\mathbb{R}^n} f(x_1, \cdots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = 1.$$

Definition (Jointly distributed random variables)**.** We say $X_1, \cdots, X_n$ are *jointly distributed continuous random variables* and have *joint pdf* f if for any set $A \subseteq \mathbb{R}^n$
$$\mathbb{P}((X_1, \cdots, X_n) \in A) = \int_{(x_1,\cdots x_n)\in A} f(x_1, \cdots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$
where
$$f(x_1, \cdots, x_n) \geq 0$$
and
$$\int_{\mathbb{R}^n} f(x_1, \cdots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = 1.$$

– `5.3 Jointly distributed random variables`

– `5 Continuous random variables`

–

– `VOCABULARY`

NOTE:

- probability.tex 175

Example. In the case where $n = 2$,

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)\ \mathrm{d}x\ \mathrm{d}y.$$

If F is differentiable, then

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

Example. In the case where $n = 2$,

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)\ \mathrm{d}x\ \mathrm{d}y.$$

If F is differentiable, then

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

- 5.3 Jointly distributed random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 176

Theorem. If X and Y are jointly continuous random variables, then they are individually continuous random variables.

Theorem. If X and Y are jointly continuous random variables, then they are individually continuous random variables.

- 5.3 Jointly distributed random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 177

Theorem. If X and Y are jointly continuous random variables, then they are individually continuous random variables.

Proof. We prove this by showing that X has a density function. We know that

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(X \in A, Y \in (-\infty, +\infty)) \\ &= \int_{x \in A} \int_{-\infty}^{\infty} f(x, y)\ \mathrm{d}y\ \mathrm{d}x \\ &= \int_{x \in A} f_X(x)\ \mathrm{d}x\end{aligned}$$

So

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

is the (marginal) pdf of X. $\square$

- 5.3 Jointly distributed random variables
- 5 Continuous random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 178

Definition (Independent continuous random variables)**.** Continuous random variables $X_1, \cdots, X_n$ are independent if

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2)\cdots$$

for all $A_i \subseteq \Omega_{X_i}$. If we let F_{X_i} and f_{X_i} be the cdf, pdf of X, then

$$F(x_1, \cdots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

and

$$f(x_1, \cdots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

are each individually equivalent to the definition above.

Definition (Independent continuous random variables)**.** Continuous random variables $X_1, \cdots, X_n$ are independent if

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2)\cdots$$

for all $A_i \subseteq \Omega_{X_i}$. If we let F_{X_i} and f_{X_i} be the cdf, pdf of X, then

$$F(x_1, \cdots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

and

$$f(x_1, \cdots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

are each individually equivalent to the definition above.

- 5.3 Jointly distributed random variables
- 5 Continuous random variables
-
- VOCABULARY

NOTE:

- probability.tex 179

Example. If (X_1, X_2) takes a random value from $[0, 1] \times [0, 1]$, then $f(x_1, x_2) = 1$. Then we can see that $f(x_1, x_2) = 1 \cdot 1 = f(x_1) \cdot f(x_2)$. So X_1 and X_2 are independent. On the other hand, if (Y_1, Y_2) takes a random value from $[0, 1] \times [0, 1]$ with the restriction that $Y_1 \leq Y_2$, then they are not independent, since $f(x_1, x_2) = 2I[Y_1 \leq Y_2]$, which cannot be split into two parts.

Example. If (X_1, X_2) takes a random value from $[0, 1] \times [0, 1]$, then $f(x_1, x_2) = 1$. Then we can see that $f(x_1, x_2) = 1 \cdot 1 = f(x_1) \cdot f(x_2)$. So X_1 and X_2 are independent. On the other hand, if (Y_1, Y_2) takes a random value from $[0, 1] \times [0, 1]$ with the restriction that $Y_1 \leq Y_2$, then they are not independent, since $f(x_1, x_2) = 2I[Y_1 \leq Y_2]$, which cannot be split into two parts.

- 5.3 Jointly distributed random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 180

Proposition. For independent continuous random variables X_i,

(i) $\mathbb{E}[\prod X_i] = \prod \mathbb{E}[X_i]$

(ii) $\mathrm{var}(\sum X_i) = \sum \mathrm{var}(X_i)$

Proposition. For independent continuous random variables X_i,

(i) $\mathbb{E}[\prod X_i] = \prod \mathbb{E}[X_i]$

(ii) $\mathrm{var}(\sum X_i) = \sum \mathrm{var}(X_i)$

- 5.3 Jointly distributed random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

6.4 Geometric probability

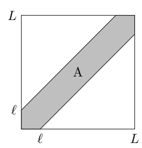
NOTE:

- probability.tex 181

Example. Two points X and Y are chosen independently on a line segment of length L. What is the probability that $|X - Y| \leq \ell$? By "at random", we mean

$$f(x, y) = \frac{1}{L^2},$$

since each of X and Y have pdf $1/L$. We can visualize this on a graph:



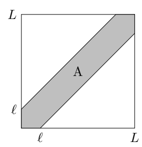
Here the two axes are the values of X and Y, and A is the permitted region. The total area of the white part is simply the area of a square with length $L - \ell$. So the area of A is $L^2 - (L - \ell)^2 = 2L\ell - \ell^2$. So the desired probability is

$$\int_A f(x, y)\ \mathrm{d}x\ \mathrm{d}y = \frac{2L\ell - \ell^2}{L^2}.$$

Example. Two points X and Y are chosen independently on a line segment of length L. What is the probability that $|X - Y| \leq \ell$? By "at random", we mean

$$f(x, y) = \frac{1}{L^2},$$

since each of X and Y have pdf $1/L$. We can visualize this on a graph:



Here the two axes are the values of X and Y, and A is the permitted region. The total area of the white part is simply the area of a square with length $L - \ell$. So the area of A is $L^2 - (L - \ell)^2 = 2L\ell - \ell^2$. So the desired probability is

$$\int_A f(x, y)\ \mathrm{d}x\ \mathrm{d}y = \frac{2L\ell - \ell^2}{L^2}.$$

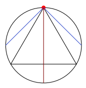
- 5.4 Geometric probability
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

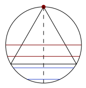
- probability.tex 182

Example (Bertrand's paradox). Suppose we draw a random chord in a circle. What is the probability that the length of the chord is greater than the length of the side of an inscribed equilateral triangle? There are three ways of "drawing a random chord".

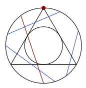
(i) We randomly pick two end points over the circumference independently. Now draw the inscribed triangle with the vertex at one end point. For the length of the chord to be longer than a side of the triangle, the other end point must between the two other vertices of the triangle. This happens with probability 1/3.



(ii) wlog the chord is horizontal, on the lower side of the circle. The mid-point is uniformly distributed along the middle (dashed) line. Then the probability of getting a long line is 1/2.



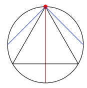
(iii) The mid point of the chord is distributed uniformly across the circle. Then you get a long line if and only if the mid-point lies in the smaller circle shown below. This occurs with probability 1/4.



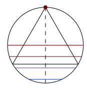
We get different answers for different notions of "random"! This is why when we say "randomly", we should be explicit in what we mean by that.

Example (Bertrand's paradox). Suppose we draw a random chord in a circle. What is the probability that the length of the chord is greater than the length of the side of an inscribed equilateral triangle? There are three ways of "drawing a random chord".

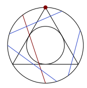
(i) We randomly pick two end points over the circumference independently. Now draw the inscribed triangle with the vertex at one end point. For the length of the chord to be longer than a side of the triangle, the other end point must between the two other vertices of the triangle. This happens with probability 1/3.



(ii) wlog the chord is horizontal, on the lower side of the circle. The mid-point is uniformly distributed along the middle (dashed) line. Then the probability of getting a long line is 1/2.



(iii) The mid point of the chord is distributed uniformly across the circle. Then you get a long line if and only if the mid-point lies in the smaller circle shown below. This occurs with probability 1/4.



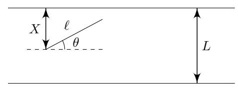
We get different answers for different notions of "random"! This is why when we say "randomly", we should be explicit in what we mean by that.

- 5.4 Geometric probability
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 183

Example (Buffon's needle). A needle of length ℓ is tossed at random onto a floor marked with parallel lines a distance L apart, where $\ell \leq L$. Let A be the event that the needle intersects a line. What is $\mathbb{P}(A)$?



Suppose that $X \sim U[0, L]$ and $\theta \sim U[0, \pi]$. Then

$$f(x, \theta) = \frac{1}{L\pi}.$$

This touches a line iff $X \leq \ell \sin \theta$. Hence

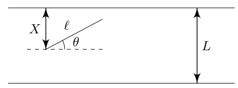
$$\mathbb{P}(A) = \int_{\theta=0}^{\pi} \underbrace{\frac{\ell \sin \theta}{L}}_{\mathbb{P}(X \leq \ell \sin \theta)} \frac{1}{\pi}\,\mathrm{d}\theta = \frac{2\ell}{\pi L}.$$

Since the answer involves π, we can estimate π by conducting repeated experiments! Suppose we have N hits out of n tosses. Then an estimator for p is $\hat{p} = \frac{N}{n}$. Hence

$$\hat{\pi} = \frac{2\ell}{\hat{p}L}.$$

We will later find out that this is a really inefficient way of estimating π (as you could have guessed).

Example (Buffon's needle). A needle of length ℓ is tossed at random onto a floor marked with parallel lines a distance L apart, where $\ell \leq L$. Let A be the event that the needle intersects a line. What is $\mathbb{P}(A)$?



Suppose that $X \sim U[0, L]$ and $\theta \sim U[0, \pi]$. Then

$$f(x, \theta) = \frac{1}{L\pi}.$$

This touches a line iff $X \leq \ell \sin \theta$. Hence

$$\mathbb{P}(A) = \int_{\theta=0}^{\pi} \underbrace{\frac{\ell \sin \theta}{L}}_{\mathbb{P}(X \leq \ell \sin \theta)} \frac{1}{\pi}\,\mathrm{d}\theta = \frac{2\ell}{\pi L}.$$

Since the answer involves π, we can estimate π by conducting repeated experiments! Suppose we have N hits out of n tosses. Then an estimator for p is $\hat{p} = \frac{N}{n}$. Hence

$$\hat{\pi} = \frac{2\ell}{\hat{p}L}.$$

We will later find out that this is a really inefficient way of estimating π (as you could have guessed).

- 5.4 Geometric probability
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

6.5 The normal distribution

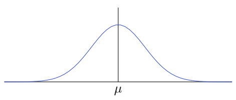
NOTE:

- probability.tex 184

Definition (Normal distribution). The *normal distribution* with parameters μ, σ^2, written $N(\mu, \sigma^2)$ has pdf

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

for $-\infty < x < \infty$. It looks approximately like this:

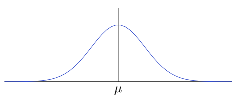


The *standard normal* is when $\mu = 0, \sigma^2 = 1$, i.e. $X \sim N(0, 1)$. We usually write $\phi(x)$ for the pdf and $\Phi(x)$ for the cdf of the standard normal.

Definition (Normal distribution). The *normal distribution* with parameters μ, σ^2, written $N(\mu, \sigma^2)$ has pdf

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

for $-\infty < x < \infty$. It looks approximately like this:



The *standard normal* is when $\mu = 0, \sigma^2 = 1$, i.e. $X \sim N(0, 1)$. We usually write $\phi(x)$ for the pdf and $\Phi(x)$ for the cdf of the standard normal.

- 5.5 The normal distribution
- 5 Continuous random variables
-
- VOCABULARY

NOTE:

- probability.tex 185

Proposition.

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,\mathrm{d}x = 1.$$

Proposition.

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,\mathrm{d}x = 1.$$

- 5.5 The normal distribution
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 186

Proposition.

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,\mathrm{d}x = 1.$$

Proof. Substitute $z = \frac{(x-\mu)}{\sigma}$. Then

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,\mathrm{d}z.$$

Then

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x \int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,\mathrm{d}y \\ &= \int_0^{\infty} \int_0^{2\pi} \frac{1}{2\pi} e^{-r^2/2} r\,\mathrm{d}r\,\mathrm{d}\theta \\ &= 1. \end{aligned}$$

□

- 5.5 The normal distribution
- 5 Continuous random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 187

Proposition. $\mathbb{E}[X] = \mu$.

Proposition. $\mathbb{E}[X] = \mu$.

- 5.5 The normal distribution
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 188

Proposition. $\mathbb{E}[X] = \mu$.

Proof.

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} xe^{-(x-\mu)^2/2\sigma^2}\,\mathrm{d}x\\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(x-\mu)e^{-(x-\mu)^2/2\sigma^2}\,\mathrm{d}x + \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\mu e^{-(x-\mu)^2/2\sigma^2}\,\mathrm{d}x.\end{aligned}$$

The first term is antisymmetric about μ and gives 0. The second is just μ times the integral we did above. So we get μ. □

- 5.5 The normal distribution
- 5 Continuous random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 189

Proposition. $\operatorname{var}(X) = \sigma^2$.

Proposition. $\operatorname{var}(X) = \sigma^2$.

- 5.5 The normal distribution
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 190

Proposition. $\operatorname{var}(X) = \sigma^2$.

Proof. We have $\operatorname{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Substitute $Z = \frac{X-\mu}{\sigma}$. Then $\mathbb{E}[Z] = 0$, $\mathbb{E}[Z^2] = \frac{1}{\sigma^2}\mathbb{E}[X^2]$. Then

$$\begin{aligned}\operatorname{var}(Z) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2 e^{-z^2/2}\,\mathrm{d}z\\ &= \left[-\frac{1}{\sqrt{2\pi}}ze^{-z^2/2}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}\,\mathrm{d}z\\ &= 0 + 1\\ &= 1\end{aligned}$$

So $\operatorname{var} X = \sigma^2$. □

- 5.5 The normal distribution
- 5 Continuous random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 191

Example. UK adult male heights are normally distributed with mean 70" and standard deviation 3". In the Netherlands, these figures are 71" and 3". What is $\mathbb{P}(Y > X)$, where X and Y are the heights of randomly chosen UK and Netherlands males, respectively? We have $X \sim N(70, 3^2)$ and $Y \sim N(71, 3^2)$. Then (as we will show in later lectures) $Y - X \sim N(1, 18)$.

$$\mathbb{P}(Y > X) = \mathbb{P}(Y - X > 0) = \mathbb{P}\left(\frac{Y - X - 1}{\sqrt{18}} > \frac{-1}{\sqrt{18}}\right) = 1 - \Phi(-$$

since $\frac{(Y-X)-1}{\sqrt{18}} \sim N(0,1)$, and the answer is approximately 0.5931. Now suppose that in both countries, the Olympic male basketball teams are selected from that portion of male whose hight is at least above 4" above the mean (which corresponds to the 9.1% tallest males of the country). What is the probability that a randomly chosen Netherlands player is taller than a randomly chosen UK player? For the second part, we have

$$\mathbb{P}(Y > X \mid X \geq 74, Y \geq 75) = \frac{\int_{x=74}^{75}\phi_X(x)\,\mathrm{d}x + \int_{x=75}^{\infty}\int_{y=x}^{\infty}\phi_Y(}{\int_{x=74}^{\infty}\phi_X(x)\,\mathrm{d}x\int_{y=75}^{\infty}\phi_Y}$$

which is approximately 0.7558. So even though the Netherlands people are only slightly taller, if we consider the tallest bunch, the Netherlands people will be much taller on average.

Example. UK adult male heights are normally distributed with mean 70" and standard deviation 3". In the Netherlands, these figures are 71" and 3". What is $\mathbb{P}(Y > X)$, where X and Y are the heights of randomly chosen UK and Netherlands males, respectively? We have $X \sim N(70, 3^2)$ and $Y \sim N(71, 3^2)$. Then (as we will show in later lectures) $Y - X \sim N(1, 18)$.

$$\mathbb{P}(Y > X) = \mathbb{P}(Y - X > 0) = \mathbb{P}\left(\frac{Y - X - 1}{\sqrt{18}} > \frac{-1}{\sqrt{18}}\right) = 1 - \Phi(-$$

since $\frac{(Y-X)-1}{\sqrt{18}} \sim N(0,1)$, and the answer is approximately 0.5931. Now suppose that in both countries, the Olympic male basketball teams are selected from that portion of male whose hight is at least above 4" above the mean (which corresponds to the 9.1% tallest males of the country). What is the probability that a randomly chosen Netherlands player is taller than a randomly chosen UK player? For the second part, we have

$$\mathbb{P}(Y > X \mid X \geq 74, Y \geq 75) = \frac{\int_{x=74}^{75}\phi_X(x)\,\mathrm{d}x + \int_{x=75}^{\infty}\int_{y=x}^{\infty}\phi_Y(}{\int_{x=74}^{\infty}\phi_X(x)\,\mathrm{d}x\int_{y=75}^{\infty}\phi_Y}$$

which is approximately 0.7558. So even though the Netherlands people are only slightly taller, if we consider the tallest bunch, the Netherlands people will be much taller on average.

- 5.5 The normal distribution
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

6.6 Transformation of random variables

NOTE:

- probability.tex 192

Theorem. If X is a continuous random variable with a pdf $f(x)$, and $h(x)$ is a continuous, strictly increasing function with $h^{-1}(x)$ differentiable, then $Y = h(X)$ is a random variable with pdf

$$f_Y(y) = f_X(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y).$$

Theorem. If X is a continuous random variable with a pdf $f(x)$, and $h(x)$ is a continuous, strictly increasing function with $h^{-1}(x)$ differentiable, then $Y = h(X)$ is a random variable with pdf

$$f_Y(y) = f_X(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y).$$

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 193

Theorem. If X is a continuous random variable with a pdf $f(x)$, and $h(x)$ is a continuous, strictly increasing function with $h^{-1}(x)$ differentiable, then $Y = h(X)$ is a random variable with pdf

$$f_Y(y) = f_X(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y).$$

Proof.

$$\begin{aligned}F_Y(y) &= \mathbb{P}(Y \leq y)\\ &= \mathbb{P}(h(X) \leq y)\\ &= \mathbb{P}(X \leq h^{-1}(y))\\ &= F(h^{-1}(y))\end{aligned}$$

Take the derivative with respect to y to obtain

$$f_Y(y) = F'_Y(y) = f(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y).$$ □

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 194

Example. Let $X \sim U[0, 1]$. Let $Y = -\log X$. Then

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(-\log X \leq y)\\ &= \mathbb{P}(X \geq e^{-y})\\ &= 1 - e^{-y}.\end{aligned}$$

But this is the cumulative distribution function of $\mathcal{E}(1)$. So Y is exponentially distributed with $\lambda = 1$.

Example. Let $X \sim U[0, 1]$. Let $Y = -\log X$. Then

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(-\log X \leq y)\\ &= \mathbb{P}(X \geq e^{-y})\\ &= 1 - e^{-y}.\end{aligned}$$

But this is the cumulative distribution function of $\mathcal{E}(1)$. So Y is exponentially distributed with $\lambda = 1$.

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 195

Theorem. Let $U \sim U[0,1]$. For any strictly increasing distribution function F, the random variable $X = F^{-1}U$ has distribution function F.

Theorem. Let $U \sim U[0,1]$. For any strictly increasing distribution function F, the random variable $X = F^{-1}U$ has distribution function F.

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 196

Theorem. Let $U \sim U[0,1]$. For any strictly increasing distribution function F, the random variable $X = F^{-1}U$ has distribution function F.

Proof.

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

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- PROOF EXERCISE

NOTE:

- probability.tex 197

Definition (Jacobian determinant)**.** Suppose $\frac{\partial s_i}{\partial y_j}$ exists and is continuous at every point $(y_1, \cdots, y_n) \in S$. Then the *Jacobian determinant* is

$$J = \frac{\partial(s_1, \cdots, s_n)}{\partial(y_1, \cdots, y_n)} = \det \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{pmatrix}$$

Definition (Jacobian determinant)**.** Suppose $\frac{\partial s_i}{\partial y_j}$ exists and is continuous at every point $(y_1, \cdots, y_n) \in S$. Then the *Jacobian determinant* is

$$J = \frac{\partial(s_1, \cdots, s_n)}{\partial(y_1, \cdots, y_n)} = \det \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{pmatrix}$$

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- VOCABULARY

NOTE:

- probability.tex 198

Proposition. $(Y_1, \cdots, Y_n)$ has density

$$g(y_1, \cdots, y_n) = f(s_1(y_1, \cdots, y_n), \cdots s_n(y_1, \cdots, y_n))|J|$$

if $(y_1, \cdots, y_n) \in S$, 0 otherwise.

Proposition. $(Y_1, \cdots, Y_n)$ has density

$$g(y_1, \cdots, y_n) = f(s_1(y_1, \cdots, y_n), \cdots s_n(y_1, \cdots, y_n))|J|$$

if $(y_1, \cdots, y_n) \in S$, 0 otherwise.

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 199

Example. Suppose (X, Y) has density

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

We see that X and Y are independent, with each having a density $f(x) = 2x$. Define $U = X/Y$, $V = XY$. Then we have $X = \sqrt{UV}$ and $Y = \sqrt{V/U}$. The Jacobian is

$$\det \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix} = \det \begin{pmatrix} \frac{1}{2}\sqrt{v/u} & \frac{1}{2}\sqrt{u/v} \\ -\frac{1}{2}\sqrt{v/u^3} & \frac{1}{2}\sqrt{1/uv} \end{pmatrix} = \frac{1}{2u}$$

Alternatively, we can find this by considering

$$\det \begin{pmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial u/\partial y \end{pmatrix} = 2u,$$

and then inverting the matrix. So

$$g(u,v) = 4\sqrt{uv}\sqrt{\frac{v}{u}}\frac{1}{2u} = \frac{2v}{u},$$

if (u, v) is in the image S, 0 otherwise. So

$$g(u,v) = \frac{2v}{u} I[(u,v) \in S].$$

Since this is not separable, we know that U and V are not independent.

Example. Suppose (X, Y) has density

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

We see that X and Y are independent, with each having a density $f(x) = 2x$. Define $U = X/Y$, $V = XY$. Then we have $X = \sqrt{UV}$ and $Y = \sqrt{V/U}$. The Jacobian is

$$\det \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix} = \det \begin{pmatrix} \frac{1}{2}\sqrt{v/u} & \frac{1}{2}\sqrt{u/v} \\ -\frac{1}{2}\sqrt{v/u^3} & \frac{1}{2}\sqrt{1/uv} \end{pmatrix} = \frac{1}{2u}$$

Alternatively, we can find this by considering

$$\det \begin{pmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial u/\partial y \end{pmatrix} = 2u,$$

and then inverting the matrix. So

$$g(u,v) = 4\sqrt{uv}\sqrt{\frac{v}{u}}\frac{1}{2u} = \frac{2v}{u},$$

if (u, v) is in the image S, 0 otherwise. So

$$g(u,v) = \frac{2v}{u} I[(u,v) \in S].$$

Since this is not separable, we know that U and V are not independent.

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 200

Example. Suppose X_1, X_2 have joint pdf $f(x_1, x_2)$. Suppose we want to find the pdf of $Y = X_1 + X_2$. We let $Z = X_2$. Then $X_1 = Y - Z$ and $X_2 = Z$. Then

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = A\mathbf{X}$$

Then $|J| = 1/|\det A| = 1$. Then

$$g(y,z) = f(y - z, z)$$

So

$$g_Y(y) = \int_{-\infty}^{\infty} f(y - z, z)\,\mathrm{d}z = \int_{-\infty}^{\infty} f(z, y - z)\,\mathrm{d}z.$$

If X_1 and X_2 are independent, $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. Then

$$g(y) = \int_{-\infty}^{\infty} f_1(z)f_2(y - z)\,\mathrm{d}z.$$

Example. Suppose X_1, X_2 have joint pdf $f(x_1, x_2)$. Suppose we want to find the pdf of $Y = X_1 + X_2$. We let $Z = X_2$. Then $X_1 = Y - Z$ and $X_2 = Z$. Then

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = A\mathbf{X}$$

Then $|J| = 1/|\det A| = 1$. Then

$$g(y,z) = f(y - z, z)$$

So

$$g_Y(y) = \int_{-\infty}^{\infty} f(y - z, z)\,\mathrm{d}z = \int_{-\infty}^{\infty} f(z, y - z)\,\mathrm{d}z.$$

If X_1 and X_2 are independent, $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. Then

$$g(y) = \int_{-\infty}^{\infty} f_1(z)f_2(y - z)\,\mathrm{d}z.$$

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 201

Example. Suppose X has pdf f. What is the pdf of $Y = |X|$? We use our definition. We have

$$\mathbb{P}(|X| \in (a,b)) = \int_a^b f(x) + \int_{-b}^{-a} f(x)\,\mathrm{d}x = \int_a^b (f(x) + f(-x))\,\mathrm{d}x$$

So

$$f_Y(x) = f(x) + f(-x),$$

which makes sense, since getting $|X| = x$ is equivalent to getting $X = x$ or $X = -x$.

Example. Suppose X has pdf f. What is the pdf of $Y = |X|$? We use our definition. We have

$$\mathbb{P}(|X| \in (a,b)) = \int_a^b f(x) + \int_{-b}^{-a} f(x)\,\mathrm{d}x = \int_a^b (f(x) + f(-x))\,\mathrm{d}x$$

So

$$f_Y(x) = f(x) + f(-x),$$

which makes sense, since getting $|X| = x$ is equivalent to getting $X = x$ or $X = -x$.

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 202

Example. Suppose $X_1 \sim \mathcal{E}(\lambda), X_2 \sim \mathcal{E}(\mu)$ are independent random variables. Let $Y = \min(X_1, X_2)$. Then

$$\begin{aligned}\mathbb{P}(Y \geq t) &= \mathbb{P}(X_1 \geq t, X_2 \geq t)\\ &= \mathbb{P}(X_1 \geq t)\mathbb{P}(X_2 \geq t)\\ &= e^{-\lambda t}e^{-\mu t}\\ &= e^{-(\lambda+\mu)t}.\end{aligned}$$

So $Y \sim \mathcal{E}(\lambda + \mu)$.

Example. Suppose $X_1 \sim \mathcal{E}(\lambda), X_2 \sim \mathcal{E}(\mu)$ are independent random variables. Let $Y = \min(X_1, X_2)$. Then

$$\begin{aligned}\mathbb{P}(Y \geq t) &= \mathbb{P}(X_1 \geq t, X_2 \geq t)\\ &= \mathbb{P}(X_1 \geq t)\mathbb{P}(X_2 \geq t)\\ &= e^{-\lambda t}e^{-\mu t}\\ &= e^{-(\lambda+\mu)t}.\end{aligned}$$

So $Y \sim \mathcal{E}(\lambda + \mu)$.

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 203

Order statistics

Definition (Order statistics)**.** Suppose that $X_1, \cdots, X_n$ are some random variables, and $Y_1, \cdots, Y_n$ is $X_1, \cdots, X_n$ arranged in increasing order, i.e. $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. This is the *order statistics*. We sometimes write $Y_i = X_{(i)}$.

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- VOCABULARY

NOTE:

- probability.tex 204

Example. Let $X_1, \cdots, X_n$ be iid $\mathcal{E}(\lambda)$, and $Y_1, \cdots, Y_n$ be the order statistic. Let

$$\begin{aligned}Z_1 &= Y_1\\ Z_2 &= Y_2 - Y_1\\ &\vdots\\ Z_n &= Y_n - Y_{n-1}.\end{aligned}$$

These are the distances between the occurrences. We can write this as a $\mathbf{Z} = A\mathbf{Y}$, with

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0\\ -1 & 1 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\end{pmatrix}$$

Then $\det(A) = 1$ and hence $|J| = 1$. Suppose that the pdf of $Z_1, \cdots, Z_n$ is, say h. Then

$$\begin{aligned}h(z_1, \cdots, z_n) &= g(y_1, \cdots, y_n)\cdot 1\\ &= n!f(y_1)\cdots f(y_n)\\ &= n!\lambda^n e^{-\lambda(y_1 + \cdots + y_n)}\\ &= n!\lambda^n e^{-\lambda(nz_1 + (n-1)z_2 + \cdots + z_n)}\\ &= \prod_{i=1}^n (\lambda i)e^{-(\lambda i)z_{n+1-i}}\end{aligned}$$

Since h is expressed as a product of n density functions, we have

$$Z_i \sim \mathcal{E}((n+1-i)\lambda).$$

with all Z_i independent.

Example. Let $X_1, \cdots, X_n$ be iid $\mathcal{E}(\lambda)$, and $Y_1, \cdots, Y_n$ be the order statistic. Let

$$\begin{aligned}Z_1 &= Y_1\\ Z_2 &= Y_2 - Y_1\\ &\vdots\\ Z_n &= Y_n - Y_{n-1}.\end{aligned}$$

These are the distances between the occurrences. We can write this as a $\mathbf{Z} = A\mathbf{Y}$, with

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0\\ -1 & 1 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\end{pmatrix}$$

Then $\det(A) = 1$ and hence $|J| = 1$. Suppose that the pdf of $Z_1, \cdots, Z_n$ is, say h. Then

$$\begin{aligned}h(z_1, \cdots, z_n) &= g(y_1, \cdots, y_n)\cdot 1\\ &= n!f(y_1)\cdots f(y_n)\\ &= n!\lambda^n e^{-\lambda(y_1 + \cdots + y_n)}\\ &= n!\lambda^n e^{-\lambda(nz_1 + (n-1)z_2 + \cdots + z_n)}\\ &= \prod_{i=1}^n (\lambda i)e^{-(\lambda i)z_{n+1-i}}\end{aligned}$$

Since h is expressed as a product of n density functions, we have

$$Z_i \sim \mathcal{E}((n+1-i)\lambda).$$

with all Z_i independent.

- 5.6 Transformation of random variables
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

6.7 Moment generating functions

NOTE:

- probability.tex 205

Moment generating function

Definition (Moment generating function)**.** The *moment generating function* of a random variable X is

$$m(\theta) = \mathbb{E}[e^{\theta X}].$$

For those θ in which $m(\theta)$ is finite, we have

$$m(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\,\mathrm{d}x.$$

- 5.7 Moment generating functions
- 5 Continuous random variables
-
- VOCABULARY

NOTE:

- probability.tex 206

Theorem. The mgf determines the distribution of X provided $m(\theta)$ is finite for all θ in some interval containing the origin.

Theorem. The mgf determines the distribution of X provided $m(\theta)$ is finite for all θ in some interval containing the origin.

- 5.7 Moment generating functions
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 207

Moment

Definition (Moment)**.** The rth *moment* of X is $\mathbb{E}[X^r]$.

- 5.7 Moment generating functions
- 5 Continuous random variables
-
- VOCABULARY

NOTE:

- probability.tex 208

Theorem. The rth moment X is the coefficient of $\frac{\theta^r}{r!}$ in the power series expansion of $m(\theta)$, and is

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

Theorem. The rth moment X is the coefficient of $\frac{\theta^r}{r!}$ in the power series expansion of $m(\theta)$, and is

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

- 5.7 Moment generating functions
- 5 Continuous random variables
-
- GENERAL KNOWLEDGE

NOTE:

– probability.tex 209

Theorem. The rth moment X is the coefficient of $\frac{\theta^r}{r!}$ in the power series expansion of $m(\theta)$, and is

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

Proof. We have

$$e^{\theta X} = 1 + \theta X + \frac{\theta^2}{2!}X^2 + \cdots.$$

So

$$m(\theta) = \mathbb{E}[e^{\theta X}] = 1 + \theta\mathbb{E}[X] + \frac{\theta^2}{2!}\mathbb{E}[X^2] + \cdots \qquad \square$$

– 5.7 Moment generating functions

– 5 Continuous random variables

–

– PROOF EXERCISE

NOTE:

– probability.tex 210

Example. Let $X \sim \mathcal{E}(\lambda)$. Then its mgf is

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

where $0 < \theta < \lambda$. So

$$\mathbb{E}[X] = m'(0) = \frac{\lambda}{(\lambda-\theta)^2}\bigg|_{\theta=0} = \frac{1}{\lambda}.$$

Also,

$$\mathbb{E}[X^2] = m''(0) = \frac{2\lambda}{(\lambda-\theta)^3}\bigg|_{\theta=0} = \frac{2}{\lambda^2}.$$

So

$$\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Example. Let $X \sim \mathcal{E}(\lambda)$. Then its mgf is

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

where $0 < \theta < \lambda$. So

$$\mathbb{E}[X] = m'(0) = \frac{\lambda}{(\lambda-\theta)^2}\bigg|_{\theta=0} = \frac{1}{\lambda}.$$

Also,

$$\mathbb{E}[X^2] = m''(0) = \frac{2\lambda}{(\lambda-\theta)^3}\bigg|_{\theta=0} = \frac{2}{\lambda^2}.$$

So

$$\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

– 5.7 Moment generating functions

– 5 Continuous random variables

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 211

Theorem. If X and Y are independent random variables with moment generating functions $m_X(\theta), m_Y(\theta)$, then $X+Y$ has mgf $m_{X+Y}(\theta) = m_X(\theta)m_Y(\theta)$.

Theorem. If X and Y are independent random variables with moment generating functions $m_X(\theta), m_Y(\theta)$, then $X+Y$ has mgf $m_{X+Y}(\theta) = m_X(\theta)m_Y(\theta)$.

– 5.7 Moment generating functions

– 5 Continuous random variables

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 212

Theorem. If X and Y are independent random variables with moment generating functions $m_X(\theta), m_Y(\theta)$, then $X+Y$ has mgf $m_{X+Y}(\theta) = m_X(\theta)m_Y(\theta)$.

Proof.

$$\mathbb{E}[e^{\theta(X+Y)}] = \mathbb{E}[e^{\theta X}e^{\theta Y}] = \mathbb{E}[e^{\theta X}]\mathbb{E}[e^{\theta Y}] = m_X(\theta)m_Y(\theta).$$

$\square$

– 5.7 Moment generating functions

– 5 Continuous random variables

–

– PROOF EXERCISE

7 More distributions

7.1 Cauchy distribution

NOTE:

– probability.tex 213

Cauchy distribution

Definition (Cauchy distribution)**.** The *Cauchy distribution* has pdf

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

for $-\infty < x < \infty$.

– 6.1 Cauchy distribution

– 6 More distributions

–

– VOCABULARY

NOTE:

– probability.tex 214

Proposition. The mean of the Cauchy distribution is undefined, while $\mathbb{E}[X^2] = \infty$.

Proposition. The mean of the Cauchy distribution is undefined, while $\mathbb{E}[X^2] = \infty$.

– 6.1 Cauchy distribution

– 6 More distributions

–

– GENERAL KNOWLEDGE

NOTE:

– probability.tex 215

Proposition. The mean of the Cauchy distribution is undefined, while $\mathbb{E}[X^2] = \infty$.

Proof.

$$\mathbb{E}[X] = \int_0^\infty \frac{x}{\pi(1+x^2)}\,\mathrm{d}x + \int_{-\infty}^0 \frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \infty - \infty$$

which is undefined, but $\mathbb{E}[X^2] = \infty + \infty = \infty$. $\square$

– 6.1 Cauchy distribution

– 6 More distributions

–

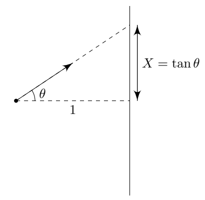
– PROOF EXERCISE

NOTE:

– probability.tex 216

Example.

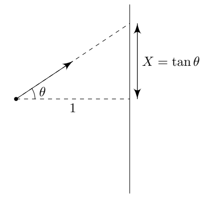
(i) If $\Theta \sim U[-\frac{\pi}{2}, \frac{\pi}{2}]$, then $X = \tan\theta$ has a Cauchy distribution. For example, if we fire a bullet at a wall 1 meter apart at a random random angle θ, the vertical displacement follows a Cauchy distribution.



(ii) If $X, Y \sim N(0,1)$, then X/Y has a Cauchy distribution.

Example.

(i) If $\Theta \sim U[-\frac{\pi}{2}, \frac{\pi}{2}]$, then $X = \tan\theta$ has a Cauchy distribution. For example, if we fire a bullet at a wall 1 meter apart at a random random angle θ, the vertical displacement follows a Cauchy distribution.



(ii) If $X, Y \sim N(0,1)$, then X/Y has a Cauchy distribution.

– 6.1 Cauchy distribution

– 6 More distributions

–

– GENERAL KNOWLEDGE

7.2 Gamma distribution

NOTE:

- probability.tex 217

Gamma distribution

Definition (Gamma distribution)**.** The *gamma distribution* $\Gamma(n, \lambda)$ has pdf

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

We can show that this is a distribution by showing that it integrates to 1.

- 6.2 Gamma distribution
- 6 More distributions
-
- VOCABULARY

7.3 Beta distribution*

NOTE:

- probability.tex 218

Beta distribution

Definition (Beta distribution)**.** The *beta distribution* $\beta(a, b)$ has pdf

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

for $0 \leq x \leq 1$. This has mean $a/(a+b)$.

- 6.3 Beta distribution*
- 6 More distributions
-
- VOCABULARY

7.4 More on the normal distribution

NOTE:

- probability.tex 219

Proposition. The moment generating function of $N(\mu, \sigma^2)$ is

$$\mathbb{E}[e^{\theta X}] = \exp\left(\theta\mu + \frac{1}{2}\theta^2\sigma^2\right).$$

Proposition. The moment generating function of $N(\mu, \sigma^2)$ is

$$\mathbb{E}[e^{\theta X}] = \exp\left(\theta\mu + \frac{1}{2}\theta^2\sigma^2\right).$$

- 6.4 More on the normal distribution
- 6 More distributions
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 220

Proposition. The moment generating function of $N(\mu, \sigma^2)$ is

$$\mathbb{E}[e^{\theta X}] = \exp\left(\theta\mu + \frac{1}{2}\theta^2\sigma^2\right).$$

Proof.

$$\mathbb{E}[e^{\theta X}] = \int_{-\infty}^{\infty} e^{\theta x}\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\,\mathrm{d}x.$$

Substitute $z = \frac{x-\mu}{\sigma}$. Then

$$\begin{aligned}\mathbb{E}[e^{\theta X}] &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}} e^{\theta(\mu+\sigma z)} e^{-\frac{1}{2}z^2}\,\mathrm{d}z\\ &= e^{\theta\mu + \frac{1}{2}\theta^2\sigma^2}\int_{-\infty}^{\infty}\underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\theta\sigma)^2}}_{\text{pdf of } N(\sigma\theta, 1)}\,\mathrm{d}z\\ &= e^{\theta\mu+\frac{1}{2}\theta^2\sigma^2}.\end{aligned}$$

□

- 6.4 More on the normal distribution
- 6 More distributions
-
- PROOF EXERCISE

NOTE:

- probability.tex 221

Theorem. Suppose X, Y are independent random variables with $X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim (\mu_2, \sigma_2^2)$. Then

(i) $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(ii) $aX \sim N(a\mu_1, a^2\sigma_1^2)$.

Theorem. Suppose X, Y are independent random variables with $X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim (\mu_2, \sigma_2^2)$. Then

(i) $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(ii) $aX \sim N(a\mu_1, a^2\sigma_1^2)$.

- 6.4 More on the normal distribution
- 6 More distributions
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 222

Theorem. Suppose X, Y are independent random variables with $X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim (\mu_2, \sigma_2^2)$. Then

(i) $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(ii) $aX \sim N(a\mu_1, a^2\sigma_1^2)$.

Proof.

(i)

$$\begin{aligned}\mathbb{E}[e^{\theta(X+Y)}] &= \mathbb{E}[e^{\theta X}]\cdot\mathbb{E}[e^{\theta Y}]\\ &= e^{\mu_1\theta + \frac{1}{2}\sigma_1^2\theta^2}\cdot e^{\mu_2\theta+\frac{1}{2}\sigma_2^2\theta}\\ &= e^{(\mu_1+\mu_2)\theta + \frac{1}{2}(\sigma_1^2+\sigma_2^2)\theta^2}\end{aligned}$$

which is the mgf of $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(ii)

$$\begin{aligned}\mathbb{E}[e^{\theta(aX)}] &= \mathbb{E}[e^{(\theta a)X}]\\ &= e^{\mu(a\theta)+\frac{1}{2}\sigma^2(a\theta)^2}\\ &= e^{(a\mu)\theta + \frac{1}{2}(a^2\sigma^2)\theta^2}\end{aligned}$$

□

- 6.4 More on the normal distribution
- 6 More distributions
-
- PROOF EXERCISE

7.5 Multivariate normal

8 Central limit theorem

NOTE:

- probability.tex 223

Theorem (Central limit theorem)**.** Let $X_1, X_2, \cdots$ be iid random variables with $\mathbb{E}[X_i] = \mu$, $\mathrm{var}(X_i) = \sigma^2 < \infty$. Define

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

Then for all finite intervals (a, b),

$$\lim_{n\to\infty}\mathbb{P}\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,\mathrm{d}t.$$

Note that the final term is the pdf of a standard normal. We say

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to_D N(0, 1).$$

Theorem (Central limit theorem)**.** Let $X_1, X_2, \cdots$ be iid random variables with $\mathbb{E}[X_i] = \mu$, $\mathrm{var}(X_i) = \sigma^2 < \infty$. Define

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

Then for all finite intervals (a, b),

$$\lim_{n\to\infty}\mathbb{P}\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,\mathrm{d}t.$$

Note that the final term is the pdf of a standard normal. We say

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to_D N(0, 1).$$

- 6.5 Multivariate normal
- 7 Central limit theorem
-
- GENERAL KNOWLEDGE

NOTE:

- probability.tex 224

Continuity theorem

Theorem (Continuity theorem)**.** If the random variables $X_1, X_2, \cdots$ have mgf's $m_1(\theta), m_2(\theta), \cdots$ and $m_n(\theta) \to m(\theta)$ as $n \to \infty$ for all θ, then $X_n \to_D$ the random variable with mgf $m(\theta)$.

- 6.5 Multivariate normal
- 7 Central limit theorem
-
- GENERAL KNOWLEDGE

NOTE:

– `probability.tex 225`

Theorem (Continuity theorem)**.** If the random variables $X_1, X_2, \cdots$ have mgf's $m_1(\theta), m_2(\theta), \cdots$ and $m_n(\theta) \to m(\theta)$ as $n \to \infty$ for all θ, then $X_n \to_D$ the random variable with mgf $m(\theta)$.

Proof. wlog, assume $\mu = 0, \sigma^2 = 1$ (otherwise replace X_i with $\frac{X_i - \mu}{\sigma}$). Then

$$\begin{aligned} m_{X_i}(\theta) &= \mathbb{E}[e^{\theta X_i}] = 1 + \theta \mathbb{E}[X_i] + \frac{\theta^2}{2!}\mathbb{E}[X_i^2] + \cdots \\ &= 1 + \frac{1}{2}\theta^2 + \frac{1}{3!}\theta^3 \mathbb{E}[X_i^3] + \cdots \end{aligned}$$

Now consider $S_n/\sqrt{n}$. Then

$$\begin{aligned} \mathbb{E}[e^{\theta S_n/\sqrt{n}}] &= \mathbb{E}[e^{\theta(X_1 + \ldots + X_n)/\sqrt{n}}] \\ &= \mathbb{E}[e^{\theta X_1/\sqrt{n}}] \cdots \mathbb{E}[e^{\theta X_n/\sqrt{n}}] \\ &= \left(\mathbb{E}[e^{\theta X_1/\sqrt{n}}]\right)^n \\ &= \left(1 + \frac{1}{2}\theta^2 \frac{1}{n} + \frac{1}{3!}\theta^3 \mathbb{E}[X^3]\frac{1}{n^{3/2}} + \cdots\right)^n \\ &\to e^{\frac{1}{2}\theta^2} \end{aligned}$$

as $n \to \infty$ since $(1 + a/n)^n \to e^a$. And this is the mgf of the standard normal. So the result follows from the continuity theorem. □

– `6.5 Multivariate normal`

– `7 Central limit theorem`

–

– `PROOF EXERCISE`

NOTE:

– `probability.tex 226`

Example. Suppose two planes fly a route. Each of n passengers chooses a plane at random. The number of people choosing plane 1 is $S \sim B(n, \frac{1}{2})$. Suppose each has s seats. What is

$$F(s) = \mathbb{P}(S > s),$$

i.e. the probability that plane 1 is over-booked? We have

$$F(s) = \mathbb{P}(S > s) = \mathbb{P}\left(\frac{S - n/2}{\sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}}} > \frac{s - n/2}{\sqrt{n}/2}\right).$$

Since

$$\frac{S - np}{\sqrt{n}/2} \sim N(0, 1),$$

we have

$$F(s) \approx 1 - \Phi\left(\frac{s - n/2}{\sqrt{n}/2}\right).$$

For example, if $n = 1000$ and $s = 537$, then $\frac{S_n - n/2}{\sqrt{n}/2} \approx 2.34$, $\Phi(2.34) \approx 0.99$, and $F(s) \approx 0.01$. So with only 74 seats as buffer between the two planes, the probability of overbooking is just $1/100$.

Example. Suppose two planes fly a route. Each of n passengers chooses a plane at random. The number of people choosing plane 1 is $S \sim B(n, \frac{1}{2})$. Suppose each has s seats. What is

$$F(s) = \mathbb{P}(S > s),$$

i.e. the probability that plane 1 is over-booked? We have

$$F(s) = \mathbb{P}(S > s) = \mathbb{P}\left(\frac{S - n/2}{\sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}}} > \frac{s - n/2}{\sqrt{n}/2}\right).$$

Since

$$\frac{S - np}{\sqrt{n}/2} \sim N(0, 1),$$

we have

$$F(s) \approx 1 - \Phi\left(\frac{s - n/2}{\sqrt{n}/2}\right).$$

For example, if $n = 1000$ and $s = 537$, then $\frac{S_n - n/2}{\sqrt{n}/2} \approx 2.34$, $\Phi(2.34) \approx 0.99$, and $F(s) \approx 0.01$. So with only 74 seats as buffer between the two planes, the probability of overbooking is just $1/100$.

– `6.5 Multivariate normal`

– `7 Central limit theorem`

–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 227`

Example. An unknown proportion p of the electorate will vote Labour. It is desired to find p without an error not exceeding 0.005. How large should the sample be? We estimate by

$$p' = \frac{S_n}{n},$$

where $X_i \sim B(1, p)$. Then

$$\begin{aligned} \mathbb{P}(|p' - p| \leq 0.005) &= \mathbb{P}(|S_n - np| \leq 0.005n) \\ &= \mathbb{P}\left(\underbrace{\frac{|S_n - np|}{\sqrt{np(1-p)}}}_{\approx N(0,1)} \leq \frac{0.005n}{\sqrt{np(1-p)}}\right) \end{aligned}$$

We want $|p' - p| \leq 0.005$ with probability ≥ 0.95. Then we want

$$\frac{0.005n}{\sqrt{np(1-p)}} \geq \Phi^{-1}(0.975) = 1.96.$$

(we use 0.975 instead of 0.95 since we are doing a two-tailed test) Since the maximum possible value of $p(1-p)$ is $1/4$, we have

$$n \geq 38416.$$

In practice, we don't have that many samples. Instead, we go by

$$\mathbb{P}(|p' < p| \leq 0.03) \geq 0.95.$$

This just requires $n \geq 1068$.

Example. An unknown proportion p of the electorate will vote Labour. It is desired to find p without an error not exceeding 0.005. How large should the sample be? We estimate by

$$p' = \frac{S_n}{n},$$

where $X_i \sim B(1, p)$. Then

$$\begin{aligned} \mathbb{P}(|p' - p| \leq 0.005) &= \mathbb{P}(|S_n - np| \leq 0.005n) \\ &= \mathbb{P}\left(\underbrace{\frac{|S_n - np|}{\sqrt{np(1-p)}}}_{\approx N(0,1)} \leq \frac{0.005n}{\sqrt{np(1-p)}}\right) \end{aligned}$$

We want $|p' - p| \leq 0.005$ with probability ≥ 0.95. Then we want

$$\frac{0.005n}{\sqrt{np(1-p)}} \geq \Phi^{-1}(0.975) = 1.96.$$

(we use 0.975 instead of 0.95 since we are doing a two-tailed test) Since the maximum possible value of $p(1-p)$ is $1/4$, we have

$$n \geq 38416.$$

In practice, we don't have that many samples. Instead, we go by

$$\mathbb{P}(|p' < p| \leq 0.03) \geq 0.95.$$

This just requires $n \geq 1068$.

– `6.5 Multivariate normal`

– `7 Central limit theorem`

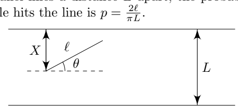
–

– `GENERAL KNOWLEDGE`

NOTE:

– `probability.tex 228`

Example (Estimating π with Buffon's needle)**.** Recall that if we randomly toss a needle of length ℓ to a floor marked with parallel lines a distance L apart, the probability that the needle hits the line is $p = \frac{2\ell}{\pi L}$.



Suppose we toss the pin n times, and it hits the line N times. Then

$$N \approx N(np, np(1-p))$$

by the Central limit theorem. Write p' for the actual proportion observed. Then

$$\begin{aligned} \hat{\pi} &= \frac{2\ell}{(N/n)L} \\ &= \frac{\pi 2\ell/(\pi L)}{p'} \\ &= \frac{\pi p}{p + (p' - p)} \\ &= \pi\left(1 - \frac{p' - p}{p} + \cdots\right) \end{aligned}$$

Hence

$$\hat{\pi} - \pi \approx \frac{p - p'}{p}.$$

We know

$$p' \sim N\left(p, \frac{p(1-p)}{n}\right).$$

So we can find

$$\hat{\pi} - \pi \sim N\left(0, \frac{\pi^2 p(1-p)}{np^2}\right) = N\left(0, \frac{\pi^2(1-p)}{np}\right)$$

We want a small variance, and that occurs when p is the largest. Since $p = 2\ell/\pi L$, this is maximized with $\ell = L$. In this case,

$$p = \frac{2}{\pi},$$

and

$$\hat{\pi} - \pi \approx N\left(0, \frac{(\pi - 2)\pi^2}{2n}\right).$$

If we want to estimate π to 3 decimal places, then we need

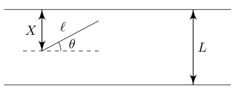
$$\mathbb{P}(|\hat{\pi} - \pi| \leq 0.001) \geq 0.95.$$

This is true if and only if

$$0.001\sqrt{\frac{2n}{(\pi - 2)(\pi^2)}} \geq \Phi^{-1}(0.975) = 1.96$$

So $n \geq 2.16 \times 10^7$. So we can obtain π to 3 decimal places just by throwing a stick 20 million times! Isn't that exciting?

Example (Estimating π with Buffon's needle)**.** Recall that if we randomly toss a needle of length ℓ to a floor marked with parallel lines a distance L apart, the probability that the needle hits the line is $p = \frac{2\ell}{\pi L}$.



Suppose we toss the pin n times, and it hits the line N times. Then

$$N \approx N(np, np(1-p))$$

by the Central limit theorem. Write p' for the actual proportion observed. Then

$$\begin{aligned} \hat{\pi} &= \frac{2\ell}{(N/n)L} \\ &= \frac{\pi 2\ell/(\pi L)}{p'} \\ &= \frac{\pi p}{p + (p' - p)} \\ &= \pi\left(1 - \frac{p' - p}{p} + \cdots\right) \end{aligned}$$

Hence

$$\hat{\pi} - \pi \approx \frac{p - p'}{p}.$$

We know

$$p' \sim N\left(p, \frac{p(1-p)}{n}\right).$$

So we can find

$$\hat{\pi} - \pi \sim N\left(0, \frac{\pi^2 p(1-p)}{np^2}\right) = N\left(0, \frac{\pi^2(1-p)}{np}\right)$$

We want a small variance, and that occurs when p is the largest. Since $p = 2\ell/\pi L$, this is maximized with $\ell = L$. In this case,

$$p = \frac{2}{\pi},$$

and

$$\hat{\pi} - \pi \approx N\left(0, \frac{(\pi - 2)\pi^2}{2n}\right).$$

If we want to estimate π to 3 decimal places, then we need

$$\mathbb{P}(|\hat{\pi} - \pi| \leq 0.001) \geq 0.95.$$

This is true if and only if

$$0.001\sqrt{\frac{2n}{(\pi - 2)(\pi^2)}} \geq \Phi^{-1}(0.975) = 1.96$$

So $n \geq 2.16 \times 10^7$. So we can obtain π to 3 decimal places just by throwing a stick 20 million times! Isn't that exciting?

– `6.5 Multivariate normal`

– `7 Central limit theorem`

–

– `GENERAL KNOWLEDGE`

9 Summary of distributions

9.1 Discrete distributions

9.2 Continuous distributions