Part IA — Probability

Based on lectures by Dr D. Yeo

Lent 2022

Contents

§0 Introduction

Text in blue is usually less important.

Example 0.1

Dice: outcomes $1, 2, \dots, 6$.

- $\mathbb{P}(2) = \frac{1}{6}$.
- $\mathbb{P}(\text{multiple of } 3) = \frac{2}{6} = \frac{1}{3}$.
- $\mathbb{P}(\text{not a multiple of } 3) = \frac{2}{3}$
- $\mathbb{P}(\text{prime}) = \frac{1}{2}$.
- $$\begin{aligned} \mathbb{P}(\text{prime or multiple of } 3) &= \cancel{\frac{1}{3} + \frac{1}{2} = \frac{5}{6}}. \\ &= \frac{4}{6} = \frac{2}{3}. \\ \mathbb{P}(\text{prime or multiple of } 3) &= \frac{1}{3} + \frac{1}{2} - \frac{1}{6} = \frac{2}{3} \end{aligned}$$

§1 Formal Setup

Definition 1.1 (Sample Space)
The **sample space** Ω is a set of outcomes.

Definition 1.2 (σ-algebra)

- Let $\mathcal{F}$ a collection of subsets of Ω (called *events*).
- $\mathcal{F}$ is a **σ-alegbra** if

  F1. $\Omega \in \mathcal{F}$.

  F2. $A \in \mathcal{F}$ then $A^c = \Omega \setminus A \in \mathcal{F}$.

  F3. $\forall$ countable collections $(A_n)_{n \in \mathbb{N}} \in \mathcal{F}$[a], the union $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}$ also.

[a] A_1 does not need to be countable, only the index

Remark 1. The motivation for F2 is so that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (the probability of not A is defined as expected).

Definition 1.3 (Probability Measure)
Given σ-algebra $\mathcal{F}$ on Ω, function $\mathbb{P} : \mathcal{F} \to [0,1]$[a] is a **probability measure** if

P2. $\mathbb{P}(\Omega) = 1$.

P3. $\forall$ countable collections $(A_n)_{n \in \mathbb{N}}$ of *disjoint* events in $\mathcal{F}$:

$$\mathbb{P}\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} \mathbb{P}(A_n).$$

Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a *probability space*.

[a] P1. $\mathbb{P}(A) \geq 0$

Example 1.1
Coming back to Example 0.1. $\Omega = \{1, 2, \ldots, 6\}$ so
$\mathbb{P}(\Omega) = \mathbb{P}(1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ or } 5 \text{ or } 6) = 1$ and $\mathcal{F}$ is all subsets of Ω.

Question

Why $\mathbb{P} : \mathcal{F} \to [0,1]$
and not $\mathbb{P} : \Omega \to [0,1]$?

If Ω is countable:

- In general: $\mathcal{F} =$ all subsets of Ω, i.e. $\mathcal{P}(\Omega)$ (the power set).
- $\mathbb{P}(2)$ is shorthand for $\mathbb{P}(\{2\})$.
- $\mathbb{P}$ is determined by $(\mathbb{P}(\{w\}),\ \forall\, w \in \Omega)$ (e.g. unfair dice).

If Ω is uncountable:

- E.g. $\Omega = [0,1]$. Want to choose a real number, all equally likely.
- If $\mathbb{P}\left(\{0\}\right) = \alpha > 0$ then $\mathbb{P}\left(\left\{0, 1, \frac{1}{2}, \dots, \frac{1}{n}\right\}\right) = (n+1)\alpha$ $\lightning$ if n large as $\mathbb{P} > 1$.
- So $\mathbb{P}(\{0\}) = 0$, or $\mathbb{P}(\{0\})$ is undefined.
- What about $\mathbb{P}\left(\left\{x : x \le \frac{1}{3}\right\}\right)$?
  - ? "Add up" all $\mathbb{P}(\{x\})$ for $x \le \frac{1}{3}$. However this range is uncountable and we can't take a sum of uncountably many terms.

Aside

Question

Can we choose uniformly from an infinite countable set? (E.g. $\Omega = \mathbb{N}$ or $\Omega = \mathbb{Q} \cap [0,1]$)

Answer

No it is not possible but that's ok there $\exists$ lots of interesting probability measures of $\mathbb{N}$!

Proof. Suppose possible

- $\mathbb{P}(\{0\}) = \alpha > 0 \quad \forall\, \omega \in \Omega$. Then $\mathbb{P}(\Omega) = \sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) = \sum_{\omega\in\Omega} \alpha = \infty$. $\lightning$ of P2 : $\mathbb{P}(\Omega) = 1$.
- $\mathbb{P}(\{0\}) = 0 \quad \forall\, \omega \in \Omega$. Then $\mathbb{P}(\Omega) = \sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) = \sum_{\omega\in\Omega} 0 = 0$.

$\square$

Proposition 1.1 (From the axioms)

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

  Proof. A, A^c are disjoint. $A \cup A^c = \Omega$.
  $\implies \mathbb{P}(A) + \mathbb{P}(A^c) \underset{P3}{=} \mathbb{P}(\Omega) \underset{P2}{=} 1$ $\square$

- $\mathbb{P}(\varnothing) = 0$
- If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

§1.1 Examples of Probability Spaces

Example 1.2 (Uniform Choice)

Ω finite, $\Omega = \{\omega_1, \ldots, \omega_n\}$, $\mathcal{F}$ = all subsets. *uniform* choice (equally likely)

$$\mathbb{P} : \mathcal{F} \to [0,1], \ \mathbb{P}(A) = \frac{|A|}{|\Omega|}.$$

In particular: $\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} \quad \forall\, \omega \in \Omega$.

Example 1.3 (Choosing without replacement)

n indistinguishable marbles labelled $\{1, \ldots, n\}$. Pick $k \leq n$ marbles uniformly at random. Here: $\Omega = \{A \subseteq \{1, \ldots, n\}, |A| = k\} \quad |\Omega| = \binom{n}{k}$

Example 1.4 (Well-shuffled deck of cards)

Uniformly chosen *permutation* of 52 cards.

$$\begin{aligned} \Omega &= \{\text{all permutations of 52 cards}\} \\ |\Omega| &= 52! \\ \mathbb{P}\left(\substack{\text{first three cards}\\\text{have the same suit}}\right) &= \frac{52 \times 12 \times 11 \times 49!}{52!} = \frac{22}{425} \\ \text{Note: } &= \frac{12}{51} \times \frac{11}{50} \end{aligned}$$

Example 1.5 (Coincident Birthdays)

There are n people; what is the probability that at least two of them share a birthday?

Assumptions:

- No leap years! (365 days)
- All birthdays are equally likely

Let $\Omega = \{1, \ldots, 365\}^n$ and $\mathcal{F} = \mathcal{P}(\Omega)$.
Let $A = \{$at least two people share the same birthday$\}$ and so
$A^c = \{$all n birthdays are different$\}$.

$$\mathbb{P}(A^c) = \frac{|A^c|}{|\Omega|} = \frac{365 \times 364 \cdots \times (365 - n + 1)}{365^n}$$
$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$$

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

KEY IDEA: Calculating $\mathbb{P}(A^c)$ is easier than $\mathbb{P}(A)$.

§2 Combinatorial Analysis

§2.1 Subsets

Question

Let Ω be finite and $|\Omega| = n$. How many ways to *partition* Ω into k disjoint subsets $\Omega_1, \dots, \Omega_k$ with $|\Omega_i| = n_i$ (with $\sum_{i=1}^{k} n_i = n$)?

Answer

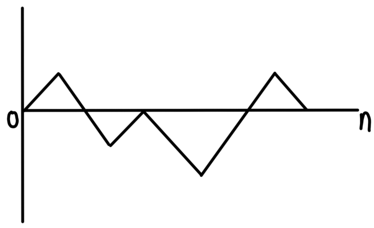
$$
\begin{aligned}
M &= \underset{\substack{\text{Choose} \\ \text{first part}}}{\binom{n}{n_1}} \underset{\substack{\text{Then choose} \\ \text{second part}}}{\binom{n-n_1}{n_2}} \binom{n-n_1-n_2}{n_3} \dots \underbrace{\binom{n-(n_1+\dots+n_{k-1})}{n_k}}_{=1} \\
&= \frac{n!}{n_1!\cancel{(n-n_1)!}} \times \frac{\cancel{(n-n_1)!}}{n_2!\cancel{(n-n_1-n_2)!}} \times \cancel{\frac{[n-(n+n_1+\dots+n_{k-1})]!}{0!n_k!}} \\
&= \frac{n!}{n_1!n_2!\dots n_k!} \\
&= \underbrace{\binom{n}{n_1, n_2, \dots, n_k}}_{\text{Multinomial coefficient}}
\end{aligned}
$$

Key sanity check

- Does ordering of the subsets matter?

E.g. Is $\Omega_2 = \{3,4,7\}, \Omega_3 = \{1,5,8\}$ equal to $\Omega_3 = \{3,4,7\}, \Omega_2 = \{1,5,8\}$? No, ordering does matter as we put elements first in the second subset then the third.

§2.2 Random Walks



$$\Omega = \{(X_0, X_1, \dots, X_n) : X_0 = 0, |X_n - X_{k-1}| = 1 \;\forall\; k = 1, \dots, n\}.$$

$|\Omega| = 2^n$ (we can go either up or down at each k)

$$\mathbb{P}(X_n = n) = \frac{1}{2^n}$$
$$\mathbb{P}(X_n = 0) = 0 \text{ if } n \text{ is odd}$$
$$\text{What about } \mathbb{P}(X_n = 0) \text{ when } n \text{ is even}$$

Idea - Choose $\frac{n}{2}$ ks for $X_k = X_{k-1} + 1$ and the rest $X_k = X_{k-1} - 1$ (i.e. go up half the time and down the other half).

$$\begin{aligned}\mathbb{P}(X_n = 0) &= 2^{-n}\binom{n}{\frac{n}{2}} \\ &= \frac{n!}{2^n\left(\frac{n}{2}!\right)^2}\end{aligned}$$

Question
What happens when n is large?

§2.3 Stirling's Formula

Notation. Let $(a_n), (b_n)$ be two sequences. Say $a_n \sim b_n$ as $n \to \infty$ if $\frac{a_n}{b_n} \to 1$ as $n \to \infty$.

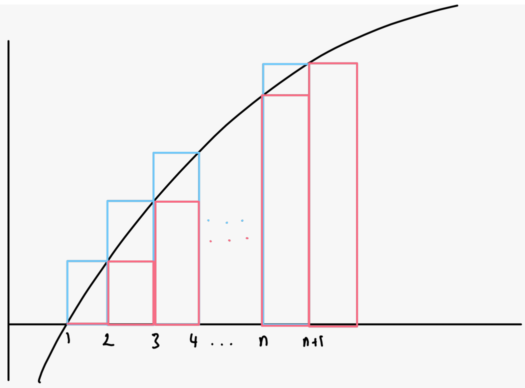
Example 2.1
$n^2 + 5n + \frac{6}{n} \sim n^2$

Example 2.2 (Non-Example)
$\exp\left(n^2 + 5n + \frac{6}{n}\right) \nsim \exp(n^2)$

Theorem 2.1 (Stirling)
$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$ as $n \to \infty$.

Theorem 2.2 (Weaker Version)
$\log n! \sim n \log n$.

Proof. $\log(n!) = \log 2 + \cdots + \log n$.



$$\overset{\text{“Upper Integral”}}{\int_1^n \log x \, dx} \le \log n! \le \overset{\text{“Lower Integral”}}{\int_1^{n+1} \log x \, dx}$$
$$\underbrace{n \log n - n + 1}_{\sim n \log n} \le \log n! \le \underbrace{(n+1)\log(n+1) - n}_{\sim n \log n}$$

Key idea: *Sandwiching* between lower/upper integrals. It was useful that

- $\log x$ is increasing
- $\log x$ has a nice integral!

□

§2.4 (Ordered) compositions

Definition 2.1 (Composition)
A **composition** of m with k parts is a sequence $(m_1, \ldots, m_k)$ of non-negative integers with $m_1 + \cdots + m_k = m$.

Example 2.3

$$3 + 0 + 1 + 2 = 6 \ne 1 + 2 + 0 + 3 = 6$$
$$\star\star\star \,||\, \star \,|\, \star\star$$

There is a bijection between compositions *and* sequences of m stars and $(k-1)$ dividers. So the number of compositions is $\binom{m+k-1}{m}$.

Comment: Easy to mistake k with $k-1$ in no. of dividers.

§3 Properties of Probability measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathbb{P} : \mathcal{F} \to [0,1]$.

Definition 3.1 (Countable additivity)
P3 : $\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$ for $(A_n)_{n\in\mathbb{N}}$ disjoint.

Question
What if the sets are not disjoint?

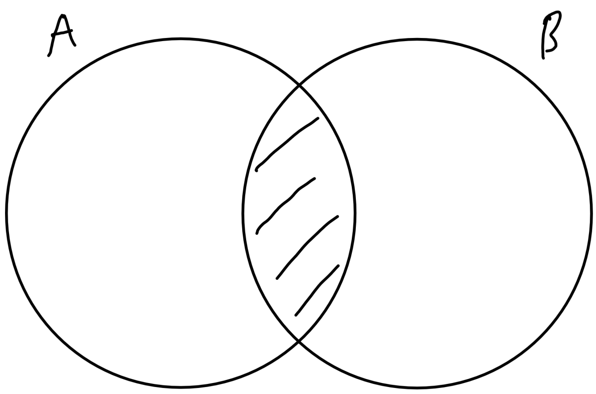
§3.1 Countable sub-additivity

Proposition 3.1 (Countable sub-additivity)
Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events in $\mathcal{F}$. Then

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

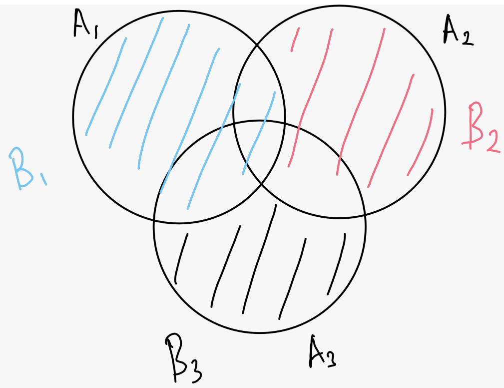
May also be called a *union bound*.

Intuition:



$\sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$ "double counts" some sub-events.

Proof. Idea: Rewrite $\bigcup_{n\in\mathbb{N}} A_n$ as a *disjoint* union. Define $B_1 = A_1$ and $B_n = \underbrace{A_n \setminus (A_1 \cup \cdots \cup A_{n-1})}_{\in\mathcal{F}\text{(by Sheet 1)}} \quad \forall\, n \geq 2$.



So

- $\bigcup_{n\in\mathbb{N}} B_n = \bigcup_{n\in\mathbb{N}} A_n$.
- $(B_n)_{n\in\mathbb{N}}$ is disjoint (by construction).
- $B_n \subseteq A_n \implies \underbrace{\mathbb{P}(B_n) \leq \mathbb{P}(A_n)}_{\text{Q4, Sheet 1}}$

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} B_n\right) \underset{P3 \text{ on } (B_n)}{=} \sum_{n\in\mathbb{N}} \mathbb{P}\left(B_n\right) \leq \sum_{n\in\mathbb{N}} \mathbb{P}\left(A_n\right)$$

□

§3.2 Continuity

Proposition 3.2 (Continuity)
Let $(A_n)_{n\in\mathbb{N}}$ be an increasing sequence of events in $\mathcal{F}$, i.e. $A_n \subseteq A_{n+1} \quad \forall\, n$. Then $\mathbb{P}(A_n) \leq \mathbb{P}(A_{n+1})$. So $\mathbb{P}(A_n)$ converges as $n \to \infty$.[a]

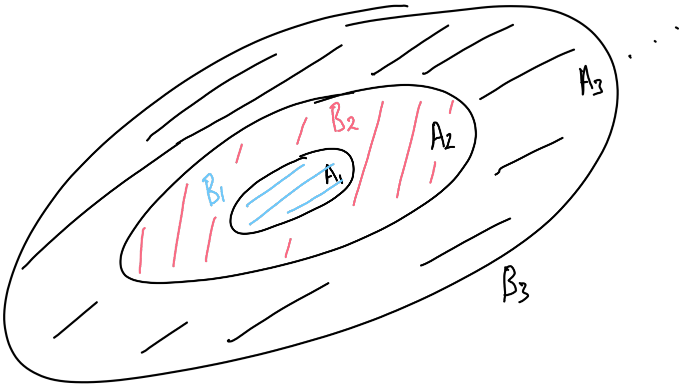
In fact: $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right)$.

[a] As probabilities are bounded above by 1 and increasing.

For motivation try Q6, Sheet 1.

Proof. Let us reuse the B_ns from the previous subsection.

- $\bigcup_{k=1}^{n} B_k = A_n$ (disjoint union).
- $\bigcup_{n\in\mathbb{N}} B_n = \bigcup_{n\in\mathbb{N}} A_n$



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

□

§3.3 Inclusion-Exclusion Principle

Background: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
Similarly: $A, B, C \in \mathcal{F}$

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

Proposition 3.3 (Inclusion-Exclusion Principle)
Let $A_1, \dots, A_n \in \mathcal{F}$, then:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{1\leq i_1<i_2\leq n} \mathbb{P}(A_{i_1} \cap A_{i_2}) + \sum_{1\leq i_1<i_2<i_3\leq n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \dots \\ &\quad + (-1)^{n+1}\mathbb{P}(A_1 \cap \dots \cap A_n) \\ &= \sum_{\substack{I\subset\{1,\dots,n\} \\ I\neq\varnothing}} (-1)^{|I|+1}\mathbb{P}\left(\bigcap_{i\in I} A_i\right)\end{aligned}$$

Note: $\sum_{1\leq i_1<i_2<i_3\leq n}$ is the sum of all triples that are distinct and unordered.

Proof. By induction. For $n = 2$ it holds (Q4e, Sheet 1).

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}\left(\left(\bigcup_{i=1}^{n-1} A_i\right) \cup A_n\right)$$

Using $n = 2$ case we get:

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

We want to break down the final element on the RHS

$$\text{Idea: } \left(\bigcup_{i=1}^{n-1} A_i\right) \cap A_n = \bigcup_{i=1}^{n-1} (A_i \cap A_n)$$

If we apply IEP to $\bigcup_{i=1}^{n-1} (A_i \cap A_n)$ we need to calculate $\bigcap_{i \in J} (A_i \cap A_n)$

$$\bigcap_{i \in J} (A_i \cap A_n) = \bigcap_{i \in J \cup \{n\}} A_i, \qquad J \subset \{1, \ldots, n-1\}$$

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \underbrace{\sum_{\substack{J \subset \{1,\ldots,n-1\} \\ J \neq \varnothing}} (-1)^{|J|+1} \mathbb{P}\left(\bigcap_{i \in J} A_i\right)}_{n-1 \text{ case}} + \mathbb{P}(A_n)$$

$$- \underbrace{\sum_{\substack{J \subset \{1,\ldots,n-1\} \\ J \neq \varnothing}} (-1)^{|J|+1} \mathbb{P}\left(\bigcap_{i \in J \cup \{n\}} A_i\right)}_{n-1 \text{ case on } (A_i \cap A_n)}$$

$$J \cup \{n\} \mapsto I.$$

$$-(-1)^{|J|+1} \mapsto (-1)^{|I|+1}$$

$$= \underbrace{\sum_{\substack{I \subset \{1,\ldots,n-1\} \\ I \neq \varnothing}} (-1)^{|I|+1} \mathbb{P}\left(\bigcap_{i \in I} A_i\right)}_{\text{Just changed the labels}} + \mathbb{P}(A_n)$$

$$+ \sum_{\substack{I \subset \{1,\ldots,n\} \\ n \in I,\ |I| \geq 2}} (-1)^{|I|+1} \mathbb{P}\left(\bigcap_{i \in I} A_i\right)$$

$$= \sum_{\substack{I \subset \{1,\ldots,n\} \\ I \neq \varnothing}} (-1)^{|I|+1} \mathbb{P}\left(\bigcap_{i \in I} A_i\right).$$

Let us check that we have indeed counted all subsets I.

- $\sum_{\substack{I \subset \{1,\dots,n-1\} \\ I \neq \varnothing}} (-1)^{|I|+1} \mathbb{P}\left(\bigcap_{i \in I} A_i\right)$ accounts for all subsets where $n \notin I$.
- $\mathbb{P}(A_n)$ accounts for $\{n\}$
- $\sum_{\substack{I \subset \{1,\dots,n\} \\ n \in I,\ |I| > 2}} (-1)^{|I|+1} \mathbb{P}\left(\bigcap_{i \in I} A_i\right)$ accounts for all subsets where $n \in I$ and $I \neq \{n\}$.

$\square$

§3.4 Bonferroni Inequalities

Question

What if you *truncate* IEP (Inclusion-Exclusion Principle)?

Proposition 3.4 (Bonferroni Inequality)

Recall: Countable sub-additivity - $\mathbb{P}(\cup A_i) \leq \sum \mathbb{P}(A_i)$.

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{k=1}^{r} (-1)^{k+1} \sum_{i_1 < i_2 < \cdots < i_k} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) \quad \text{if } r \text{ is odd}$$

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{k=1}^{r} (-1)^{k+1} \sum_{i_1 < i_2 < \cdots < i_k} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) \quad \text{if } r \text{ is even}$$

Proof. By induction on r and n. Let r be odd

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

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) \leq \underbrace{\sum_{\substack{J \subset \{1,\dots,n-1\} \\ 1 \leq |J| \leq r}} (-1)^{|J|+1} \mathbb{P}\left(\bigcap_{i \in J} A_i\right)}_{n-1 \text{ case.}}$$

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} (A_i \cap A_n)\right) \geq \underbrace{\sum_{\substack{J \subset \{1,\dots,n-1\} \\ 1 \leq |J| \leq r-1^{a}}} (-1)^{|J|+1} \mathbb{P}\left(\bigcap_{i \in J \cup \{n\}} A_i\right)}_{r-1 \text{ case on } (A_i \cap A_n)}$$

Using the same rearranging as in the proof of Inclusion-Exclusion Principle

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{\substack{I\subset\{1,\dots,n\}\\ 1\leq |I|\leq r}} (-1)^{|I|+1}\mathbb{P}\left(\bigcap_{i\in I} A_i\right).$$

The case of r being even is similar, simply note all three inequalities are reversed. ☐

[a] J doesn't include n and we only want r elements in the intersection

Question

When is it good to truncate at e.g. r = 2?

§3.5 Counting with IEP

Uniform probability measure on Ω, $|\Omega| < \infty$. $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ $\forall\ A \subseteq \Omega$. Then $\forall\ A_1, \dots, A_n \subseteq \Omega$

$$|A_1 \cup \dots \cup A_n| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1<\dots<i_k} |A_i \cap \dots \cap A_{i_k}|$$

(and similarly for Bonferroni Inequalities).

Example 3.1 (Surjections)

What is the probability that a function $f : \{1,\dots,n\} \to \{1,\dots,m\}$, $n \geq m$ is a surjection? Let $\Omega = \{f : \{1,\dots,n\} \to \{1,\dots,m\}\}$ and $A = \{f \in \Omega : \text{Image}(f) = \{1,\dots,m\}\}$.
$\forall\ i \in \{1,\dots,m\}$ define $B_i = \{f \in \Omega : i \notin \text{Image}(f)\}$.

Key observations:

- $$\begin{aligned} A &= B_1^c \cap \dots B_m^c \\ &= (B_1 \cup \dots \cup B_m)^c \end{aligned}$$

- $|B_{i_1} \cap \dots \cap B_{i_k}|$ is nice to calculate.

$$\begin{aligned} |B_{i_1} \cap \dots \cap B_{i_k}| &= |\{f \in \Omega : i_1, \dots, i_k \notin \text{Image}(f)| \\ &= (m-k)^n \end{aligned}$$

$$IEP \to |B_1 \cup \dots \cup B_m| = \sum_{k=1}^{m} (-1)^{k+1} \sum_{i_1<\dots<i_k} \underbrace{|B_{i_1} \cap \dots \cap B_{i_k}|}_{\text{same for all } i_1,\dots,i_k}$$

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

Example 3.2 (Derangements)

What is the probability that a permutation has no fixed points? Derangements can be useful in a Secret Santa.

$\Omega = \{\text{permuations of } \{1, \dots, n\}\}$ and the derangements, D, are $\{\sigma \in \Omega : \sigma(i) \neq i \quad \forall\, i = 1, \dots, n\}$.

Question

Is $\mathbb{P}(D) = \frac{|D|}{|\Omega|}$ large or small (e.g. when $n \to \infty$?)

$\forall\, i \in \{1, \dots, n\} : A_i = \{\sigma \in \Omega : \sigma(i) = i\}$.

Key observations:

- $D = A_1^c \cap \dots A_n^c = \left(\bigcup_{i=1}^n A_i\right)^c$.
- $$
\begin{aligned}
\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) &= \frac{(n-k)!}{n!} \\
IEP \to \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < \cdots < i_k} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) \\
&= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \frac{1}{k!} \\
\mathbb{P}(D) &= 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \\
&= 1 - \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} \\
&= \sum_{k=0}^{n} \frac{(-1)^{k}}{k!} \\
\lim_{n \to \infty} \mathbb{P}(D) &= \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k!} \\
&= e^{-1} \approx 0.37.
\end{aligned}
$$

Remark 2.

- What if instead $\Omega' = \{\text{all functions } f : \{1, \dots, n\} \text{ to itself}\}$?

$$\begin{aligned} D &= \{f \in \Omega' : f(i) \neq i \quad \forall\, i = 1, \dots, n\}. \\ \mathbb{P}(D) &= \frac{(n-1)^n}{n^n} \\ &= \left(1 - \frac{1}{n}^n\right) \\ \lim_{n\to\infty} D &= e^{-1}. \end{aligned}$$

- We would liked to have calculated $\mathbb{P}(D)$ by doing $\left(\frac{n-1}{n}\right)^n$ as we have n choices each with probability $\frac{n-1}{n}$. We will be allowed to do this soon! See Example 3.6
- $f(i)$ is a random quantity associated to Ω. We will be allowed to study $f(i)$ as a *random variable* soon.
- We are allowed to toss a fair coin n times, $\Omega = \{H, T\}^n$. But we have not yet studied tossing an unfair coin n times.

§3.6 Independence

$(\Omega, \mathcal{F}, \mathbb{P})$ as before.

Definition 3.2 (Indepence)
Events $A, B \in \mathcal{F}$ are **independent** ($A \perp\!\!\!\perp B$) if

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

A countable[a] collection of events (A_n) are **independent** if $\forall$ distinct $i_1, \dots, i_k$[b] we have:

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}).$$

[a] including finite
[b] k is finite

Remark 3 (Caution). "Pairwise independence" does not imply independence.

Example 3.3

$$\begin{aligned}
\Omega &= \{(H,H),(H,T),(T,H),(T,T)\} \\
\mathbb{P}(\{\omega\}) &= \frac{1}{4} \quad \forall\, \omega \in \Omega. \\
A &= \text{first coin is } H = \{(H,H),(H,T)\}. \\
B &= \text{second coin is } H = \{(T,H),(H,H)\}. \\
C &= \text{both coins have the same outcome} = \{(T,T),(H,H)\}. \\
\mathbb{P}(A) = \mathbb{P}(B) &= P(C) = \frac{1}{2}. \\
A \cap B = A \cap C &= B \cap C = \{(H,H)\}. \\
\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) &= \mathbb{P}(B \cap C) = \frac{1}{4}. \quad \text{Pairwise independence} \checkmark \\
\mathbb{P}(A \cap B \cap C) &= \frac{1}{4} \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C). \quad \text{Independence} \times
\end{aligned}$$

Example 3.4 (Independence)

- $$\begin{aligned}
\Omega' &= \{\text{all functions } f : \{1,\dots,n\} \text{ to itself}\} \\
A_i &= \{f \in \Omega' : f(i) = i\}. \\
\mathbb{P}(A_i) &= \frac{n^{n-1}}{n^n} = \frac{1}{n} \\
\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) &= \frac{n^{n-k}}{n^n} \\
&= \frac{1}{n^k} \\
&= \prod_{j=1}^{k} \mathbb{P}(A_{i_j})
\end{aligned}$$

  Here: (A_i) are independent events.

- $$\begin{aligned}
\Omega &= \{\sigma : \text{ permutation of } \{1,\dots,n\}\} \\
A_i &= \{\sigma \in \Omega : \sigma(i) = i\} \\
\mathbb{P}(A_i) &= \frac{(n-1)!}{n!} = \frac{1}{n}. \\
i \neq j \quad \mathbb{P}(A_i \cap A_j) &= \frac{(n-2)!}{n!} \\
&= \frac{1}{n(n-1)} \\
&\neq \mathbb{P}(A_i)\mathbb{P}(A_j)
\end{aligned}$$

  Here: (A_i) are not independent events.

§3.6.1 Properties

Claim 3.1

If A is independent of B, then A is also independent of B^c.

Proof.

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

□

Claim 3.2

A is independent of $B = \Omega$ and of $C = \varnothing$

Proof. $\mathbb{P}(A \cap \Omega) = P(A) = \mathbb{P}(A)\underbrace{\mathbb{P}(\Omega)}_{=1}$ and so $A \perp\!\!\!\perp \varnothing$ by Claim 1. □

§3.7 Conditional Probability

$(\Omega, \mathcal{F}, \mathbb{P})$ as before.

Consider $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0,\ A \in \mathcal{F}$

Definition 3.3 (Conditional Probability)

The **conditional probability** of A given B is $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

"The probability of A if we know B happened". (e.g. revealing information in succession)

Example 3.5

A, B independent.

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

"Knowing whether B happened doesn't affect the probability of A".

§3.7.1 Properties

- P1 - $\mathbb{P}(A \mid B) \geq 0$
- P2 - $\mathbb{P}(B \mid B) = 1 = \mathbb{P}(\Omega \mid B)$
- P3 - (A_n) disjoint events $\in \mathcal{F}$:

Claim 3.3

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n \mid B\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n \mid B)$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n \mid B\right) &= \frac{\mathbb{P}\left(\left(\bigcup_n A_n\right) \cap B\right)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}\left(\bigcup_n (A_n \cap B)\right)}{\mathbb{P}(B)}, \quad \cup(A_n \cap B) \text{ is a disjoint union} \\
&= \frac{\sum_n \mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} \\
&= \sum_{n\in\mathbb{N}} \mathbb{P}(A_n \mid B)
\end{aligned}$$

Summary: Use definition and apply P1, P2, P3 to the numerator. □

$\mathbb{P}(\bullet \mid B)$ is a function from $\mathcal{F} \to [0,1]$ that satisfies the rules to be a probability measure on Ω.

Aside?

Consider $\Omega' = B$ (especially in finite or countable setting). Let $\mathcal{F}' = \mathcal{P}(B)$. Then $(\Omega', \mathcal{F}', \mathbb{P}(\bullet \mid B))$ also satisfies rules to be a probability measure on Ω'.

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A) \tag{1}$$

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

Example 3.6

Uniform choice of a permutation $(\sigma(1), \sigma(2), \ldots, \sigma(n)) \in \Sigma_n$.

Claim 3.4

$$\mathbb{P}\left(\sigma(k) = i_k \mid \sigma(1) = i, \ldots, \sigma(k-1) = i_{k-1}\right), \quad i_1, \ldots, i_{k-1} \text{ distinct.}$$
$$= \begin{cases} 0 & \text{if } i_k \in \{i_1, \ldots, i_{k-1}\}. \\ \frac{1}{n-k+1}{}^{a} & \text{else} \end{cases}$$

[a]This is an example of (Ordered) compositions

Proof.

$$\begin{aligned} \mathbb{P}\left(\sigma(k) = i_k \mid \sigma(1) = i, \ldots, \sigma(k-1) = i_{k-1}\right) &= \frac{\mathbb{P}(\sigma(1) = i, \ldots, \sigma(k) = i_k)}{\mathbb{P}(\sigma(1) = i, \ldots, \sigma(k-1) = i_{k-1})} \\ &= \frac{\frac{(n-k)!}{n!}}{\frac{(n-k+1)!}{n!}} \\ &= \frac{(n-k)!}{(n-k+1)!} \\ &= \frac{1}{n-k+1} \\ \mathbb{P}(\sigma(1) = i, \ldots, \sigma(k) = i_k) &= 0 \quad \text{if } i_k \in \{i_1, \ldots, i_{k-1}\}. \end{aligned}$$

□

§3.7.2 Law of Total Probability and Bayes' Formula

Definition 3.4 (Partition)

$(B_1, B_2, \ldots)^a \in \Omega$ is a **partition** of Ω if:

- $\Omega = \bigcup_n B_n$
- (B_n) are disjoint

[a]finite or countable

Theorem 3.1 (Law of Total Probability)

(B_n) a finite or countable partition of Ω with $B_n \in \mathcal{F}\ \forall\ n$ s.t. $\mathbb{P}(B_n) > 0$. Then

$\forall\ A \in \mathcal{F}$:

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

Also know as "Partition Theorem".

Proof. Note that $\bigcup_n (A \cap B_n) = A$.

$$\begin{aligned}\mathbb{P}(A) &= \sum_{n\in\mathbb{N}} \mathbb{P}(A \cap B_n) \\ &= \sum_n \mathbb{P}(A \mid B_n)\mathbb{P}(B_n) \quad \text{by Equation (1)}\end{aligned}$$

□

Theorem 3.2 (Bayes' Formula)

Same setup as above

$$\mathbb{P}(B_n \mid A) = \frac{\mathbb{P}(A \mid B_n)\mathbb{P}(B_n)}{\sum_m \mathbb{P}(A \mid B_m)\mathbb{P}(B_m)}.$$

Let $n = 2$: $\mathbb{P}(B \mid A)\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(A \cap B)$

Proof.

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

□

Example 3.7 (Lecture course)

Consider a Lecture course which has $2/3$ of the lectures on weekdays and $1/3$ on weekends. Let

$$\begin{aligned}\mathbb{P}(\text{forget notes} \mid \text{weekday}) &= \frac{1}{8} \\ \mathbb{P}(\text{forget notes} \mid \text{weekend}) &= \frac{1}{2}\end{aligned}$$

What is $\mathbb{P}(\text{weekend} \mid \text{forget notes})$? Let $B_1 = \{\text{weekday}\}$, $B_2 = \{\text{weekend}\}$ and

$A = \{\text{forget notes}\}$.
By LTP (Law of Total Probability): $\mathbb{P}(A) = \frac{2}{3} \times \frac{1}{8} + \frac{1}{3} \times \frac{1}{2} = \frac{1}{4}$.
By Bayes' Formula: $\mathbb{P}(B_2 \mid A) = \frac{1}{3} \times \frac{1}{2} / \frac{1}{4} = \frac{2}{3}$.

Example 3.8 (Disease Testing)
Suppose p are infected and $(1-p)$ are not. $\mathbb{P}(\text{tests positive} \mid \text{infected}) = 1 - \alpha$ and $\mathbb{P}(\text{tests positive} \mid \text{not infected}) = \beta$ where $\alpha, \beta \in (0, 1)$.

We want to work out $\mathbb{P}(\text{infected} \mid \text{test positive})$.

By LTP: $\mathbb{P}(\text{test positive}) = p(1-\alpha) + (1-p)\beta$.
By Bayes' Formula: $\mathbb{P}(\text{infected} \mid \text{test positive}) = \frac{p(1-\alpha)}{p(1-\alpha)+(1-p)\beta}$.

Suppose $p \ll \beta$ then $p(1-\alpha) \ll (1-p)\beta$ so $\mathbb{P}(\text{infected} \mid \text{test positive}) \sim \frac{p(1-\alpha)}{(1-p)\beta} \sim \frac{p}{\beta}$ which is small.

Example 3.9 (Simpson's Paradox)
Scientists ask: do jelly beans make you tongue change colour?

Oxford	Change	No change	% change	
Blue	15	22	41%	$\Delta = 3\%$
Green	5	8	38%	

Cambridge	Change	No change	% change	
Blue	10	3	77%	$\Delta = 15\%$
Green	23	14	62%	

Total	Change	No change	% change	
Blue	25	25	50%	$\Delta = -6\%$
Green	28	22	56%	

The conclusion from this example should be that the Cambridge methodology is different to the Oxford one rather than anything about blue/ green jelly beans.[a]

Let $A = \{\text{change colour}\}$, $B = \{\text{blue}\}$, $B^C = \{\text{green}\}$, $C = \{\text{Cambridge}\}$, $C^c = \{\text{Oxford}\}$.

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

$$\not\Longrightarrow \mathbb{P}(A \mid B) > \mathbb{P}(A \mid B^c)$$

[a]Obviously this is a frivolous example however if we changed Oxford to November 2021, Cambridge to January 2021 and we were measuring vaccine efficacy for different vaccines we would get similar results. And it would be reasonable to conclude that the main underlying factor was a change in the viral landscape rather than waning efficacy.

Theorem 3.3 (Law of Total Probability for Conditional Probabilities)
Suppose $C_1, C_2, \ldots$ a partition of B.

$$\mathbb{P}(A \mid B) = \sum_n \mathbb{P}(A \mid C_n)\mathbb{P}(C_n \mid B)$$

Proof.

$$\begin{aligned}
\mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}\left(A \cap \left(\bigcup_n C_n\right)\right)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}\left(\bigcup_n (A \cap C_n)\right)}{\mathbb{P}(B)} \\
&= \frac{\sum_n \mathbb{P}(A \cap C_n)}{\mathbb{P}(B)} \\
&= \frac{\sum_n \mathbb{P}(A \mid C_n)\mathbb{P}(C_n)}{\mathbb{P}(B)} \\
&= \sum_n \mathbb{P}(A \mid C_n)\frac{\mathbb{P}(B \cap C_n)}{\mathbb{P}(B)} \qquad C_n \subset B \implies B \cap C_n = C_n \\
&= \sum_n \mathbb{P}(A \mid C_n)\mathbb{P}(C_n \mid B)
\end{aligned}$$

$\square$

Non Examinable

Special Case:

- If all $\mathbb{P}(C_n)$ are equal then so are $\mathbb{P}(C_n \mid B)$. Note $\sum_n \mathbb{P}(C_n \mid B) = 1$.[1]
- If $\mathbb{P}(A \mid C_n)$ are all equal.

[1]$\mathbb{P}(C_n \mid B) = \frac{\mathbb{P}(C_n \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(C_n)}{\mathbb{P}(B)}$.

Then $\mathbb{P}(A \mid B) = \mathbb{P}(A \mid C_n)$.

Example 3.10 (Well-shuffled deck of cards)

Uniformly chosen *permutation*, $\sigma \in \Sigma_{52}$, of 52 cards. $\{1, 2, 3, 4\}$ are *aces*. Let $A = \{\sigma(1), \sigma(2) \text{ are aces}\}$, $B = \{\sigma(1) \text{ is an ace}\} = \{\sigma(1) \le 4\}$, $C_1 = \{\sigma(1) = 1\}, \dots, C_4 = \{\sigma(1) = 4\}$.

Note:

- $$\begin{aligned} \mathbb{P}(A \mid C_i) &= \mathbb{P}(\sigma(2) \in \{1, 2, 3, 4\} \mid \sigma(1) = i) \quad i \le 4 \\ &= \frac{3}{51} \text{ by Example 3.6} \end{aligned}$$
- $\mathbb{P}(C_1) = \cdots = \mathbb{P}(C_4) = \frac{1}{52}$.

So $\mathbb{P}(A \mid B) = \frac{3}{51}$ and $\mathbb{P}(A) = \mathbb{P}(B)\mathbb{P}(A \mid B) = \frac{4}{52} \times \frac{3}{51}$

§4 Discrete Random Variables

Motivation: Roll two dice. $\Omega = \{1, \dots, 6\}^2 = \{(i,j) : 1 \le i, j \le 6\}$. If we restrict our attention to:

- the first dice e.g. $\{(i,j) : i = 3\}$.
- the sum of the dice e.g. $\{(i,j) : i + j = 8\}$.
- the max of the dice e.g. $\{(i,j) : i, j \le 4, i$ or $j = 4\}$.

This is annoying and we want to move on from sets.

Goal: "Random real-valued measurements", we want the value of the first dice to be X and the sum to be $X + Y$...

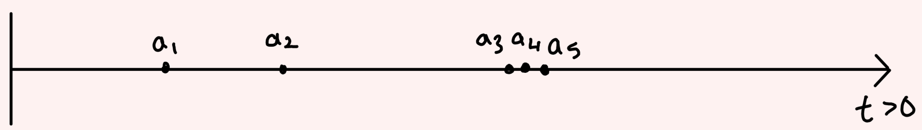
Definition 4.1 (Discrete Random Variable)

A **discrete random variable** X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$ s.t.

- $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$
- $\mathrm{Image}(X)$ if finite or countable (subset of $\mathbb{R}$).
- We abbreviate $\{\omega \in \Omega : X(\omega) = x\}$ as $\{X = x\}$. So $\mathbb{P}(X = x)$ is valid.
- Often $\mathrm{Image}(X) = \mathbb{Z}$ or $\mathbb{N}_0$ or $\{0, 1\}$ etc. *not* {Heads or Tails}.

If Ω is finite or countable and $\mathcal{F} = \mathcal{P}(\Omega)$ both blue bullet points hold automatically.

Example 4.1 (Part II Applied Probability)



"random arrival process". Let $\Omega = \{$countable subsets $(a_1, a_2, \dots)$ of $(0, \infty)\}$ and $N_t =$ number of arrivals by time $t = |\{a_i : a_i \le t\}| \in \mathbb{N}_0$ is a discrete RV (random variable) for each time t.

Definition 4.2 (Probability Mass Function)

The **probability mass function** of discrete RV X is the function $p_X : \mathbb{R} \to [0, 1]$ given by $p_X(x) = \mathbb{P}(X = x) \quad \forall\, x \in \mathbb{R}$.

Note.

- if $x \notin \text{Image}(X)$ then $p_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}) = \mathbb{P}(\varnothing) = 0$.
- $$\begin{aligned}\sum_{x \in \text{Im}(X)} p_X(x) &= \sum_{x \in \text{Im}(X)} \mathbb{P}(\underbrace{\{\omega \in \Omega : X(\omega) = x\}}_{\text{disjoint}}) \\ &= \mathbb{P}\left(\bigcup_{x \in \text{Im}(X)} \{\omega \in \Omega : X(\omega) = x\}\right) \\ &= \mathbb{P}(\Omega) \\ &= 1\end{aligned}$$

Example 4.2 (Indicator function)
For event $A \in \mathcal{F}$, define $1_A : \Omega \to \mathbb{R}$ by

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

1_A is a discrete RV with Image $= \{0, 1\}$. $p_{1_A}(1) = \mathbb{P}(1_A = 1) = \mathbb{P}(A)$, $p_{1_A}(0) = \mathbb{P}(1_A = 0) = \mathbb{P}(A^c)$ and $p_{1_A}(x) = 0 \quad \forall\, x \notin \{0, 1\}$.

This encodes "did A happen" as a real number.

Remark 4. Given a pmf p_X (probability mass function), we can always construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a RV defined on it with this pmf.

- $\Omega = \text{Im}(X)$ i.e. $\{x \in \mathbb{R} : p_X(x) > 0\}$
- $\mathcal{F} = \mathcal{P}(\Omega)$
- $\mathbb{P}(\{x\}) = p_X(x)$ and extend to all $A \in \mathcal{F}$

§4.1 Discrete Probability Distributions

§4.1.1 Finite Ω

Definition 4.3 (Bernoulli Distribution - "(biased) coin toss")
If $X \sim \text{Bern}(p)$ where $p \in [0, 1]$ then $\text{Im}(X) = \{0, 1\}$, $p_X(1) = \mathbb{P}(X = 1) = p$ and $p_X(0) = 1 - p$.

Example 4.3

$1_A \sim \mathrm{Bern}(p)$ with $p = \mathbb{P}(A)$.

Definition 4.4 (Binomial Distribution)

If $X \sim \mathrm{Bin}(n,p)$ where $n \in \mathbb{Z}^+$ and $p \in [0,1]$ then $\mathrm{Im}(X) = \{0,1,\dots,n\}$, $p_X(k) = \mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$. $\sum_{k=0}^n p_X(k) = (p+(1-p))^n = 1$.

The binomial distribution can be used to model the number of heads when a coin is tossed n times.

§4.1.2 More than one RV

Motivation: Roll a dice with outcome $X \in \{1,2,\dots,6\}$. Events: $A = \{1 \text{ or } 2\}$, $B = \{1 \text{ or } 2 \text{ or } 3\}$, $C = \{1 \text{ or } 3 \text{ or } 5\}$. $1_A \sim \mathrm{Bern}\left(\frac{1}{3}\right)$, $1_B \sim \mathrm{Bern}\left(\frac{1}{2}\right)$, $1_C \sim \mathrm{Bern}\left(\frac{1}{2}\right)$.
Note: $1_A \le 1_B$ for all outcomes
but $1_A \le 1_C$ for all outcomes *is false*.

Definition 4.5 (Independent RVs)

Let $X_1,\dots,X_n$ be discrete RVs. We say $X_1,\dots,X_n$ are *independent* if:

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1)\dots\mathbb{P}(X_n = x_n) \quad \forall\, x_1,\dots,x_n \in \mathbb{R}.$$

(suffices to check $\forall\, x_i \in \mathrm{Im}(X_i)$)

Example 4.4

$X_1,\dots,X_n$ independent RVs each with the Bernoulli(p) distribution. Study $S_n = X_1 + \dots + X_n$. Then

$$\begin{aligned}
\mathbb{P}(S_n = k) &= \sum_{\substack{x_1+\dots+x_n=k \\ x_i \in \{0,1\}}} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \\
&= \sum_{x_1+\dots+x_n=k} \mathbb{P}(X_1 = x_1)\dots\mathbb{P}(X_n = x_n) \\
&= \sum_{x_1+\dots+x_n=k} p^{|\{i : x_i = 1\}|}(1-p)^{|\{i : x_i = 0\}|} \\
&= \sum_{x_1+\dots+x_n=k} p^k (1-p)^{n-k}
\end{aligned}$$

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

So $S_n \sim \mathrm{Bin}(n,p)$.

Example 4.5 (Non-example)
$(\sigma(1), \sigma(2), \ldots, \sigma(n))$ a uniform permutation.

Claim 4.1
$\sigma(1)$ and $\sigma(2)$ are *not* independent.

Suffices to find i_1, i_2 s.t. $\mathbb{P}(\sigma(1) = i_1, \sigma(2) = i_2) \neq \mathbb{P}(\sigma(1) = i_1)\mathbb{P}(\sigma(2) = i_2)$. E.g. $\mathbb{P}(\sigma(1) = 1, \sigma(2) = 1) = 0 \neq \underbrace{\mathbb{P}(\sigma(1) = 1)\mathbb{P}(\sigma(2) = 1)}_{=1/n \times 1/n}$

Consequence of definition

Let $X_1, \ldots, X_n$ be independent. Then $\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X \in A_1) \ldots \mathbb{P}(X_n \in A_n) \quad \forall\, A_1, \ldots, A_n \subset \mathbb{R}$ countable.

§4.1.3 $\Omega = \mathbb{N}$ - "Ways of choosing a random integer"

Definition 4.6 (Geometric Distribution ("Waiting for success"))
If $X \sim \mathrm{Geo}(p)$ where $p \in (0,1)$. $\mathrm{Im}(X) = \{1, 2, \ldots\}$, $p_X(k) = \mathbb{P}(\text{(k-1) failures, then success on the kth trial}) = (1-p)^{k-1}p$. Check: $\sum_{k\geq 1}(1-p)^{k-1}p = p\sum_{t\geq 0}(1-p)^t = \frac{p}{1-(1-p)} = 1$.

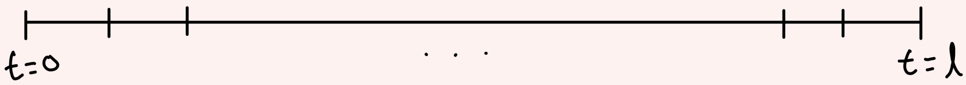
Alternatively: "Count how many failures before a success"
$\mathrm{Im}(Y) = \{0, 1, 2, \ldots\}$, $p_Y(k) = \mathbb{P}(\text{k failures, then success on the (k+1)th trial})$. Check: $\sum_{k\geq 0}(1-p)^k p = 1$.

The geometric distribution can be used to model the number of coin tosses until we get a head.

Definition 4.7 (Poisson Distribution)
If $X \sim \mathrm{Po}(\lambda)$ (or $\mathrm{Poi}(\lambda)$) with $\lambda \in (0, \infty)$. $\mathrm{Im}(X) = \{0, 1, 2, \ldots\}$ and $\mathbb{P}(X = k) = e^{-\lambda}\lambda^k / k! \quad \forall\, k \geq 0$. Check: $\sum_{k\geq 0}\mathbb{P}(X = k) = e^{\lambda}\sum_{k\geq 0}\frac{\lambda^k}{k!} = e^{\lambda}e^{\lambda}$.

Motivation: Consider $X_n \sim \mathrm{Bin}(n, \frac{\lambda}{n})$

Example 4.6 ("Arrival proccess")



Split time interval $[0, \lambda]$ into n small intervals.

- Probability of an arrival in each interval is p, independently across intervals.
- Total no. of arrivals is X_n.

$$\mathbb{P}(X_n = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Fix k and let $n \to \infty$

$$= \frac{n!}{n^k (n-k)!} \times \underbrace{\frac{\lambda^k}{k!}}_{\text{no n}} \times \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \times \underbrace{\left(1 - \frac{1}{n}\right)^{-k}}_{\to 1}$$

$$\begin{aligned} \frac{n!}{n^k (n-k)!} &= \frac{n(n-1)\dots(n-k+1)}{n^k} \\ &= 1 \times \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \dots \times \left(1 - \frac{k-1}{n}\right) \\ &\to 1 \quad \text{There are a fixed number of terms all converging to 1} \end{aligned}$$

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

We might want to say $\mathrm{Bin}(n, \frac{\lambda}{n})$ converges to $\mathrm{Po}(\lambda)$, but what does convergence of random variables mean?

§4.2 Expectation

$(\Omega, \mathcal{F}, \mathbb{P})$ and X a discrete RV. For now: X only takes non-negative values. "$X \geq 0$"

Definition 4.8 (Expectation)

The **expectation of X** (or **expected value** or **mean**)

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

"Average of values taken by X, weighted by p_X".

Example 4.7 (Uniform Dice)

X uniform on $\{1, 2, \ldots, 6\}$

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 2 + \cdots + \frac{1}{6}\cdot 6 \\ &= 3.5\end{aligned}$$

Note. $\mathbb{E}[X]$ need not be in $\mathrm{Im}(X)$.

Example 4.8 (Binomial Distribution)

Let $X \sim \mathrm{Binomial}(n, p)$

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

$$\textit{Trick:}\quad \begin{aligned} k\binom{n}{k} &= \frac{k \times n!}{k! \times (n-k)!} \\ &= \frac{n!}{(k-1)!(n-k)!} \\ &= n\binom{n-1}{k-1}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X] &= n\sum_{k=1}^{n}\binom{n-1}{k-1}p^k(1-p)^k \\ &= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^k \\ &= np\sum_{l=0}^{n-1}\underbrace{\binom{n-1}{l}p^l(1-p)^{(n-1)-l}}_{\text{pmf of }\mathrm{Bin}(n-1,p)} \\ &= np(p + (1-p))^{n-1} \\ &= np.\end{aligned}$$

Note. We would like to say:

$$\mathbb{E}[\mathrm{Bin}(n, p)] = \mathbb{E}[\mathrm{Bern}(p)] + \cdots + \mathbb{E}[\mathrm{Bern}(p)]$$

We will be able to do this soon.

Example 4.9 (Poisson Distribution)
Let $X \sim \text{Poisson}(\lambda)$

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k \ge 0} k\mathbb{P}(X = k) \\
&= \sum_{k \ge 0} k \cdot e^{-\lambda} \frac{\lambda^k}{k!} \\
&= \sum_{k \ge 1} e^{-\lambda} \frac{\lambda^k}{(k-1)!} \\
&= \lambda \sum_{k \ge 1} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} \qquad \text{pmf of Poisson}(\lambda) \\
&= \lambda.
\end{aligned}$$

Note. We would like to say

$$\mathbb{E}[\text{Poisson}(\lambda)] \approx \mathbb{E}\left[\text{Bin}\left(n, \frac{\lambda}{n}\right)\right] = \lambda$$

It is not true in general that $\mathbb{P}(X_n = k) \approx \mathbb{P}(X = k) \implies \mathbb{E}[X_n] \approx \mathbb{E}[X]$

Not important

If X can take on any real value (not necessarily $X \ge 0$)

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

unless: A, $\sum_{\substack{x>0 \\ x \in \text{Im}(X)}} x\mathbb{P}(X = x) = +\infty$ and B, $\sum_{\substack{x<0 \\ x \in \text{Im}(X)}} x\mathbb{P}(X = x) = -\infty$.
Then we say $\mathbb{E}[X]$ is not defined. Do we really want to study $\infty + \frac{2}{3}(-\infty)$

Summary:

- A and B, $\mathbb{E}[X]$ is not defined.
- A but not B, $\mathbb{E}[X] = +\infty$.[2]

[2]Some people say not defined instead of letting $\mathbb{E}[X] = \pm\infty$

- B but not A, $\mathbb{E}[X] = -\infty$.
- neither A nor B, X is then integrable i.e. $\mathbb{E}[X]$ absolutely converges.

Example 4.10

Most examples in the course are integrable *except*:

- $\mathbb{P}(X = n) = \frac{6}{\pi^2}\frac{1}{n^2}$ for $n \geq 1$. Note that $\sum \mathbb{P}(X = n) = 1$.
  Then $\mathbb{E}[X] = \sum \frac{6}{\pi^2}\frac{1}{n} = +\infty$.
- $\mathbb{P}(X = n) = \frac{3}{\pi^2}\frac{1}{n^2}$ for $n \in \mathbb{Z} \setminus \{0\}$. Then $\mathbb{E}[X]$ is not defined. "It's symmetric so $\mathbb{E}[X] = 0$", we have decided that this is wrong to prevent many things going wrong in second and third year courses in probability.

Example 4.11 (Indicator Function)

$\mathbb{E}[1_A] = \mathbb{P}(A)$.

§4.2.1 Properties of Expectation

Proposition 4.1

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

Proof. $\mathbb{E}[X] = \sum_{\substack{x \in \mathrm{Im}(X) \\ x \neq 0}} x\mathbb{P}(X = x)$ □

Proposition 4.2 (Linearity of expectation)

Given random variables X, Y (both integrable) on same probability space $\forall\ \lambda, \mu \in \mathbb{R}$

$$\mathbb{E}[\lambda X + \mu Y] = \lambda\mathbb{E}[X] + \mu\mathbb{E}[Y]$$

$$\text{Similarly} \quad \mathbb{E}[\lambda_1 X_1, + \cdots + \lambda_n X_n] = \lambda_1\mathbb{E}[X_1] + \cdots + \lambda_n\mathbb{E}[X_n]^{a}$$

[a] holds for countably infinite collection though proof is omitted until more analysis experience.

Note. *Independence* is *NOT* a condition.

Proof. If Ω is countable:

$$
\begin{aligned}
\mathbb{E}[\lambda X + \mu Y] &= \cancel{\sum_{z \in \mathrm{Im}(\lambda X + \mu Y)} z\mathbb{P}(\lambda X + \mu Y = z)} \quad \text{awkward} \\
&= \sum_{\omega \in \Omega} (\lambda X(\omega) + \mu Y(\omega))\mathbb{P}(\{\omega\}) \\
&= \lambda \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}) + \mu \sum_{\omega \in \Omega} Y(\omega)\mathbb{P}(\{\omega\}) \\
&= \lambda\mathbb{E}[X] + \mu\mathbb{E}[Y].
\end{aligned}
$$

□

Aside - Special Cases

1. If $\lambda, c \in \mathbb{R}$ then:
   a) $\mathbb{E}[X + c] = \mathbb{E}[X] + c$
   b) $\mathbb{E}[\lambda X] = \lambda\mathbb{E}[X]$
2. a) X, Y random variables (both integrable) on same probability space. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
   b) in fact $\lambda, \mu \in \mathbb{R}$ $\mathbb{E}[\lambda X + \mu Y] = \lambda\mathbb{E}[X] + \mu\mathbb{E}[Y]$ (similarly $\mathbb{E}[\lambda_1, X_1, + \cdots + \lambda_n X_n] = \lambda_1\mathbb{E}[X_1] + \cdots + \lambda_n\mathbb{E}[X_n]$)

Corollary 4.1

$X \geq Y$[a] then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

[a] $X(\omega) \geq Y(\omega) \quad \forall\, \omega \in \Omega$

Proof.

$$
\begin{aligned}
X &= (X - Y) + Y \\
\mathbb{E}[X] &= \mathbb{E}[X - Y] + \mathbb{E}[Y] \\
X - Y \geq 0 &\implies \mathbb{E}[X - Y] \geq 0 \quad \text{by Proposition 4.1.}
\end{aligned}
$$

□

Example 4.12 (Counting Problems)

$(\sigma(1), \dots, \sigma(n))$ uniform on Σ_n. $Z = |\{i : \sigma(i) = i\}|$ = number of fixed points. Let $A_i = \{\sigma(i) = i\}$, recall from Example 3.4 A_is *not* independent.

Key step:

$$\begin{aligned} Z &= 1_{A_1} + \dots + 1_{A_n} \\ \mathbb{E}[Z] &= \mathbb{E}[1_{A_1} + \dots + 1_{A_n}] \\ &= \mathbb{E}[1_{A_1}] + \dots + \mathbb{E}[1_{A_n}] \quad \text{by Linearity of expectation} \\ &= \mathbb{P}(A_1) + \dots + \mathbb{P}(A_n) \quad \text{by Example 4.11} \\ &= \frac{1}{n} n = 1. \end{aligned}$$

Note. Same answer as $\mathrm{Bin}(n, \frac{1}{n})$

Proposition 4.3

If X takes values in $\{0, 1, 2, \dots\}$ then

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

Proof. One can carefully re-arrange the summands which is left as an exercise to the reader. □

Alternative. Write $X = \sum_{k \geq 1} 1_{X \geq k}$[a] then take $\mathbb{E}[X]$:

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}\left[\sum 1_{X \geq k}\right] \\ &= \sum \mathbb{E}[1_{X \geq k}] \\ &= \sum \mathbb{P}(X \geq k) \end{aligned}$$

□

[a] Sanity check: let $X = 7$, $1_{X \geq 1} = \dots = 1_{X \geq 7} = 1$ whilst $1_{X \geq 8} = 1_{X \geq 9} = \dots = 0$.

Claim 4.2 (Markov's Inequality)

Let $X \geq 0$ be a random variable. Then $\forall\, a > 0$:

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

The LHS is interesting, e.g. if we want to bound the probability of an extreme outcome, whilst the RHS is easy to study.

Note. Is $a = \frac{\mathbb{E}[X]}{2}$ useful? No, we already know probabilities are less than 2. If a is large it might be useful

Proof. Observe $X \geq a 1_{X \geq a}$[a] and take $\mathbb{E}$

$$\begin{aligned} \mathbb{E}[X] &\geq a\mathbb{E}[1_{X \geq a}] \\ &= a\mathbb{P}(X \geq a) \end{aligned}$$

□

[a] Check: If $X \in [0, a)$ then RHS $= 0$ else RHS $= a$.

Note. Markov's Inequality is also true for continuous RVs.

§4.2.1 Studying $\mathbb{E}[f(X)]$

Let $f : \mathbb{R} \to \mathbb{R}$ be a function. Then $f(X)$ is also a *random variable*[3].

Claim 4.3

$$\mathbb{E}[f(X)] = \sum_{x \in \mathrm{Im}(X)} f(x)\mathbb{P}(X = x)^{a}.$$

[a] if it exists

Proof. Let $A = \mathrm{Im}(f(x)) = \{f(x) : x \in \mathrm{Im}(X)\}$. Starting with RHS

$$\begin{aligned} \sum_{x \in \mathrm{Im}(X)} f(x)\mathbb{P}(X = x) &= \sum_{y \in A} \sum_{\substack{x \in \mathrm{Im}(X) \\ f(x) = y}} f(x)\mathbb{P}(X = x) \\ &= \sum_{y \in A} y \sum_{\substack{x \in \mathrm{Im}(X) \\ f(x) = y}} \mathbb{P}(X = x) \\ &= \sum_{y \in A} y\mathbb{P}(f(X) = y) \quad \text{by additivity} \\ &= \mathbb{E}[f(X)] \end{aligned}$$

□

[3] $X : \Omega \to \mathbb{R}$ so $f(X) : \Omega \to \mathbb{R}$.

§4.3 Variance

Motivation

$$\begin{aligned}U_n &\sim \text{Uniform}(\{-n, -n+1, \dots, n\}) \\ V_n &\sim \text{Uniform}(\{-n, n\}) \\ Z_n &= 0 \\ S_n &= \text{random walk for } n \text{ steps} \\ &\sim n - 2\,\text{Bin}\left(n, \frac{1}{2}\right)\end{aligned}$$

All of these have $\mathbb{E} = 0$.

Variance is a way to "measure how concentrated a RV is around its mean".

Definition 4.9

The **variance** of X is:

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

Proposition 4.4

$\text{Var}(X) \geq 0$ with equality $\iff$ $\mathbb{P}(X = \mathbb{E}[X]) = 1$ (as $(X - \mathbb{E}[X])^2$ so by Proposition 4.1).

Definition 4.10 (Alternative characterisation)

$$\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \quad (\geq 0)$$

Proof. Write $\mu = \mathbb{E}[X]$

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}\left[(X-\mu)^2\right] \\ &= \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] \\ &= \mathbb{E}[X^2] - 2\mu \underbrace{\mathbb{E}[X]}_{\mu} + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2\end{aligned}$$

□

Proposition 4.5 (Properties)
If $\lambda, c \in \mathbb{R}$

- $\operatorname{Var}(\lambda X) = \lambda^2 \operatorname{Var}(X)$.
- $\operatorname{Var}(X + c) = \operatorname{Var}(X)$

Proof.

$$\begin{aligned}
\mathbb{E}[X + c] &= \mu + c \\
\operatorname{Var}(X + c) &= \mathbb{E}\left[\left(X + c - (\mu + c)^2\right)\right] \\
&= \mathbb{E}\left[(X - \mu)^2\right] \\
&= \operatorname{Var}(X).
\end{aligned}$$

□

Example 4.13 (Poisson Distribution)
Let $X \sim \operatorname{Poisson}(\lambda)$

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

"Falling factorial trick": sometimes easier to calculate $\mathbb{E}[X(X-1)]$ than $\mathbb{E}[X^2]$

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k \geq 2} \underbrace{k(k-1)}_{\text{function}} \underbrace{e^{-\lambda} \frac{\lambda^k}{k!}}_{\text{PMF}} \\
&= \lambda^2 e^{-\lambda} \underbrace{\sum_{k \geq 2} \frac{\lambda^{k-1}}{(k-2)!}}_{e^{\lambda}} \\
&= \lambda^2 \\
\mathbb{E}[X^2] &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] \\
&= \lambda^2 + \lambda \\
\operatorname{Var}(X) &= \lambda
\end{aligned}$$

Example 4.14 (Geometric Distribution)

Let $Y \sim \text{Geom}(p)$ where $Y \in \mathbb{N}$.

$$\mathbb{E}[Y] = \frac{1}{p}, \ \text{Var}(Y) = \frac{1-p}{p^2}.$$

Proof left as an exercise.

Note. λ large: $\text{Var}(x) = \mathbb{E}[X]$, more concentrated
p small: $\text{Var}(Y) \approx \frac{1}{p^2} = (\mathbb{E}[X])^2$.

Example 4.15 (Bernouli Distribution)
Let $X \sim \text{Bern}(p)$.

$$\begin{aligned}
\mathbb{E}[X] &= 1 \times p = p \\
\mathbb{E}[X^2] &= 1^2 \times p = p \\
\text{Var}(X) &= p - p^2 \\
&= p(1-p)
\end{aligned}$$

Example 4.16 (Binomial Distribution)
Let $X \sim \text{Bin}(n, p)$

$$\begin{aligned}
\mathbb{E}[X] &= np \\
\mathbb{E}\left[X^2\right] &= \text{ugly}
\end{aligned}$$

§4.3.1 Sums of RVs

Goal: Study $\text{Var}(X_1 + \cdots + X_n)$. Do the X_is need to be independent.

Proposition 4.6 (Preliminary: Expectation of Product of RVs)
If X, Y are independent RVs and f, g are functions $\mathbb{R} \to \mathbb{R}$.
Then: $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$ "Splits as a product" .

Example 4.17
$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Example 4.18

Let $f(x) = g(x) = z^x$ (or e^{tx}).

Proof. (X, Y discrete)

$$\begin{aligned}
\text{LHS} &= \sum_{x,y\in\text{Im}} f(x)g(y)\mathbb{P}(X=x, Y=y) \\
&= \sum_{x,y\in\text{Im}} f(x)g(y)\mathbb{P}(X=x)\mathbb{P}(Y=y) \\
&= \left[\sum_{x\in\text{Im}} f(x)\mathbb{P}(X=x)\right]\left[\sum_{y\in\text{Im}} g(x)\mathbb{P}(Y=y)\right] \\
&= \mathbb{E}[f(X)]\mathbb{E}[g(Y)]
\end{aligned}$$

□

Proposition 4.7 (Sums of independent RVs)

Let $X_1, \ldots, X_n$ be independent. Then

$$\text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n)$$

Proof. (Suffices to Prove (STP) $n = 2$). Say $\mathbb{E}[X] = \mu$, $\mathbb{E}[Y] = \nu$

$$\begin{aligned}
\text{Var}(X+Y) &= \mathbb{E}\left[(X+Y-\mu-\nu)^2\right] \\
&= \mathbb{E}\left[(X-\mu)^2\right] + \mathbb{E}\left[(Y-\nu)^2\right] + 2\mathbb{E}[(X-\mu)(Y-\nu)]. \\
&= \text{Var}(X) + \text{Var}(Y) + 2\mathbb{E}[X-\mu]\mathbb{E}[Y-\nu] \\
&= \text{Var}(X) + \text{Var}(Y).
\end{aligned}$$

□

Example 4.19 (Binomial)

Going back to Binomial Distribution, $\text{Var}(\text{Bin}(n,p)) = np(1-p)$.

Goal: Study $\text{Var}(X+Y)$ when X, Y not independent.

Definition 4.11 (Covariance)

Let X, Y be two RVs. Their **covariance** is

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

"Measures how dependent X, Y are and in which direction" (X large $\implies$ Y larger if $\mathrm{Cov} > 0$ else Y smaller).

Proposition 4.8 (Properties)

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{Cov}(Y,X)\\ \mathrm{Cov}(X,X) &= \mathrm{Var}(X)\end{aligned}$$

Definition 4.12 (Alternative Characterisation)

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

Proof.

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}\left[(X-\mu)(Y-\nu)\right]\\ &= \mathbb{E}[XY] - \mu\underbrace{\mathbb{E}[Y]}_{\nu} - \nu\underbrace{\mathbb{E}[X]}_{\mu} + \mu\nu\\ &= \mathbb{E}[XY] - \mu\nu.\end{aligned}$$

□

Proposition 4.9 (More Properties)

Let $\lambda \in \mathbb{R}$

$$\begin{aligned}\mathrm{Cov}(\lambda, X) &= 0\\ \mathrm{Cov}(X+\lambda, Y) &= \mathrm{Cov}(X,Y)\\ \mathrm{Cov}(\lambda X, Y) &= \lambda\,\mathrm{Cov}(X,Y)\\ \mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)\end{aligned}$$

Covariance is linear in each argument i.e. $\mathrm{Cov}(\sum \lambda_i X_i, Y) = \sum \lambda_i \mathrm{Cov}(X_i, Y)$ and $\mathrm{Cov}(\sum \lambda_i X_i, \sum \mu_j Y_j) = \sum_{i=1}^n \sum_{j=1}^m \lambda_i \mu_j \mathrm{Cov}(X_i, Y_j)$.

Proposition 4.10 (Sums of RVs)

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) &= \mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\right)\\ &= \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j)^{a}\end{aligned}$$

[a] or $2\sum_{i<j} \mathrm{Cov}(X_i, X_j)$

Proposition 4.11

X, Y independent $\implies \mathrm{Cov}(X, Y) = 0$, the converse is false.

Example 4.20

Let $Y = -X$, $\mathrm{Var}(Y) = \mathrm{Var}(X)$. So $\mathrm{Var}(X+Y) = \mathrm{Var}(0) = 0 \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Example 4.21 (Uniform Permutation)

Let $(\sigma(1), \dots, \sigma(n))$ be uniformly chosen on Σ_n. Let $A_i = \{\sigma(i) = i\}$ and $N = 1_{A_1} + \cdots + 1_{A_n}$ (N is the number of fixed points).

$$\mathbb{E}[N] = n \times \frac{1}{n} = 1$$

A_i and A_j are *not* independent

$$\begin{aligned}\mathrm{Var}(1_{A_1}) &= \frac{1}{n}\left(1 - \frac{1}{n}\right)\\ \mathrm{Cov}(1_{A_i}, 1_{A_j}) &= \mathbb{E}[1_{A_i}1_{A_j}] - \mathbb{E}[1_{A_i}]\mathbb{E}[1_{A_j}]\\ &= \mathbb{E}[1_{A_i\cap A_j}] - \mathbb{E}[1_{A_i}]\mathbb{E}[1_{A_j}]\\ &= \mathbb{P}(A_i\cap A_j) - \mathbb{P}(A_i)\mathbb{P}(A_j)\\ &= \frac{1}{n(n-1)} - \frac{1}{n}\times\frac{1}{n}\\ &= \frac{1}{n^2(n-1)} > 0\end{aligned}$$

Note: Cov doesn't depend on i, j

$$\mathrm{Var}(N) = \sum_{i=1}^n \mathrm{Var}(1_{A_i}) + \sum_{i\neq j}\mathrm{Cov}(1_{A_i}, 1_{A_j})$$

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

Compare with Bin $\left(n, \frac{1}{n}\right)$: $\mathbb{E} = 1$, $\text{Var} = n \times \frac{1}{n}\left(1 - \frac{1}{n}\right) = 1 - \frac{1}{n}$.

§4.3.2 Chebyshev's Inequality

Proposition 4.12 (Chebyshev's Inequality)
Let X be a RV, $\mathbb{E}[X] = \mu$ finite, $\text{Var}(X) = \sigma^2 < \infty$.

$$\mathbb{P}(|X-\mu| \geq \lambda) \leq \frac{\text{Var}(X)}{\lambda^2}$$

Remember the proof, not the statement.

Proof. Idea: Apply Markov's Inequality to $(X-\mu)^2$.

$$
\begin{aligned}
\mathbb{P}\left((X-\mu)^2 \geq \lambda^2\right) &\leq \frac{\mathbb{E}\left[(X-\mu)^2\right]}{\lambda^2} \\
&= \frac{\text{Var}(X)}{\lambda^2}
\end{aligned}
$$

□

Danger: Applying Markov's Inequality to $|X - \mu|$, $\mathbb{E}\left[|X-\mu|\right]$ is less nice than $\mathbb{E}\left[(X-\mu)^2\right]$.

Comments

- Chebyshev's Inequality gives better bounds than Markov's Inequality (decays with λ^2 instead of λ).
- We can apply it to all RVs, not just those ≥ 0.
- Caveat: We need $\text{Var}(X) < \infty$ which is a stronger condition than $\mathbb{E}[X] < \infty$.

Definition 4.13 (Standard Deviation)
$\sqrt{\text{Var}(X)}$ is the **standard deviation**, σ, of X.

It has the same "units" as X but not many nice properties so Var is generally preferred.

We can rewrite Chebyshev as $\mathbb{P}(|X-\mu| \geq k\sigma) \leq 1/k^2$

§4.4 Conditional Expectation

Setting: $(\Omega, \mathcal{F}, \mathbb{P})$.
Recall the definition of Conditional Probability.

Definition 4.14 (Conditional Expectation)
$B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, X a RV. The **conditional expectation** is

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

Example 4.22 (Uniform Dice)
Let X be a dice, uniform on $\{1, \dots, 6\}$.

$$\begin{aligned}\mathbb{E}[X \mid X \text{ prime}] &= \frac{\frac{1}{6}[0+2+3+0+5+0]}{\frac{1}{2}} \\ &= \frac{1}{3}(2+3+5) \\ &= \frac{10}{3}.\end{aligned}$$

Definition 4.15 (Alternative Characterisation)

$$\mathbb{E}[X \mid B] = \sum_{x \in \operatorname{Im} X} x \mathbb{P}(X = x \mid B).$$

Proof.

$$\begin{aligned}\text{RHS } &= \sum \frac{x\mathbb{P}(\{X = x\} \cap B)}{\mathbb{P}(B)} \\ &= \sum_{\substack{x \neq 0 \\ x \in \operatorname{Im} X}} \frac{x \mathbb{P}(X 1_B = x)}{\mathbb{P}(B)} \\ \text{Note: } \mathbb{E}[X 1_B] &= \sum_{\substack{x \neq 0 \\ x \in \operatorname{Im} X}} x \mathbb{P}(X 1_B = x)\end{aligned}$$

□

Proposition 4.13 (Law of Total Expectation)

Let $(B_1, B_2, \dots)$ be a finite or countably-infinite partition of Ω with $B_n \in \mathcal{F} \quad \forall\, n$ s.t. $\mathbb{P}(B_n) > 0$. X a RV.

$$\mathbb{E}[X] = \sum_n \mathbb{E}[X \mid B_n]\mathbb{P}(B_n).$$

Example 4.23

Let $X = 1_A$ we recover the Law of Total Probability.

Proof.

$$\begin{aligned} \text{RHS} &= \sum_n \mathbb{E}[X 1_{B_n}] \\ &= \mathbb{E}\left[X \cdot (1_{B_1} + \cdots + 1_{B_n})\right] \quad \text{by Linearity of expectation} \\ &= \mathbb{E}[X \cdot 1] \\ &= \mathbb{E}[X]. \end{aligned}$$

□

Application: Two stage randomness where (B_n) describes what happens in stage 1.

Example 4.24 (Sums of random number of terms)

Let $(X_n)_{n\geq 1}$ be IID and $N \in \{0, 1, 2, \dots\}$ be a random index independent of (X_n). $S_n = X_1 + \cdots + X_n$ with $\mathbb{E}[X_n] = \mu$ so $\mathbb{E}[S_n] = n\mu$. Then

$$\begin{aligned} \mathbb{E}[S_N] &= \sum_{n\geq 0} \mathbb{E}[S_N \mid N = n]\mathbb{P}(N = n) \\ &= \sum \mathbb{E}[S_n]\mathbb{P}(N = n)^a \\ &= \sum_{n\geq 0} n\mu\mathbb{P}(N = n) \\ &= \mu\mathbb{E}[N]. \end{aligned}$$

[a] KEY STEP $\mathbb{E}[S_N \mid N = n] = \mathbb{E}[S_n \mid N = n] = \mathbb{E}[S_n]$ last step follows as S_n and $\{N = n\}$ are independent.

§4.5 Random Walks

Definition 4.16 (Random Walk)

Let $(X_n)_{n\geq 1}$ be IID RVs then $S_n = x_0 + X_1 + \cdots + X_n$. $(S_0, S_1, S_2, \dots)$ is a random process called a **Random Walk** started from x_0.

§4.5.1 Simple Random Walk (SRW) on $\mathbb{Z}$ - Main example in our course

Definition 4.17 (Simple Random Walk)

A **simple random walk** has $P(X_i = +1) = p$ and $\mathbb{P}(X = -1) = q = 1 - p$. $x_0 \in \mathbb{Z}$ and is often 0.
It is "symmetric" in the special case where $p = q = \frac{1}{2}$.

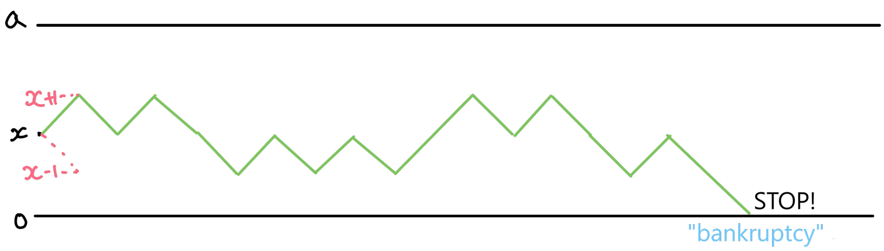
Example 4.25

$\mathbb{P}(S_2 = x_0) = pq + qp = 2pq$

Useful interpretation: A gamble repeatedly plays a game where he wins £1 with $\mathbb{P} = p$, loses £1 with $\mathbb{P} = q$.
Often: stops at £0.

Question

Suppose the gambler starts with £x at time 0. What is the probability he reaches £a before £0. $(0 < x < a)$



Notation. $\mathbb{P}_x(\cdot) = \mathbb{P}(\cdot \mid x_0 = x)$ "measure of RW started from x_0."

Answer

Key idea: Conditional on $S_1 = z$, $(S_1, S_2, \dots)$ is a random walk started from z.
Apply Law of Total Probability

$$\begin{aligned}
\mathbb{P}_x(S \text{ hits a before } 0) &= \sum^{a} \overbrace{\mathbb{P}_x(S \text{ hits a before } 0 \mid S_1 = z)}^{\text{Interpret this.}} \mathbb{P}_x(S_1 = z) \\
&= \sum_z \mathbb{P}_z(S \text{ hits a before } 0)\mathbb{P}_x(S_1 = z) \\
\text{let } h_x &= \mathbb{P}_x(S \text{ hits a before } 0). \\
S_1 &= x \pm 1 \\
h_x &= p h_{x+1} + q h_{x-1}.
\end{aligned}$$

Important to specify boundary conditions: $h_0 = 0$, $h_a = 1$

Solving Linear Recurrence Equations
$ph_{x+1} - h_x + qh_{x-1} = 0$ is a homogenous equation whose solutions form a vector space. We want to find two LI solutions and we guess $h_x = \lambda^x$ So

$$\begin{aligned} p\lambda^{x+1} - \lambda^x + q\lambda^{x-1} &= 0 \\ p\lambda^2 - \lambda + q &= 0 \\ \lambda &= 1, \frac{q}{p} \end{aligned}$$

Case $q \neq p$

$$h_x = A + B\left(\frac{q}{p}\right)^x$$

Use BCs to find A,B:

$$\begin{aligned} x = 0 : h_0 &= 0 = A + B \\ x = a : h_a &= 1 = A + B\left(\frac{q}{p}\right)^a \\ h_x &= \frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^a - 1}. \end{aligned}$$

Case $p = q = \frac{1}{2}$:
Note $h_x = x$, "x is the average of $x+1$ and $x-1$".
General solution: $h_x = A + Bx$

$$\begin{aligned} h_0 &= 0 = A \\ h_a &= 1 = Ba \\ h_x &= \frac{x}{a} \end{aligned}$$

[a] Subscript: $z \in \mathrm{Im}(S_1)$ but we will not bother with that any more

Probability sanity check: $p = q = \frac{1}{2}$. "Fair game"
Study: Expected profit if you start from £x and play until time T.

$$\begin{aligned} \mathbb{E}_x[S_T] &= a\mathbb{P}_x(S_T = a) + 0 \times \mathbb{P}_x(S_T = 0) \\ &= ah_x = x \end{aligned}$$

Fits our intuition for fair games. ✓

Question

Suppose the gambler starts with £x at time 0. What is the expected absorption time, $T = \min\{n \geq 0 : S_n = 0 \text{ or } S_n = a\}$. "first time S hits $\{0, a\}$"

Answer

Apply Law of Total Expectation

We want $\mathbb{E}_x[T]$ ($\mathbb{E}[T]$ when we start from x) which we label as τ_x

$$\begin{aligned}\tau_x &= \mathbb{E}_x[T] \\ &= p\overbrace{\mathbb{E}_x[T \mid S_1 = x+1]}^{\text{Interpret this.}}{}^{a} + q\mathbb{E}_x[T \mid S_1 = x-1] \\ &= p\mathbb{E}_{x+1}[T+1] + q\mathbb{E}_{x-1}[T-1] \\ &= p(1 + \mathbb{E}_{x+1}[T]) + q(1 + \mathbb{E}_{x-1}[T]) \\ &= 1 + p\tau_{x+1} + q\tau_{x-1}\end{aligned}$$

Boundary conditions: $\tau_0 = \tau_a = 0$ "We're already there" .

We already solved the homogenous case of this equation previously. We want to find a *particular solution*, guess: "one level more complicated than general solution"

$\underline{p \neq q}$: Guess: $\tau_x = \frac{x}{q-p}$ works as a particular solution

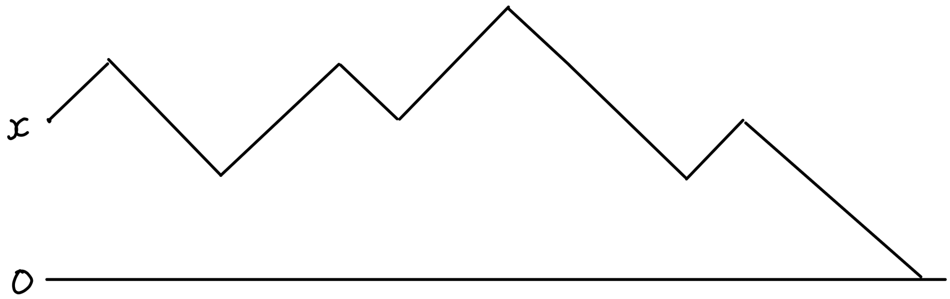
$\underline{p = q = \frac{1}{2}}$: Guess $\tau_x = Cx^2$ *might* work

$$\begin{aligned}\text{Sub in: } \frac{C}{2}(x+1)^2 - Cx^2 + \frac{C}{2}(x-1)^2 &= -1 \\ C &= -1 \\ \tau_x &= A + Bx - x^2 \\ \tau_0 = \tau_a = 0 \text{ and } \tau_x &\geq 0 \\ \therefore \tau_x &= x(a-x)\end{aligned}$$

[a] "As if we started from $(x+1)$ and incremented time by one unit."

§4.5.2 Unbounded RW: "Gambler's Ruin'

$$\begin{aligned}\mathbb{P}_x(\text{hit } 0) &= \lim_{a\to\infty} \mathbb{P}_x(\text{hit } 0 \text{ before } a) \\ &= \lim_{a\to\infty} 1 - h_x\end{aligned}$$



$$= \begin{cases} (\frac{q}{p})^x & p > q \\ 1 & p \leq q \end{cases}$$

$$\text{When } p = \frac{1}{2}: \quad \mathbb{E}_x[\text{time to hit } 0] \geq \mathbb{E}_x[\text{time to hit } 0 \text{ or } a]$$
$$= x(a-x) \to \infty \text{ as } a \to \infty$$

Key conclusion: T_x (time to hit 0 from x) is for $p = \frac{1}{2}$ finite with probability 1 and has infinite expectation.

Comment (non - examinable)

Alternative derivation of $\mathbb{E}[T_1] = \infty$.
$\mathbb{E}[T_2] = 2\mathbb{E}[T_1]$ as going from $2 \to 1$ is the same as going from $1 \to 0$.

$$\begin{aligned} \mathbb{E}[T_1] &= \frac{1}{2} \times 1 + \frac{1}{2}(1 + \mathbb{E}[T_2]) \\ &= 1 + \mathbb{E}[T_1] \end{aligned}$$

We conclude that $\mathbb{E}[T_1] = \infty$.
insertpicture

---

§4.6 Generating Functions

Definition 4.18 (Probability Generating Function)
Let X be a RV taking values in $\{0, 1, 2, \dots\}$. The **probability generating function** of X is

$$G_X(z) = \mathbb{E}\left[z^X\right] = \sum_{k \geq 0} z^k \mathbb{P}(X = k).$$

Analytic comment: $G_X : (-1, 1) \to \mathbb{R}$.

Idea: "A pgf *encodes* the distribution of X as a function with nice analytic properties."

Example 4.26 (Bernoulli)
Let $X \sim \mathrm{Bern}(p)$

$$\begin{aligned} G_X(z) &= z^0\mathbb{P}(X=0) + z\mathbb{P}(X=1) \\ &= (1-p) + pz. \end{aligned}$$

Example 4.27 (Poisson)
Let $X \sim \mathrm{Poi}(\lambda)$

$$\begin{aligned} G_X(z) &= \sum_{k\geq 0} z^k e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k\geq 0}\frac{(\lambda z)^k}{k!} \\ &= e^{-\lambda}e^{\lambda z} \\ &= e^{\lambda(z-1)} \end{aligned}$$

Note. $G_X(0) = 0^0\mathbb{P}(X=0) = \mathbb{P}(X=0)$.

Proposition 4.14 (Recovering PMF from PGF)
$\mathbb{P}(X=n) = \frac{1}{n!}G_X^{(n)}(0)$

Proof. Idea: Differentiate n times

$$\begin{aligned} \frac{d^n}{dz^n}G_X(z) &= \sum_{k\geq 0}\frac{d^n}{dz^n}(z^k)\mathbb{P}(X=k) \\ &= \sum_{k\geq 0}\frac{k!}{(k-n)!}z^{k-n}\mathbb{P}(X=k) \\ &= \sum_{k\geq n}\frac{k!}{(k-n)!}z^{k-n}\mathbb{P}(X=k) \\ &= \sum_{l\geq 0}\frac{(l+n)!}{l!}z^{l}\mathbb{P}(X=l+n) \\ \frac{d^n}{dz^n}G_X(0) &= n!\mathbb{P}(X=n). \end{aligned}$$

□

Key fact: PGF *determines* PMF/distribution exactly.

Note. $G_X(1)$[4] $= \sum_{k\geq 0} \mathbb{P}(X=k) = 1$.

Proposition 4.15 (Recoving other probablistic qunatities)

$$G_X^{(n)}(1) = \mathbb{E}[\underbrace{X(X-1)\dots(X-n+1)}_{\text{Falling Factorial.}}]$$

Proof.

$$\begin{aligned} G_X^{(n)}(1) &= \sum_{k\geq n} \underbrace{k(k-1)\dots(k-n+1)}_{\text{function of } k} \mathbb{P}(X=k) \\ &= \mathbb{E}[X(X-1)\dots(X-n+1)] \end{aligned}$$

□

Proposition 4.16 (Variance in terms of pgf)

$$\mathrm{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$$

Proof.

$$\begin{aligned} \mathbb{E}\left[X^2\right] &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] \\ &= G''_X(1) + G'_X(1) \end{aligned}$$

□

Idea: Find general $\mathbb{E}[P(X)]$ ($P(X)$ is a polynomial) using $\mathbb{E}[\text{falling factorials of } X]$.

Linear Algebra aside

The falling factorials $1, X, X(X-1), \dots$ form a *basis* for $\mathbb{R}[X]$ (vector space of polynomials).

Proposition 4.17 (PGF for sum of independent RVs)

[4]Technical Comment: $G_X(1)$ means $\lim_{z\uparrow 1} G_X(z)$ if the domain is $(-1,1)$. In particular $G'_x(1)$ is possible

Let $X_1, \dots, X_n$ be independent RVs with pgfs $G_{X_1}, \dots, G_{X_n}$. Let $X = X_1 + \dots + X_n$.

$$G_X(z) = G_{X_1}(z) \dots G_{X_n}(z)$$

Special case: X_i are IID, $G_X(z) = (G_{X_1}(z))^n$. Much nicer than PMF of X!

Proof.

$$\begin{aligned}
G_X(z) &= \mathbb{E}\left[z^X\right] \\
&= \mathbb{E}\left[z^{X_1 + \dots X_N}\right] \\
&= \mathbb{E}\left[\underbrace{z^{X_1}}_{\substack{\text{function}\\ \text{of } X_1}} z^{X_2} \dots \underbrace{z^{X_n}}_{\substack{\text{function}\\ \text{of } X_n}}\right] \\
&= \mathbb{E}\left[z^{X_1}\right] \dots \mathbb{E}\left[z^{X_n}\right] \qquad \text{as } X_i \text{ are independent so by Proposition 4.6} \\
&= G_{X_1}(z) \dots G_{X_n}(z).
\end{aligned}$$

□

Example 4.28 (Binomial)
Let $X \sim \text{Bin}(n, p)$

$$\begin{aligned}
X &= X_1 + \dots + X_n, \quad X_i \text{ IID Bern}(p) \\
G_X(z) &= (1 - p + pz)^n.
\end{aligned}$$

Example 4.29 (Q6, Sheet 3)
Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$.

$$\begin{aligned}
G_X(z) &= e^{\lambda(z-1)}, \; G_Y(z) = e^{\mu(z-1)} \\
\text{Study } Z &= X + Y \\
G_Z(z) &= G_X(z) G_Y(z) \\
&= e^{\lambda(z-1)} e^{\mu(z-1)} \\
&= \underbrace{e^{(\lambda+\mu)(z-1)}}_{\text{PGF of Poi}(\lambda+\mu)}
\end{aligned}$$

So $X + Y \sim \text{Poi}(\lambda + \mu)$.

Proposition 4.18 (PGF for random sums)
Let $X_1, X_2, \ldots$ be IID with the same distribution as X. X takes values in $\{0, 1, 2, \ldots, \}$ and let N be a RV taking values in $\{0, 1, 2, \ldots\}$ *independent* of (X_n).

$$G_{X_1+\cdots+X_N} = G_N(G_X(z))$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[z^{X_1+\ldots X_N}\right] &= \sum_{n\geq 0} \mathbb{E}\left[z^{X_1+\cdots+X_N} \mid N=n\right] \mathbb{P}(N=n) \quad \text{by Law of Total Expectation} \\
&= \sum_{n\geq 0} \mathbb{E}\left[z^{X_1+\cdots+X_n} \mid N=n\right] \mathbb{P}(N=n) \quad \text{Replace with conditioning} \\
&= \sum_{n\geq 0} \mathbb{E}\left[z^{X_1+\cdots+X_n}\right] \mathbb{P}(N=n) \quad (N, X_i) \text{ independent so get rid of conditioning} \\
&= \sum_{n\geq 0} \mathbb{E}\left[z^{X_1}\right] \ldots \mathbb{E}\left[z^{X_n}\right] \mathbb{P}(N=n) \quad (X_i)\text{s independent} \\
&= \sum_{n\geq 0} (G_X(z))^n \, \mathbb{P}(N=n) \\
&= G_N(G_X(z))
\end{aligned}$$

□

Example 4.30 (Bernoulli - Q7, Sheet 3)
Let $X_i \sim \mathrm{Bern}(p)$ and $N \sim \mathrm{Poi}(\lambda)$.

$$G_{X_i}(z) = (1-p) + pz$$
$$G_N(s) = e^{\lambda(s-1)}$$

Interpretation: "Poisson thinning"
"$\mathrm{Poi}(\lambda)$ misprints, each gets found with $\mathbb{P} = 1-p$"

$$\begin{aligned}
Y &= X_1 + \cdots + X_N \\
G_Y(z) &= G_N(G_{X_i}(z)) \\
&= e^{\lambda(1-p+pz-1)} \\
&= \underbrace{e^{\lambda p(z-1)}}_{\text{PGF of Poi}(\lambda p)}
\end{aligned}$$

In general the PMF $X_1 + \cdots + X_N$ is horrible whilst $G_N(G_X(z))$ is nice.

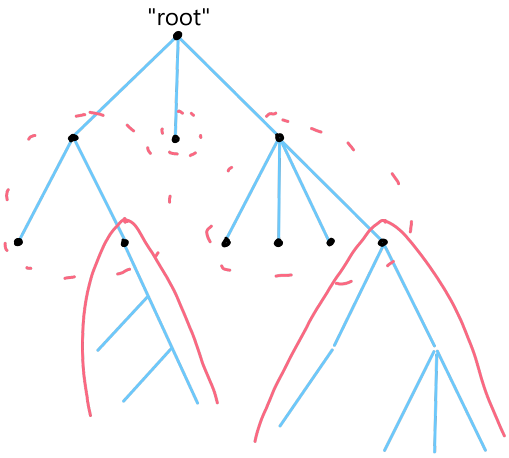
§4.7 Branching Process

"Modelling growth of a population"

Definition 4.19 (Random branching tree)
Let X be a RV on $\{0,1,2,\dots\}$. There is one individual at generation 0 and each individual has a random number of children with distribution X.

Goal:

- Study number of individuals in each generation
- Total population size - is it finite or infinite?



Reduction: Let Z_n = be the number of individuals in generation n. $Z_0 = 1$, $Z_1 \sim X$, $Z_{n+1} = X_1^{(n)} + \cdots + X_{Z_n}^{(n)}$ where $X_k^{(n)}$ are IID with distribution X and independent of Z_n. "$X_k^{(n)}$ = number of children of kth individual in generation n."

Note. If $Z_n = 0 \implies Z_{n+1} = Z_{n+2} = \cdots = 0$.

Theorem 4.1
$\mathbb{E}[Z_n] = (\mathbb{E}[X])^n$

Proof. Z_{n+1} is a random sum so $\mathbb{E}[Z_{n+1}] = \mathbb{E}[X]\mathbb{E}[Z_n]$. By induction the result follows. □

Notation. $\mu = \mathbb{E}[X] \implies \mathbb{E}[Z_n] = \mu^n$.

Notation. Let G be the PGF of X and G_n the PGF of Z_n.

Theorem 4.2
$G_n(z) = G(\dots G(z)\dots)$[a].

[a] Sometimes written as $G^n(z)$, but this is confusing notation so we won't use it

Proof. $G_{n+1}(z) = G_n(G(z))$ by PGF for random sums and so result follows by induction. $\square$

Question
What is the probability the population has goes extinct?

Definition 4.20 (Extinction Probability by generation n)
The probability that the population is extinct by generation n is $q_n = \mathbb{P}(Z_n = 0)$.

Definition 4.21 (Extinction Probability)
The probability that the population goes extinct is $q = \mathbb{P}(Z_n = 0 \text{ for any } n \geq 1)$, i.e. the population size is finite.

Note. $\{Z_n = 0\} \subseteq \{Z_{n+1} = 0\}$ as $Z_n = 0 \implies Z_{n+1} = 0$. Also $\{Z_n = 0 \text{ for any } n \geq 1\} = \bigcup_{n\geq 1}\{Z_n = 0\}$.

Theorem 4.3
$\mathbb{P}(Z_n = 0) \uparrow \mathbb{P}(\bigcup_{n\geq 1}\{Z_n = 0\})$[a], i.e. $q_n \uparrow q$ as $n \to \infty$.

[a] $\uparrow$ is convergence with an increasing sequence.

Proof. By Continuity. $\square$

There are 3 cases to consider

- $\mu < 1$ - subcritical
- $\mu = 1$ - critical
- $\mu > 1$ - supercritical

The degenerate case $\mathbb{P}(X = 1) = 1$ is boring so we exclude it.

Theorem 4.4

$q = 1 \iff \mu = \mathbb{E}[X] \leq 1.$[a]

[a] This does not hold in the degenerate case obviously.

Remark 5. Interesting that q depends on X only through $\mathbb{E}[X]$.

Interpretation: Consider a pandemic spreading through a population, obviously it cannot infect infinite people. Instead we take "finite" to mean e.g. 100 people are infected out of a large population and "infinite" might mean the model stops making sense/ a significant positive proportion are infected.

Baby Proof - (Subcritical).

$$\begin{aligned} \mathbb{P}(Z_n \geq 1) &\leq \frac{\mathbb{E}[Z_n]}{1} \quad \text{by Markov's Inequality} \\ &= \mu^n \to 0. \end{aligned}$$

For supercritical case, note $\mathbb{E}[Z_n] \to \infty$ does not imply $\mathbb{P}(Z_n = 0) \not\to 1$. □

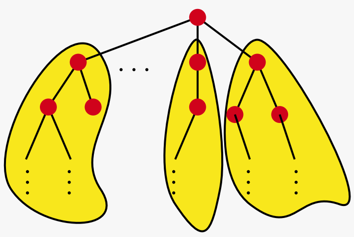
Recall G is the PGF of X and G_n the PGF of Z_n, $q_n = \mathbb{P}(Z_n = 0) = G_n(0)$ and that q is the extinction probability.

Claim 4.4

$G(q) = q$.

Proof 1. $q_{n+1} = G(q_n)$ by Theorem 4.2, $q_n \to q$ and G is continuous as it's a power series so $G(q_n) \to G(q)$ so $q = G(q)$ as $n \to \infty$. □

Proof 2 - LTP (revision of random sums). Conditional on $Z_1 = k$, we get k independent branching processes.



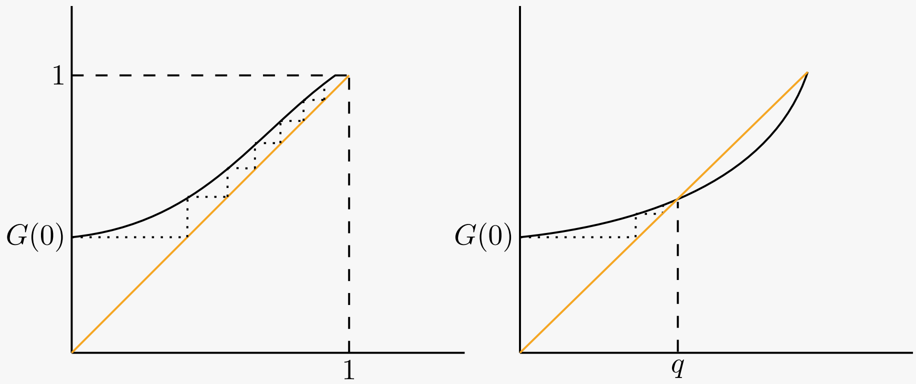
The total population is finite $\iff$ all subtrees[a] of 1^{st} generation are finite.

$$q = \mathbb{P}(\text{finite})$$

$$
\begin{aligned}
&= \sum_{k\geq 0} \mathbb{P}(\text{all finite} \mid Z_1 = k)\mathbb{P}(Z_1 = k) \\
&= \sum_{k\geq 0} [\mathbb{P}(\text{finite})]^k \mathbb{P}(Z_1 = k) \\
&= \sum q^k \mathbb{P}(Z_1 = k) \\
&= G(q).
\end{aligned}
$$

Facts about G

- $G(0) = \mathbb{P}(X = 0) \geq 0$
- $G(1) = 1$
- $G'(1) = \mathbb{E}[X] = \mu$.
- G is smooth, all derivatives ≥ 0 on $[0, 1)$ as all coefficients of the power series are non-negative.



The 1[st] graph has gradient < 1 so only one solution at $q = 1$.
In the 2[nd] graph, the gradient is > 1 so there is only one solution on $[0, 1)$ by IVT on $G(z) - z$. □

[a] Each subtree has the same distribution as the original tree.

Theorem 4.5

q, the extinction probability, is the minimal solution to $z = G(z)$ in $[0, 1]$.[a]

[a] Assuming $\mathbb{P}(X = 1) \neq 1$.

Corollary 4.2

$q = 1 \iff \mu \leq 1$

Proof. Let t be the minimal solution to $t = G(t)$. Reminder: G is increasing.

$$\begin{aligned} t &\geq 0 \\ \implies G(t) &\geq G(0) \\ \implies G(G(t)) &\geq G(G(0)) \\ \implies G(\dots G(t) \dots) &\geq G(\dots G(0) \dots) \\ t &\geq q_n \\ t &\geq q \ \text{ as } n \to \infty. \end{aligned}$$

q is a solution by Claim 4.4 and is bounded above by the minimal solution so $t = q$. ☐

§5 Continuous Probability

We will focus on the case where $\text{Im}(X)$ is an interval in $\mathbb{R}$.
Why?

- Natural for measuring physical quantities, proportions ...
- "Limits" of discrete RV.
- Calculus tools for nice calculations.

Definition 5.1 (Random Variable - Redefintion)
A **random variable** X on $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$ s.t. $\{X \leq x\} \in \mathcal{F}$.

This is consistent with the previous definition when Ω is countable (or $\text{Im}(X)$ is countable).

Drawback: We cannot take $\mathcal{F} = \mathcal{P}(\mathbb{R})$.

Definition 5.2 (Cumulative Distribution Function)
The **cdf** of RV X is

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

Example 5.1 (A 6-sided dice)
insertpicture

§5.1 Properties of CDF

Claim 5.1
F_X increasing i.e. $x \leq y \implies F_X(x) \leq F_y(y)$.

Proof. $F_X(x) = \mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y) = F_X(y)$. □

Claim 5.2
$\mathbb{P}(X > x) = 1 - F_X(x)$

Claim 5.3

$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$

Proof (Not Lectured).

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

□

Claim 5.4

F_X is right-continuous and the left limit exists, i.e. $\lim_{y \downarrow x} F_X(y) = F_X(x)$ and $\lim_{y \uparrow x} F_X(y) = F_X(x-) = \mathbb{P}(X < x)$.

Proof - Right Continuous. STP $F_X\left(x + \frac{1}{n}\right) \to F_X(x)$ as $n \to \infty$.
Let $A_n = \{x < X \leq x + \frac{1}{n}\}$. Then (A_n) are decreasing events and $\bigcap_n A_n = \varnothing$.
So $\mathbb{P}(A_n) = F_X\left(x + \frac{1}{n}\right) - F_X(x)$ and $\mathbb{P}(A_n) \to \mathbb{P}(\varnothing) = 0$. □

Proof - Left Limits. $F_X\left(x - \frac{1}{n}\right)$ is an increasing sequence bounded above by $F_X(x)$. Consider $B_n = \left\{X \leq x - \frac{1}{n}\right\}$ then (B_n) increasing and $\bigcup_n B_n = \{X < x\}$. So $F_X\left(x - \frac{1}{n}\right) = \mathbb{P}(B_n) \to \mathbb{P}(X < x)$. □

Claim 5.5

$\lim_{x \to \infty} F_X = 1$, $\lim_{x \to -\infty} F_X = 0$.

Proof. Let $A_n = X \leq n$, so (A_n) are increasing events and $\cup_n A_n = \Omega$. So $F_X(n) = \mathbb{P}(X \leq n) \to \mathbb{P}(\Omega) = 1$ by Continuity.

Similar for $\lim_{x \to -\infty} F_X = 0$. □

§5.2 Continuous RVs

Definition 5.3 (Continuous RV)

A RV is **continuous** if F is continuous.

Claim 5.6

If X is a continuous RV then $\mathbb{P}(X = x) = 0 \quad \forall\, x$.

Proof.

$$\begin{aligned} & F_X(x) = F_X(x-) \\ \iff\ & \mathbb{P}(X \leq x) = \mathbb{P}(X < x) \quad \forall\, x \\ \iff\ & \mathbb{P}(X = x) = 0 \quad \forall\, x. \end{aligned}$$

□

Note. In this course we assume that F is also differentiable so that $F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\,du$.

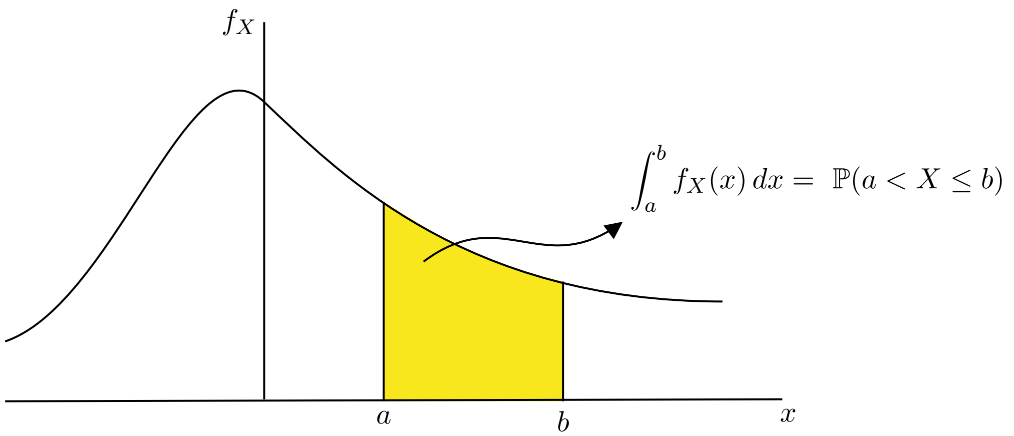
Definition 5.4 (Probability Density Function)

The **pdf** of RV X is $f_X : \mathbb{R} \to \mathbb{R}$ with properties

$$f_X(x) \geq 0 \quad \forall\, x$$
$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1.$$

Intuitive meaning

$$\begin{aligned} \mathbb{P}(x < X \leq x + \delta x) &= \int_x^{x+\delta x} f_X(u)\,du \approx \delta x \cdot f_X(x) \\ \mathbb{P}(a < X \leq b) &= \int_a^b f_X(x)\,dx \\ &= \mathbb{P}(a \leq X < b) \text{ since } \mathbb{P}(X = a) = \mathbb{P}(X = b) = 0. \end{aligned}$$



So for $S \subset \mathbb{R}$, $\mathbb{P}(X \in S) = \int_S f_X(u)\,du$. ($S$ "nice" e.g. interval or countable union of intervals)

Key takeaways:

- The CDF is a collection of probabilities
- The PDF is not a probability. We use them by integrating to get a probability.

Example 5.2 (Uniform Distribution)
Let $X \sim \mathrm{U}[a, b]$ where $a, b \in \mathbb{R}$ and $a < b$.

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

$$\begin{aligned} F_X(x) &= \int_a^x f_X(u)\,du \\ &= \frac{x-a}{b-a} \quad \text{for } a \le x \le b. \end{aligned}$$

Question
In what sense is this a "limit of discrete uniform RVs"?

Example 5.3 (Exponential Distribution)
Let $X \sim \mathrm{Exp}(\lambda)$ where $\lambda > 0$.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

It is easy to check that $f_X(x)$ is a pdf.

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \le x) \\ &= \int_0^x \lambda e^{-\lambda u}\,du \\ &= 1 - e^{-\lambda x}. \end{aligned}$$

The exponential distribution is the "limit of (rescaled) geometric distribution". It is a good way to model arrival times, "how long to wait before something happens" - link to Poisson usage which will be explored in Part II.

Claim 5.7 (Memoryless property)

(Conditional $\mathbb{P}$ works as before). Let $X \sim \text{Exp}(\lambda)$ and $s, t > 0$.

$$\mathbb{P}(X \geq s + t \mid X \geq s) = \mathbb{P}(X \geq t).$$

Proof.

$$\begin{aligned}
\mathbb{P}(X \geq s + t \mid X \geq s) &= \frac{\mathbb{P}(X \geq s + t)}{\mathbb{P}(X \geq s)} \\
&= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\
&= e^{-\lambda t} \\
&= \mathbb{P}(X \geq t).
\end{aligned}$$

□

Note. The only continuous memoryless distribution (with a density) is the exponential distribution.

§5.3 Expectation of Continuous RVs

Definition 5.5 (Expectation)

The **expectation** of X is $\int_{-\infty}^{\infty} x f_X(x)\,dx$ and $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$.[a]

[a] Technical comment: assumes at most one of $\int_{-\infty}^{0} |x| f_X(x)\,dx$ and $\int_0^{\infty} x f_X(x)\,dx$ is infinite.

Claim 5.8 (Linearity of Expectation)

$\mathbb{E}[\lambda X + \mu Y] = \lambda \mathbb{E}[X] + \mu \mathbb{E}[Y]$.

Claim 5.9

If $X \geq 0$ then $\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X \geq x)\,dx$.

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^{\infty} x f_X(x)\,dx \\
&= \int_{x=0}^{\infty} \left(\int_{u=0}^{x} 1\,du \right) f_X(x)\,dx \\
&= \int_{u=0}^{\infty} du \int_{x=u}^{\infty} f_X(x)\,dx^{a}
\end{aligned}$$

$$= \int_{u=0}^{\infty} du \mathbb{P}(X \geq u)$$

$\square$

[a]Consider the region of (u, x) space the second line integrates over and see this is the same as the third line.

§5.4 Variance of Continuous RVs

Definition 5.6 (Variance)
$\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$.

Claim 5.10
$\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$.

Example 5.4 (Uniform Distribution)
Let $X \sim \mathrm{U}[a, b]$.

$$\begin{aligned}
\mathbb{E}[X] &= \int_a^b x \frac{1}{b-a} \, dx \\
&= \frac{a+b}{2} \\
\mathbb{E}[X^2] &= \int_a^b x^2 \frac{1}{b-a} \, dx \\
&= \frac{1}{3}\left(a^2 + ab + b^2\right) \\
\operatorname{Var}(X) &= \frac{1}{3}\left(a^2 + ab + b^2\right) - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{(b-a)^2}{12}.
\end{aligned}$$

Example 5.5 (Exponential Distribution)
Let $X \sim \operatorname{Exp}(\lambda)$.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x \lambda e^{-\lambda x} \, dx \\
&= \left[-x e^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} \, dx
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\lambda}. \\
\mathbb{E}[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x}\, dx \\
&= \left[-x^2 e^{-\lambda x}\right]_0^\infty + 2\int_0^\infty x e^{-\lambda x}\, dx \\
&= 0 + \frac{2}{\lambda^2}. \\
\mathrm{Var}(X) &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\
&= \frac{1}{\lambda^2}.
\end{aligned}
$$

§5.5 Transformations of Continuous RVs

Goal:

- Let $U \sim \mathrm{Unif}[a,b]$ and $\tilde{U} \sim \mathrm{Unif}[0,1]$. We want to be able to write $U = (b-a)\tilde{U} + a$ and carry all calculations over.
- View $g(X)$ as a continuous RV with its own density.

Theorem 5.1

- Let X be a continuous RV with density f.
- Let $g : \mathbb{R} \to \mathbb{R}$ be continuous s.t.
  - g is either strictly increasing or strictly decreasing.
  - g^{-1} is differentiable.

Then $g(X)$ is a continuous RV with density

$$\hat{f}(x) = f(g^{-1}(x)) \cdot \left| \frac{d}{dx} g^{-1}(x) \right| \tag{2}$$

Remark 6.

- Density is? Something to integrate over to get a probability.
- Equation (2) <u>is</u> integration by substitution.
- The following proof uses CDFs (which <u>are</u> probabilities).

Proof. Assume g is strictly increasing, g strictly decreasing case is similar.

$$F_{g(X)}(x) = \mathbb{P}(g(X) \le x)$$

$$
\begin{aligned}
&= \mathbb{P}(X \le g^{-1}(x)) \\
&= F_X(g^{-1}(x)) \\
F'_{g(X)}(x) &= F'_X(g^{-1}(x))\frac{d}{dx}g^{-1}(x) \\
&= f(g^{-1}(x))\frac{d}{dx}g^{-1}(x).
\end{aligned}
$$

$\square$

Sanity check: We've got two expressions for $\mathbb{E}[g(X)]$, $\int_{-\infty}^{\infty} x\hat{f}(x)\,dx$ and $\int_{-\infty}^{\infty} g(x)f(x)\,dx$. (Assume $\mathrm{Im}(X) = \mathrm{Im}(g(X)) = (-\infty, \infty)$).

$$
\int_{-\infty}^{\infty} x\hat{f}(x)\,dx = \int_{-\infty}^{\infty} xf(g^{-1}(x))\frac{d}{dx}g^{-1}(x)\,dx.
$$

Substitute: $g^{-1}(x) = u$ so $du = dx\frac{d}{dx}g^{-1}(x)$.

$$
= \int_{-\infty}^{\infty} g(u)f(u)\,du.
$$

Example 5.6 (Exponential Distribution)

- Let $X \sim \mathrm{Exp}(\lambda)$ and $Y = cX$.

$$
\begin{aligned}
\mathbb{P}(Y \le x) &= \mathbb{P}\left(X \le \frac{x}{c}\right) \\
&= 1 - e^{-\lambda\frac{x}{c}} \\
&= 1 - e^{-\frac{\lambda}{c}x} - \text{CDF of Exp}\left(\frac{\lambda}{c}\right).
\end{aligned}
$$

- $$\hat{f}(x) = \frac{1}{c}f\left(\frac{x}{c}\right) = \frac{1}{c}\lambda e^{-\lambda x/c} = \frac{\lambda}{c}e^{-\frac{\lambda}{c}x} - \text{PDF of Exp}\left(\frac{\lambda}{c}\right).$$

Definition 5.7 (The Normal Distribution)

The range is $(-\infty, \infty)$. It has two parameters: $\mu \in (-\infty, \infty)$ and $\sigma^2 \in (0, \infty)$.

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

Definition 5.8 (Standard Normal)
The **standard normal distribution** is $Z \sim \mathcal{N}(0,1)$[a] where

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

[a]Sometimes may be referred to as $\varphi(x)$.

Notation. Φ denotes the CDF of Z.

Remark 7.

- $\frac{1}{\sqrt{2\pi}}$ is a "normalising constant" to ensure $\int f(x)\,dx = 1$.
- $e^{-x^2/2}$ has a very rapid decay as $x \to \pm\infty$, this helps ensure that the expected value of $g(\mathcal{N})$ is defined as $\int_0^\infty g()$ is finite.
- $\mathcal{N}(\mu, \sigma^2)$ used for modelling non-negative quantities, this is fine because if μ is large $\mathbb{P}(\mathcal{N}(\mu, \sigma^2) < 0)$ is very small.

Proof. Let us check that f_Z is actually a density

$$I = \int_{-\infty}^{\infty} e^{-x^2/2}\,dx$$

Clever idea is to use I^2 instead

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-u^2/2} e^{-v^2/2}\,du\,dv \\ &= \int \int e^{-\frac{u^2+v^2}{2}}\,du\,dv \end{aligned}$$

Polar coordinates: $u = r\cos\theta$ and $v = r\sin\theta$.

$$\begin{aligned} &= \int_{r=0}^{\infty} \int_{\theta=0}^{2\pi} r e^{-r^2/2}\,d\theta\,dr \\ &= 2\pi \int_{r=0}^{\infty} r e^{-r^2/2}\,dr \\ &= 2\pi. \end{aligned}$$

□

Claim 5.11
$\mathbb{E}[Z] = 0$.

Proof. Clear by symmetry, density is symmetric about origin, and its expectation is well-defined as tail decays rapidly. $\square$

Claim 5.12

$\mathrm{Var}(Z) = 1$.

Proof. STP: $\mathbb{E}[Z^2] = 1$.

$$\begin{aligned}
\mathbb{E}[Z^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot x e^{-x^2/2}\, dx \\
&= \underbrace{\left[-x \cdot \frac{e^{-x^2/2}}{\sqrt{2\pi}} \right]_{-\infty}^{\infty}}_{=0} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx \\
&= 1.
\end{aligned}$$

$\square$

§5.5.1 Studing $\mathcal{N}(\mu, \sigma^2)$ via linear transformation

Claim 5.13 (Facts about $X \sim \mathcal{N}(\mu, \sigma^2)$)

1. X has the same distribution as $\mu + \sigma Z$ where $Z \sim \mathcal{N}(0, 1)$.
2. X has CDF $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.
3. $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$.

Proof.

1. Let $g(x) = \mu + \sigma z$ so $g^{-1}(x) = \frac{x-\mu}{\sigma}$.
   Then $g(Z)$ has density

$$\begin{aligned}
f_{g(Z)}(x) &= f_Z(g^{-1}(x)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| \quad \text{by Theorem 5.1} \\
&= \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma} \right) \\
&= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.
\end{aligned}$$

2.
$$
\begin{aligned}
F_{g(Z)} &= \mathbb{P}(g(Z) \leq x) \\
&= \mathbb{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{x-\mu}{\sigma}\right).
\end{aligned}
$$

3. Use part (i) to get

$$
\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[\mu + \sigma Z] \\
&= \mu + \sigma \mathbb{E}[Z] \\
&= \mu \\
\mathrm{Var}(\mu + \sigma Z) &= \sigma^2 \,\mathrm{Var}(Z) \\
&= \sigma^2.
\end{aligned}
$$

□

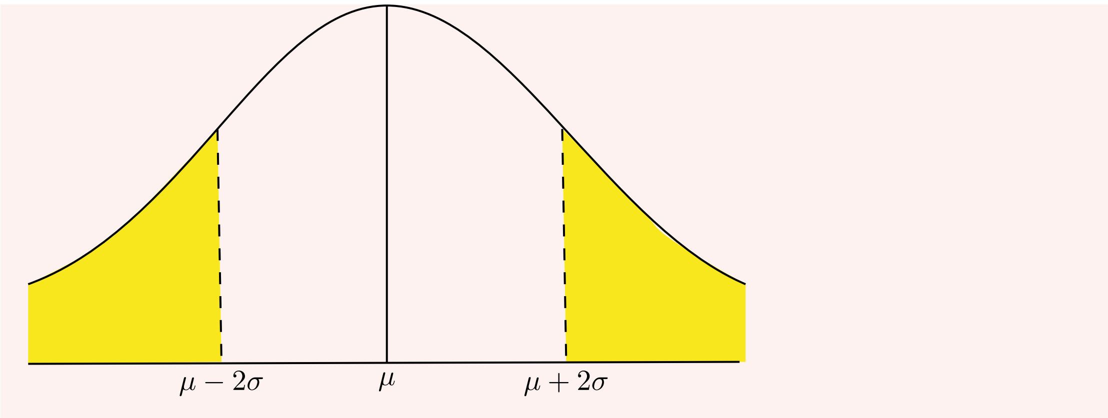
Remark 8. To calculate the cdf we only need to know Φ so you would only need to print out a table of values for Φ.

Example 5.7

Let $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$
\begin{aligned}
\mathbb{P}(a \leq X \leq B) &= \mathbb{P}\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\
&= \mathbb{P}\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).
\end{aligned}
$$

Special Case: $a = \mu - k\sigma$ and $b = \mu + k\sigma$ where $k \in \mathbb{N}$.
Then $\mathbb{P}(a \leq X \leq b) = \Phi(k) - \Phi(-k)$. "within k standard deviations of the mean".



Definition 5.9 (Median)
Suppose that X is a continuous RV, the **median** of X is the number m s.t.

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

In other words

$$\int_{-\infty}^{m} f(x)\ dx = \int_{0}^{\infty} f(x)\ dx = \frac{1}{2}$$

Remark 9.

- For $X \sim \mathcal{N}(\mu, \sigma^2)$ and other distributions symmetric about their mean, the median $m = \mathbb{E}[X]$.
- Sometimes $|X - m|$ better than $|X - \mu|$ for interpretation.

§5.6 More than one continuous RVs

Allow RVs to take values in $\mathbb{R}^n$.
E.g. $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ a RV.

Definition 5.10 (Multivarite Density Function)
We say that X has **multivariate density** $f : \mathbb{R} \to [0, \infty)$ if

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \underbrace{\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n}}_{\text{Integrate over } (-\infty, x_1] \times \cdots \times (-\infty, x_n]} f(u_1, \dots, u_n) \underbrace{\prod du_i}_{\text{i.e. } du_1 \dots du_n}$$

$$= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(u_1, \ldots, u_n) \prod du_i.$$

f is sometimes also called (especially for $n = 2$) a joint density function.

Consequence: This generalises: $\mathbb{P}((X_1, \ldots, X_n) \in A) = \int_A f(\boldsymbol{u})\, d\boldsymbol{u}$ for all "measurable" $A \in \mathbb{R}^n$.

Definition 5.11 (Independence)
We say that $X_1, \ldots, X_n$ are **independent** if $\forall x_1, \ldots, x_n$,

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

Goal: convert to statement about densities

Definition 5.12 (Marginal Density)
Let $X = (X_1, \ldots, X_n)$ have density f. The **marginal density** f_{X_i} of X_i is

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) \prod_{j \neq i} dx_j$$

"The density of X_i viewed as a RV by itself. We fix x_i and let everything else vary."

Theorem 5.2
Let $X = (X_1, \ldots, X_n)$ has density f.

1. If $X_1, \ldots, X_n$ are independent with marginals $f_{X_1}, \ldots, f_{X_n}$. Then $f(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n)$
2. Suppose that f factorises as $f(x_1, \ldots, x_n) = g_1(x_1) \ldots g_n(x_n)$ for some non-negative functions (g_i). Then $X_1, \ldots, X_n$ are independent and marginal $f_{X_i} \propto g_i$.

Proof.

1.
$$\begin{aligned}\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) &= \mathbb{P}(X_1 \leq x_1) \ldots \mathbb{P}(X_n \leq x_n) \\ &= \left[\int_{-\infty}^{x_1} f_{X_1}(u_1)\, du_1\right] \ldots \left[\int_{-\infty}^{x_n} f_{X_n}(u_n)\, du_n\right] \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \underbrace{\prod f_{X_i}(u_i)}_{\text{matches definition of } f} \prod du_i\end{aligned}$$

2. Idea:

   - replace $g_i(x)$ with $h_i(x) = \frac{g_i(x)}{\int g_i(u)\,du}$ so h_i *is* a density.
   - compute integrals for $\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$ and $\mathbb{P}(X_1 \leq x_1) \ldots \mathbb{P}(X_n \leq x_n)$ and show equality.

$\square$

§5.7 Transformation of Multiple RVs

Example 5.8

Let X and Y be independent RVs with densities f_X and f_Y respectively.
Goal: density of $Z = X + Y$.

$$\mathbb{P}(X + Y \leq z) = \iint_{\{x+y\leq z\}} f_{X,Y}(x,y)\,dx\,dy$$
$$= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z-x} f_X(x) f_Y(y)\,dx\,dy$$

Substitute $y' = y + x$

$$= \int_{x=-\infty}^{\infty} \left(\int_{y'=-\infty}^{z} f_X(x) f_Y(y' - x)\,dy' \right) dx$$

$y' \mapsto y$

$$= \int_{y=-\infty}^{z} dy \left(\int_{x=-\infty}^{\infty} f_X(x) f_Y(y - x)\,dx \right)$$

So the density of Z is:

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z - x) f_X(x)\,dx$$

We call this function the *convolution* of f_X and f_Y.

For X, Y discrete, non-negative and independent we would have

$$\mathbb{P}(X + Y = k) = \sum_{l=0}^{k} \mathbb{P}(X = \ell)\mathbb{P}(Y = k - \ell)$$

Definition 5.13 (Gamma Distribution)

The **gamma distribution** has two parameters $\lambda > 0$ and $n \in \mathbb{N}$. Its range is $[0, \infty)$. We say $X \sim \Gamma(n, \lambda)$ and has density

$$\begin{aligned} f_X(x) &= e^{-\lambda x} x^{n-1} \frac{\lambda^n}{(n-1)!} \\ n &= 1 \mapsto \mathrm{Exp}(\lambda) \\ n &= 2 \mapsto \lambda^2 x e^{-\lambda x} \end{aligned}$$

Example 5.9 (Exponential Distribution)
Let $X, Y \sim \mathrm{Exp}(\lambda)$ be IID and $Z = X + Y$.

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} \lambda e^{-\lambda(z-x)} \lambda e^{-\lambda x} \, dx \\ &= \int_0^z \lambda^2 e^{-\lambda z} \, dx \\ &= \lambda^2 z e^{-\lambda z}. \end{aligned}$$

So $X + Y \sim \Gamma(2, \lambda)$ and in fact the sum of n IID $\mathrm{Exp}(\lambda)$ has distribution $\Gamma(n, \lambda)$.

Example 5.10 (Normal Distribution)
Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent.
Then $X_1 + X_2 \sim \mathcal{N}(\mu_1, +\mu_2, \sigma_1^2 + \sigma_2^2)$ (We already know what the mean and variance of $X_1 + X_2$ is, the interesting part is that it is normal).
We can prove this using convolution but we will prove it using generating functions soon, Example 5.13.

Theorem 5.3
Let $X = (X_1, \ldots, X_n)$ be a RV on $D \in \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^n$ well-behaved and $U = g(X) = (U_1, \ldots, U_n)$. Assume the joint density $f_X(x)$ is continuous.
Then the joint density

$$f_U(\boldsymbol{u}) = f_X\left(g^{-1}(\boldsymbol{u})\right) |J(\boldsymbol{u})|$$

where J, the Jacobean, is

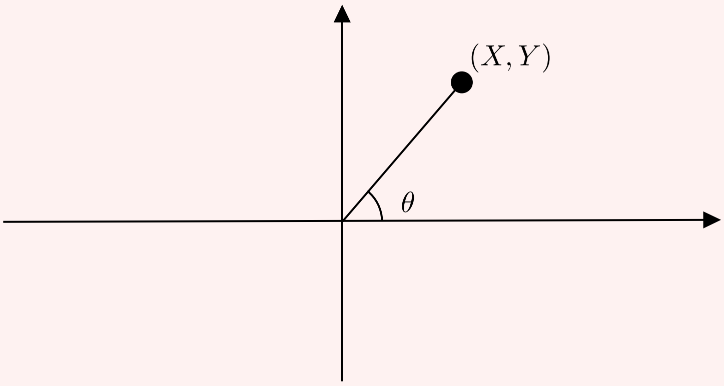
$$J = \det\left(\underbrace{\left(\frac{\partial[g^{-1}(\boldsymbol{u})]_i}{\partial u_j}\right)_{i,j=1}^{n}}_{n\times n \text{ matrix}}\right)$$

"Proof". Definition of multivariate integration by substitution. □

Top tip: $|\text{Jacobean of } g^{-1}| = \frac{1}{|\text{Jacobean of } g|}$.

Example 5.11 (Radial symmetry)

Let $X, Y \sim \mathcal{N}(0,1)$ be IID.
Let $(X,Y) = \underbrace{(R\cos\Theta, R\sin\Theta)}_{g^{-1}}$.
Range: $R > 0, \Theta \in [0, 2\pi)$.



$$\begin{aligned}
f_{X,Y}(x,y) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}} \\
&= \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}} \\
J &= \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} \\
&= r. \\
\implies f_{R,\Theta}(r,\theta) &= \frac{1}{2\pi}e^{-\frac{r^2}{2}} \times r \\
f_\Theta(\theta) &= \frac{1}{2\pi}
\end{aligned}$$

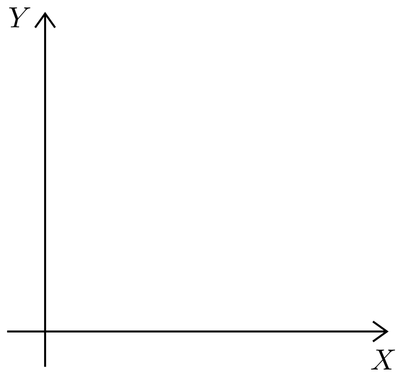
$$f_R(r) = e^{-\frac{r^2}{2}} r$$

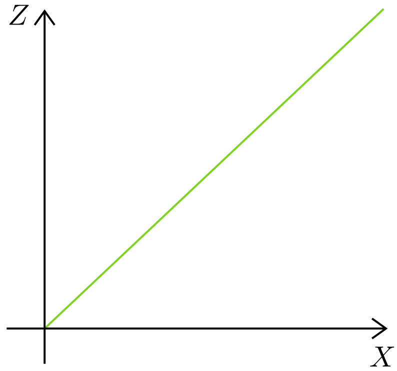
Thus Θ, R are independent and Θ is uniform on $[0, 2\pi)$.

Warning: Change of range.
Eg: $X, Y \geq 0$. $Z = X + Y$.

$$f_{X,Z}(x, z) = ?(x, z) 1_{z \geq x}$$

so X, Z are *not* independent even if ? splits as a product.





[5]

The support needs to be a rectangle as then the indicator function will split as a product.

§5.8 Moment Generating Function

Definition 5.14 (Moment Generating Function)
Let X have density f. The **MGF** of X is

$$\begin{aligned} m_X(\theta) &= \mathbb{E}[e^{\theta X}] \\ &= \int_{-\infty}^{\infty} e^{\theta x} f(x)\, dx \end{aligned}$$

whenever this is finite.

Note. $m_X(0) = 1$.

Theorem 5.4

[5]This graphic doesn't seem to render properly in edge.

The MGF uniquely determines distribution of a RV whenever it exists $\forall\, \theta \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$.

Definition 5.15 (Moment)
The nth moment of X is $\mathbb{E}[X^n]$.

Theorem 5.5
Suppose $m(\theta)$ exists $\forall\, \theta \in (-\varepsilon, \varepsilon)$. Then $m^{(n)}(0) = \frac{d^n}{d\theta^n} m(\theta)\Big|_{m=0} = \mathbb{E}[X^n]$.

Proof comment: Use $\frac{\partial^n e^{\theta x}}{\partial \theta^n} = x^n e^{\theta x}$.

Claim 5.14
Let $X_1, \ldots, X_n$ be independent and $X = X_1 + \cdots + X_n$. Then

$$\begin{aligned} m_X(\theta) &= \mathbb{E}[e^{\theta(X_1+\cdots+X_n)}] \\ &= \mathbb{E}[e^{\theta X_1}] \ldots \mathbb{E}[e^{\theta X_n}] \text{ by independence} \\ &= \prod m_{X_i}(\theta). \end{aligned}$$

Example 5.12 (Gamma Distribution)
Let $X \sim \Gamma(n, \lambda)$.

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

Goal: Reduce to integral of pdf over range!

$$\begin{aligned} &= \int_0^\infty e^{-(\lambda-\theta)x} x^{n-1} \times \frac{\lambda^n}{(n-1)!}\, dx \\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^n \int_0^\infty \underbrace{e^{-(\lambda-\theta)} x^{n-1} \frac{(\lambda-\theta)^n}{(n-1)!}}_{\text{pdf of } \Gamma(n,\lambda-\theta)^{a}}\, dx \\ &= \begin{cases} \left(\frac{\lambda}{\lambda-\theta}\right)^n & \theta < \lambda \\ \infty & \theta \geq \lambda. \end{cases} \end{aligned}$$

So $\mathrm{Exp}(\lambda)$ has MGF $\frac{\lambda}{\lambda-\theta}$. And we've proved that the sum of n IID $\mathrm{Exp}(\lambda)$ has distribution $\Gamma(n, \lambda)$.

[a] provided $\theta < \lambda$.

Example 5.13 (Normal Distribution)
Let $X \sim \mathcal{N}(\mu, \sigma^2)$.

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

The proof is left as an exercise, try relating integral to integral of pdf of some normal distribution.
Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent. Then

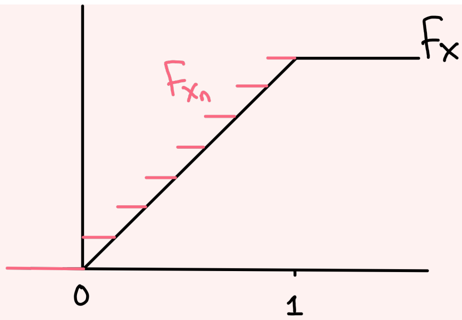
$$\begin{aligned} m_{X_1+X_2}(\theta) &= \exp\left(\theta\mu_1 + \frac{\theta^2\sigma_1^2}{2}\right) \exp\left(\theta\mu_2 + \frac{\theta^2\sigma_2^2}{2}\right) \\ &= \underbrace{\exp\left(\theta(\mu_1+\mu_2) + \frac{\theta^2}{2}(\sigma_1^2+\sigma_2^2)\right)}_{\text{MGF of } \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)} \end{aligned}$$

§5.9 Convergence of RVs

Definition 5.16 (Convergence in Distribution)

Let $(X_n)_{n\geq 1}$ and X be RVs. X_n *converges to* X *in distribution*, $X_n \xrightarrow{d} X$, if $F_{X_n}(x) \to F_X(x)$ for all $x \in \mathbb{R}$ which are continuity points of F_X.

Example 5.14
Let $X_n = \frac{1}{n}\,\mathrm{Unif}\,(1, \ldots, n)$ and $X \sim \mathrm{Unif}[0, 1]$.



F_X continuous, $F_{X_n} \to F_X(x)$ holds $\forall\, x \in [0,1]$ by picture so $X_n \xrightarrow{d} X$.

Example 5.15

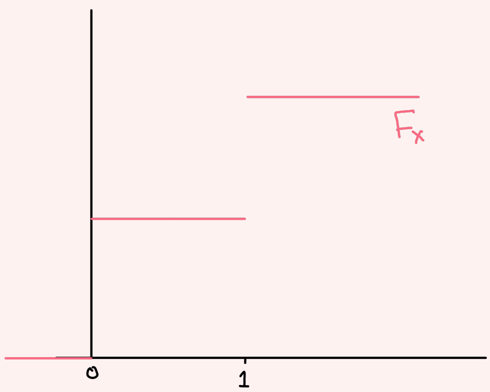
$$X_n = \begin{cases} 0 & \mathbb{P} = \frac{1}{2} \\ 1 + \frac{1}{n} & \mathbb{P} = \frac{1}{2} \end{cases}$$

$$F_{X_n}(x) = \begin{cases} \frac{1}{2} & x \in (0,1) \\ \frac{1}{2} & x = 1 \\ 1 & x > 1 \text{ when } n \text{ is large} \end{cases}$$

$$\text{Let } X \sim \mathrm{Bern}\left(\frac{1}{2}\right)$$

$$F_X(1) = 1.$$

But F_X has a discontinuity at $x = 1$ so $X_n \xrightarrow{d} X$.



(I.e. deterministic convergence of a sequence of real numbers is an example of convergence in distribution)

Claim 5.15

If X is a constant c, then convergence in distribution is equivalent to: $\forall\ \varepsilon > 0$: $\mathbb{P}(|X_n - c| > \varepsilon) \to 0$ as $n \to \infty$. "convergence in probability to constant".

Claim 5.16

If X is a continuous RV with $X_n \xrightarrow{d} X$ then $\mathbb{P}(a \le X_n \le b) \to \mathbb{P}(a \le X \le b)$ for all $a, b \in [-\infty, \infty]$.

Warning: Does not say that densities converge, e.g. in Example 5.14, X_n does not have a density.

§5.10 Laws of Large Numbers

$$\frac{S_n}{n}\text{“} \to \text{”}\mu$$

Theorem 5.6 (Weak LLN)

Let $(X_n)_{n\ge1}$ be IID with $\mu = \mathbb{E}[X_1]$ finite. Set $S_n = X_1 + \cdots + X_n \quad \forall\, n \ge 0$. Then

$$\forall\, \varepsilon > 0 : \ \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty$$

Proof. (Assume $\mathrm{Var}(X_1) = \sigma^2 < \infty$)

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{S_n}{n} - \mu^{a}\right| > \varepsilon\right) &= \mathbb{P}\left(|S_n - n\mu| > \varepsilon n\right) \\
&\le \frac{\mathrm{Var}(S_n)}{\varepsilon^2 n^2} \quad \text{by Chebyshev's Inequality} \\
&= \frac{n\sigma^2}{\varepsilon^2 n^2} \to 0 \ \text{ as } n \to \infty.
\end{aligned}$$

$\varepsilon > 0$ is fixed we are not taking limit as $\varepsilon \to 0$. □

[a] Also $\mathbb{E}\left[\frac{S_n}{n}\right]$

§5.11 Central Limit Theorem

Theorem 5.7 (Central Limit Theorem)

Let $(X_n)_{n\ge1}$ be IID with $\mu = \mathbb{E}[X_1]$ finite and $\sigma^2 < \infty$. Set $S_n = X_1 + \cdots + X_n \quad \forall\, n \ge$

0. Then

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathcal{N}(0,1) \text{ as } n \to \infty.$$

Discussion: Three stage summary

1. Distribution of S_n concentrated on $n\mu$ - we already know this from WLLN.
2. Fluctuations around $n\mu$ have order $\sqrt{n}$ - new and important
3. Shape is normal - detail.

We use CLT by

1. $S_n \overset{d}{\approx} \mathcal{N}(n\mu, n\sigma^2)$
2. $$\begin{aligned} \mathbb{P}(a \le S_n \le b) &= \mathbb{P}\left(\frac{a-n\mu}{\sqrt{n\sigma^2}} \le \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \le \frac{b-n\mu}{\sqrt{n\sigma^2}}\right). \\ &= \mathbb{P}\left(\frac{a-n\mu}{\sqrt{n\sigma^2}} \le Z \le \frac{b-n\mu}{\sqrt{n\sigma^2}}\right). \end{aligned}$$

Theorem 5.8 (Continuity Theorem for MGFs)
Let (X_n) and X have MGFs m_{X_n} and m_X. If

- $m_X(\theta) < \infty$ for $\theta \in (-\varepsilon, \varepsilon)$.
- $m_{X_n}(\theta) \to m_X(\theta) \quad \forall\, \theta$ s.t. $m_X(\theta) < \infty$.

Then $X_n \xrightarrow{d} X$.

Proof. Part II Probability and Measure. □

Idea: Expand $m_X(\theta)$ as a Taylor series around 0.

$$\begin{aligned} m_X(\theta) &= 1 + m_X'(0)\theta + \frac{m_X''(0)}{2!}\theta^2 + \dots \\ &= 1 + \theta\mathbb{E}[X] + \frac{1}{2}\theta^2\mathbb{E}[X^2] + o(\theta^2) \end{aligned}$$

Proof - WLLN via MGFs.

Comment: We know MGF of S_n, we want to study the MGF of S_n/n.

$$m_{S_n/n}(\theta) = \mathbb{E}\left[\exp\left(\theta\frac{S_n}{n}\right)\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\exp\left(\frac{\theta}{n}S_n\right)\right] \text{ Key step.} \\
&= m_{S_n}(\theta/n) \\
&= m_{X_1}(\theta/n)\dots m_{X_n}(\theta/n) \\
&= \left(1 + \mu\frac{\theta}{n} + o\left(\frac{1}{n}\right)\right)^n \\
&\to e^{\mu\theta}
\end{aligned}
$$

$e^{\mu\theta}$ is the MGF of the RV $X = \mu$ with $\mathbb{P} = 1$, so $\frac{S_n}{n} \xrightarrow{d} \mu$ by the continuity theorem. □

Proof - CLT via MGFs.

Assume WLOG $\mu = 0$ and $\sigma^2 = 1$. (so $\mathbb{E}[X_i^2] = 1$). In general $X \mapsto \frac{X-\mu}{\sqrt{\sigma^2}}$

STP: $S_n/\sqrt{n} \xrightarrow{d} \mathcal{N}(0,1)$.

$$
\begin{aligned}
m_{X_i}(\theta) &= 1 + \frac{\theta^2}{2} + o(\theta^2) \\
m_{S_n/\sqrt{n}}(\theta) &= \mathbb{E}\left[\exp\left(\theta\frac{S_n}{\sqrt{n}}\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\theta}{\sqrt{n}}S_n\right)\right] \\
&= m_{S_n}(\theta/\sqrt{n}) \\
&= m_{X_1}(\theta/\sqrt{n})\dots m_{X_n}(\theta/\sqrt{n}) \\
&= \left(1 + \frac{\theta^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \\
&\to e^{\theta^2/2}
\end{aligned}
$$

$e^{\theta^2/2}$ is the MGF of the $\mathcal{N}(0,1)$. □

Theorem 5.9 (Strong LLN)
Let $(X_n)_{n\geq 1}$ be IID with $\mu = \mathbb{E}[X_1]$ finite and $\sigma^2 < \infty$. Set $S_n = X_1 + \cdots + X_n \quad \forall\, n \geq 0$. Then

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

"almost sure convergence" or "convergence with probability 1".

§5.12 Inequalities for $\mathbb{E}[f(X)]$

Motivation: For $f(x) = x^2$ we know $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ as $\mathrm{Var}(X) \geq 0$.

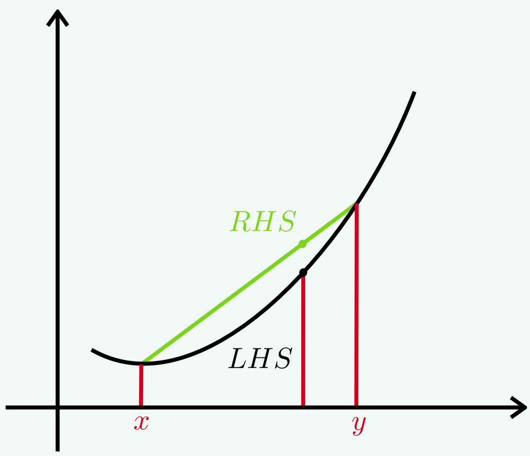
Question

What about general f?

Definition 5.17 (Convex Function)

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

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



Aside: The region above f is convex in $\mathbb{R}^2$

Definition 5.18 (Stricly Convex Function)

A function $f : \mathbb{R} \to \mathbb{R}$ is **stricly convex** if $\forall\, x, y \in \mathbb{R}$ and $t \in (0, 1)$

$$f\left(tx + (1-t)y\right) < tf(x) + (1-t)f(y).$$

Lemma 5.1 (Existence Of Subdifferential)

If $f : \mathbb{R} \to \mathbb{R}$ convex then $\forall\, y \quad \exists$ line $l(x) = mx + c$ s.t.

- $l(x) \leq f(x) \quad \forall\, x$
- $l(y) = f(y)$

Warning: not yet claiming l is a tangent.

Proof. Convexity $\implies \forall\, x < y < z$ insertpicture

$$\frac{f(y)-f(x)}{y-x} \leq \frac{f(z)-f(y)}{z-y}$$
$$\text{Let } M^- = \sup_{x<y} \frac{f(y)-f(x)}{y-x}$$
$$M^+ = \inf_{z>y} \frac{f(z)-f(y)}{z-y}$$
$$M^- \leq M^+.$$

Any value $m \in [M^-, M^+]$ works as the gradient of l. $\square$

Definition 5.19 (Concave Function)

f is concave iff $-f$ is convex.

Claim 5.17

If f is twice differentiable:

$$f \text{ convex } \iff f''(x) \geq 0 \quad \forall\, x.$$

Example 5.16

$$f(x) = \frac{1}{x} \text{ is convex on } (0, \infty)$$
$$\text{is concave on } (-\infty, 0).$$

Theorem 5.10 (Jensen's Inequality)

Let X be a RV and f a convex function:

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

(reverse the inequality if f concave).

Proof. Set $y = \mathbb{E}[X]$ as in Existence Of Subdifferential, so $\exists\, l(x) = mx + c$ s.t. $l(y) = f(y) = f(\mathbb{E}[X])$ and $f \geq l$.

$$\begin{aligned}\mathbb{E}[f(X)] &\geq \mathbb{E}[l(X)] \\ &= \mathbb{E}[mX + c]\end{aligned}$$

$$
\begin{aligned}
&= m\mathbb{E}[X] + c \\
&= my + c \\
&= l(y) \\
&= f(\mathbb{E}[X]).
\end{aligned}
$$

□

Claim 5.18

If f is strictly convex then $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$ iff $X = \mathbb{E}[X]$ with $\mathbb{P} = 1$, i.e. constant RV.

Informal comment: Jensen's inequality is better than most other inequalities as many can be derived from Jensen's.

§5.13 Application to sequences

Definition 5.20 (AM - GM inequality)

Let $x_1, \ldots, x_n \in (0, \infty)$.

$$
\frac{x_1 + \cdots + x_n}{n} \geq \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}
$$

Proof of $n = 2$.

$$
\begin{aligned}
0 &\leq (x - y)^2 \\
&= x^2 - 2xy + y^2 \\
&= x^2 + 2xy + y^2 - 4xy \\
&= (x + y)^2 - 4xy
\end{aligned}
$$

□

Proof. Let X be a RV taking values $x_1, \ldots, x_n$ each with probability $\frac{1}{n}$. Let $f(x) = -\log x$, which is convex by second derivative.

$$
\begin{aligned}
\mathbb{E}[f(X)] &\geq f(\mathbb{E}[X]) \quad \text{by Jensen's Inequality} \\
-\frac{\log x_1 + \ldots \log x_n}{n} &\geq -\log\left(\frac{x_1 + \cdots + x_n}{n}\right) \\
\log\left((x_1 \ldots x_n)^{\frac{1}{n}}\right) &\leq \log \frac{x_1 + \ldots x_n}{n}
\end{aligned}
$$

$\log x$ and e^x are increasing.

$$\left(\prod x_i\right)^{\frac{1}{n}} \leq \frac{x_1 + \cdots + x_n}{n}.$$

□

§5.14 Sampling a Continuous RV

Theorem 5.11

Let X be a continuous RV with CDF F. Then if $U \sim \mathrm{U}[0,1]$, we have $Y = F^{-1}(U) \sim X$[a].

[a] Remember that U is a RV and F is just a increasing function

Proof. Goal: Find CDF of Y.

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

rearrange within $\mathbb{P}()$

$$\begin{aligned} &= \mathbb{P}(U \leq F(x)) \\ &= F(x) \end{aligned}$$

CDF of Y = CDF of X, so $Y \sim X$. □

§5.15 Rejection Sampling

Sampling uniformly on $[0,1]^d$ is easy, we simply take $(U^{(1)}, \ldots, U^{(d)})$ IID where $U^i \sim \mathrm{U}[0,1]$.

Question

How do we sample uniformly on A? insertpicture

Goal:

$$\begin{aligned} f(x) &= \begin{cases} \frac{1}{\mathrm{area}(A)} & x \in A \\ 0 & x \notin A. \end{cases} \\ &= \frac{1_A}{\mathrm{area}(A)}. \end{aligned}$$

In higher dimensions, use $\mathrm{volume}(A)^{-1}$.

Let $U_1, U_2, \dots$ be IID uniform on $[0,1]^d$ and let $N = \min\{n : U_n \in A\}$.

Claim 5.19

U_N is uniform on A. (i.e. has density f)

Proof. Note $\mathbb{P}(N < \infty) = 1$ if $\text{area}(A) > 0$.
STP: $\mathbb{P}(U_N \in B) = \int_B f(x)\,dx = \frac{\text{area}(B)}{\text{area}(A)} \quad \forall\, B \subset A$ with a well-defined area.

$$\begin{aligned}
\mathbb{P}(U_N \in B) &= \sum_{n\geq 1} \mathbb{P}(U_N \in B, N = n) \ \text{ by Law of Total Probability} \\
&= \sum_{n\geq 1} \mathbb{P}(U_1 \notin A, \dots, U_{n-1} \notin A, U_n \in B) \\
&= \sum_{n\geq 1} \mathbb{P}(U_1 \notin A)^{n-1}\mathbb{P}(U_n \in B) \ \text{ as } U_i \text{ are independent} \\
&= \sum_{n\geq 1} (1 - \text{area}(A))^{n-1}\,\text{area}(B) \\
&= \frac{\text{area}(B)}{1 - (1 - \text{area}(A))} \\
&= \frac{\text{area}(B)}{\text{area}(A)}.
\end{aligned}$$

□

Claim 5.20

Let X be a continuous RV on $[0,1]$ with *bounded* density f_X.
Let $A = \{(x,y) : x \in [0,1], y \leq f_X(x)\}$, i.e. the green region.
Let $U = (U^{(1)}, U^{(2)})$ be uniform on A.
Then $U^{(1)} \sim X$.

insertpicture

Proof.

$$\begin{aligned}
\mathbb{P}(U^1 \leq u) &= \mathbb{P}(U \in \text{blue region}) \\
&= \text{area}(\{(x,y) : x \leq u, y \leq f_X(x)\}) \\
&= \int_0^u f_X(x)\,dx \\
&= F_X(u).
\end{aligned}$$

So the CDF of $U^{(1)}$ is F_X. □

Claim 5.21

Let X be a continuous RV on $[-k,k]^d$ with *bounded* density f_X.
Let $A = \{(\boldsymbol{x}, y) : \boldsymbol{x} \in [-k,k]^d, y \leq f_X(x)\} \in \mathbb{R}^{d+1}$.
Let $U = (\boldsymbol{U}, U^+)$ be uniform on A.
Then $\boldsymbol{U} \sim X$.

§5.16 Multivariate Normal Distribution

Definition 5.21 (Gaussian RV)
A RV X is **Gaussian** if $X \sim \mathcal{N}(\mu, \sigma^2)$.[a]

[a] X is a one dimensional normal.

Recall if X, Y are independent Gaussians then $bX + cY$ is Gaussian, Example 5.13.

Question
Does there $\exists$ joint RVs (X, Y) s.t. X, Y both Gaussian but $X + Y$ is not?

Answer
Yes but the answer is annoying and doesn't have any real physical interpretation.

Question
Can we have dependent X, Y s.t. $bX + cY$ still holds?

Answer
Yes.

Definition 5.22 (Gaussian Random Vector)
A random vector (X, Y) is **Gaussian** if $bX + cY$ are a Gaussian RV $\forall\, b, c \in \mathbb{R}$.

§5.16.1 Linear Algebra Rewrite

Definition 5.23 (Gaussian Random Vector)
Random vector $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is **Gaussian** if $u^T X$ is a Gaussian RV $\forall\, u \in$

$\mathbb{R}^n$.

Let $\mu = \mathbb{E}[X] \in \mathbb{R}^n$.

Definition 5.24 (Covariance Matrix)
The **covariance matrix** V is

$$V = (\mathrm{Cov}(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$$

For $n = 2 : V = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y) \end{pmatrix}$.

Claim 5.22
The covariance matrix is symmetric.

Claim 5.23
If X is a Gaussian random vector then $u^T X \sim \mathcal{N}(u^T\mu, u^T V u)$.

Definition 5.25 (Moment Generating Function in $\mathbb{R}^n$)
Let $X \in \mathbb{R}^n$ be a RV. The **MGF** of X is

$$m_X(u) = \mathbb{E}\left[e^{u^T X}\right]$$

whenever this is finite.

Theorem 5.12
The MGF uniquely determines distribution of a RV whenever it exists $\forall\ u \in (-\varepsilon, \varepsilon)^n$ for some $\varepsilon > 0$.

Claim 5.24
If X Gaussian $m_X(u) = m_{u^T X}(1) = \exp\left(u^T\mu + \frac{1}{2}u^T V u\right)$.

Logical overview: $X \in \mathbb{R}^n$ Gaussian.

- distribution defined by MGF
- MGF defined by μ and V
  $\implies$ distribution of X defined by μ and V.

Remark 10. Density is

$$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sqrt{\det(V)}} \exp\left(-\frac{1}{2}(x-\mu)^T V(x-\mu)\right)$$

Claim 5.25

Return to $n = 2$: For a Gaussian vector (X_1, X_2) it is independent $\iff$ $\mathrm{Cov}(X_1, X_2) = 0$. $\Longleftarrow$ is false in general.

Why useful? Imagine X_1, X_2 describe real-world parameters, e.g. height vs 1 km rowing time.

- Independence would be an interesting conclusion
- Cov can be sampled.

Proof. $X = (X_1, X_2)$ is independent iff $m_X((u_1, u_2))$ splits as a product $m_1(u_1)m_2(u_2)$.

$$\begin{aligned} \exp\left(u^T\mu\right) &= \exp(u_1\mu_1)\exp(u_2\mu_2) \\ \exp\left(\frac{1}{2}u^T V u\right) &= \exp\left(\frac{1}{2}u_1^2\sigma_1^2\right)\exp\left(\frac{1}{2}u_2^2\sigma_2^2\right)\exp(u_1u_2\,\mathrm{Cov}(X_1, X_2)) \end{aligned}$$

Therefore splits iff $\mathrm{Cov} = 0$. □

Motivation: $\mathrm{Cov}(100X_1, X_2) = 100\,\mathrm{Cov}(X_1, X_2)$, so "large covariance" doesn't imply "very dependent".

Definition 5.26 (Correlation)

The **correlation** of X, Y is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \in [-1, 1].$$

Proposition 5.1

If (X, Y) Gaussian, then $Y = aX + Z$ where Z is Gaussian and (X, Z) independent.

Proof. Define $Z = Y - aX$ for $a \in \mathbb{R}$.

Claim 5.26

(X, Z) is Gaussian

Proof.

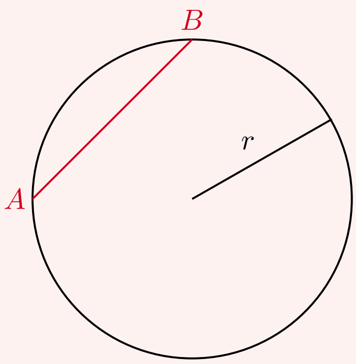
$$\begin{aligned} u_1X + u_2Z &= u_1X + u_2(Y - aX) \\ &= (u_1 - au_2)X + u_2Y. \end{aligned}$$

□

Goal: find a s.t. $\mathrm{Cov}(X, Z) = 0$ $\mathrm{Cov}(X, Z) = \mathrm{Cov}(X, Y - aX) = \mathrm{Cov}(X, Y) - a\,\mathrm{Var}(X)$ so take $a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$. Then $\mathrm{Cov}(X, Z) = 0 \implies X, Z$ independent. □

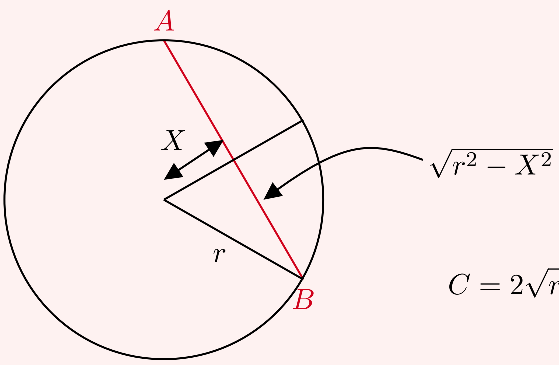
§5.17 Bertrand's Paradox

Example 5.17



Draw a chord at random.
What is the probability it has length $\leq r$?

1^{st} interpretation: Let $X \sim U[0, r]$



$C = 2\sqrt{r^2 - X^2}$

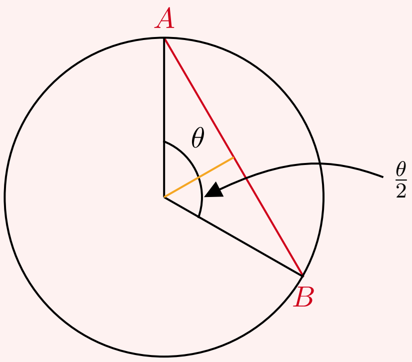
Let $C = |AB|$. What is $\mathbb{P}(C \leq r)$?

$$C = 2\sqrt{r^2 - X^2}$$

$$\begin{aligned}
\mathbb{P}(C \le r) &= \mathbb{P}(2\sqrt{r^2 - X^2} \le r) \\
&= \mathbb{P}(4(r^2 - X^2) \le r^2) \\
&= \mathbb{P}(4X^2 \ge 3r^2) \\
&= \mathbb{P}(X \ge \sqrt{3}3/2) \\
&= 1 - \frac{\sqrt{3}}{2} \\
&\approx 0.134
\end{aligned}$$

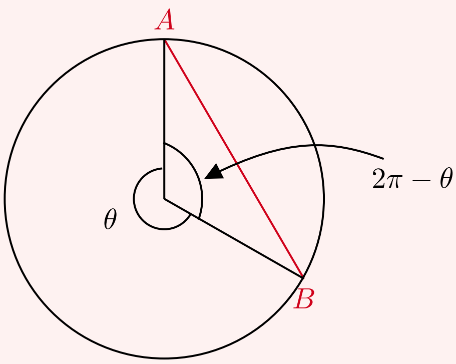
Example 5.18 (cont.)

2nd interpretation: Let $\theta \sim [0, 2\pi]$
Let $C = |AB|$
If $\theta \in [0, \pi]$:



$$C = 2r \sin \frac{\theta}{2}$$

If $\theta \in [\pi, 2\pi]$:

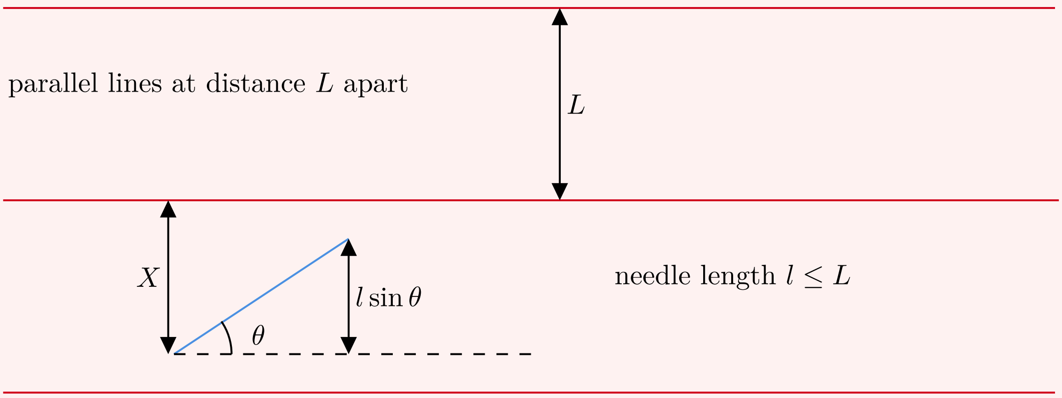


$$C = 2r \sin \frac{2\pi - \theta}{2} = 2r \sin \frac{\theta}{2}$$

$$\begin{aligned}
\mathbb{P}(C \le r) &= \mathbb{P}(2r\sin\frac{\theta}{2} \le r) \\
&= \mathbb{P}(\sin\frac{\theta}{2} \le \frac{1}{2}) \\
&= \mathbb{P}(\theta \le \frac{\pi}{3}) + \mathbb{P}(\theta \ge \frac{\pi}{3}) \\
&= \frac{1}{6} + \frac{1}{6} \\
&= \frac{1}{3} \\
&\approx 0.333\ldots
\end{aligned}$$

§5.18 Buffon's Needle

Example 5.19



Throw the needle at random. What is the probability it intersects at least one line?

$$\theta \sim U[0,\pi],\ X \sim U[0,L] \text{ indep.}$$

It intersects a line iff $X \le l\sin\theta$.

$$\mathbb{P}(\text{intersection}) = \mathbb{P}(X \le l\sin\theta) = \int_0^L \int_0^\pi \frac{1}{\pi L} 1(x \le l\sin\theta)\,\mathrm{d}x\,\mathrm{d}\theta = \frac{2l}{\pi L}$$

So $p = \frac{2l}{\pi L}$

$$\implies \pi = \frac{2l}{pL}$$

Want to use this experiment to approximate π. Throw n needles indep. and let $\hat{p}_n$

be the proportion intersecting a line. Then $\hat{p}_n$ approximates p and so

$$\hat{\pi}_n = \frac{2l}{\hat{p}_n} L \text{ approximates } \pi$$

Suppose

$$\mathbb{P}(|\hat{\pi}_n - \pi| \leq 0.001) \geq 0.99$$

How large should n be?

Example 5.20 (cont.)

$S_n =$ number of needles intersecting a line

$$S_n \sim \text{ Bin}(n, p)$$

By the CLT, $S_n \sim np + \sqrt{np(1-p)} \cdot Z, Z \sim N(0,1)$

$$\hat{p}_n = \frac{S_n}{n} \approx p + \sqrt{\frac{p(1-p)}{n}} \cdot Z$$

So

$$\hat{p}_n - p \approx \sqrt{\frac{p(1-p)}{n}}.$$

Define $f(x) = \frac{2l}{xL}$. Then $f(p) = \pi$ and $f'(p) = -\pi/p$ and $\hat{\pi}_n = f(\hat{p}_n)$.
By Taylor expansion, $\hat{\pi}_n = f(\hat{p}_n) \approx f(p) + (\hat{p}_n - p)f'(p)$

$$\implies \hat{\pi}_n \approx \pi - (\hat{p}_n - p) \cdot \frac{\pi}{p}$$

$$\implies \hat{\pi}_n - \pi \approx -\frac{\pi}{p}\sqrt{\frac{p(1-p)}{n}} = -\pi\sqrt{\frac{1-p}{pn}} \cdot Z$$

We want

$$\mathbb{P}\left(\pi\sqrt{\frac{1-p}{pn}} \cdot |Z| \leq 0.001\right) \geq 0.99$$

Have $\mathbb{P}(|Z| \geq 2.58) = 0.01$ and $\pi^2 \cdot \frac{1-p}{pn}$ decreasing in p. Minimise $\pi^2 \cdot \frac{1-p}{pn}$ by taking $l = L \implies p = \frac{2}{\pi}$ and

$$= \frac{\pi^2}{n}\left(\frac{\pi}{2} - 1\right)$$

Taking

$$\sqrt{\frac{\pi^2}{n}\left(\frac{\pi}{2} - 1\right)} \cdot 2.58 = 0.001 \implies n = 3.75 \times 10^7$$