# Probability

Hasan Baig

Lent 2021

## Contents

# 1 Probability Spaces

**Definition.** Suppose $\Omega$ is a set and $\mathcal{F}$ is a collection of subsets of $\Omega$.
We call $\mathcal{F}$ a ***$\sigma$*-algebra** if:
(i) $\Omega \in \mathcal{F}$
(ii) if $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$
(iii) for any countable collection $(A_n)_{n\geq 1}$ with $A_n \in \mathcal{F}\ \forall n$, we must also have that $\bigcup\limits_n A_n \in \mathcal{F}$

**Definition.** Suppose $\mathcal{F}$ is a $\sigma$-algebra on $\Omega$. A function $\mathbb{P} : \mathcal{F} \to [0,1]$ is called a **probability measure** if
(i) $\mathbb{P}(\Omega) = 1$
(ii) for any countable disjoint collection $(A_n)_{n\geq 1}$ in $\mathcal{F}$ with $A_n \in \mathcal{F}\ \forall n$, we have

$$\mathbb{P}(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n)$$

We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space. $\Omega$ is the sample space
$\mathcal{F}$ a $\sigma$-algebra
$\mathbb{P}$ the probability measure

**Note.** We say $\mathbb{P}(A)$ is the probability of $A$

**Remark.** When $\Omega$ countable, we take $\mathcal{F}$ to be all subsets of $\Omega$

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

**Remark.** We talk about probability of events and not outcomes.
Pick a uniform number from $[0,1]$

Properties of $\mathbb{P}$ (immediate from the definition):
- $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(\varnothing) = 0$
- 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.** Rolling a fair die
$\Omega = \{1, 2, \ldots, 6\}$, $\mathcal{F}$ = all subsets of $\Omega$.
$\mathbb{P}(\{\omega\}) = \frac{1}{6}\ \forall w \in \Omega$ and if $A \subseteq \Omega$, then $\mathbb{P}(A) = \frac{|A|}{6}$
(all outcomes equally likely)

**Example.** Equally likely outcomes
Let $\Omega$ be a finite set, $\Omega = \{\omega_1, \ldots, \omega_n\}$, $\mathcal{F}$ = all subsets.
Define $\mathbb{P} : \mathcal{F} \to [0,1]$ by $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$
In classical probability, this models picking a random element of $\Omega$.
$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}\ \forall \omega \in \Omega$

**Example.** Picking balls from a bag
Suppose we have $n$ balls with $n$ labels from $\{1, \ldots, n\}$ indistinguishable by touch.
Pick $k \leq n$ balls at random (all outcomes equally likely) without replacement.
Take $\Omega = \{A \subseteq \{1, \ldots, n\} : |A| - k\}$ $|\Omega| = \binom{n}{k}$
$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}$.

**Example.** Deck of cards
Take a well-shuffled (all possible permutations equally likely) deck of 52 cards.
$\Omega = \{\text{all permutations of 52 cards}\}$ $|\Omega| = 52!$
$\mathbb{P}(\text{top 2 cards are aces}) = \frac{4 \times 3 \times 50!}{52!} = \frac{1}{221}$

**Example.** Largest digit
Consider a string of $n$ random digits from $0, \ldots, 9$ (every digit can be any from $0, \ldots, 9$)
$\Omega = \{0, 1, \ldots, 9\}^n$ $|\Omega| = 10^n$
$A_k = \{\text{no digit exceeds } k\}$ and $B_k = \{\text{largest digit is } k\}$
$\mathbb{P}(B_k) = \frac{|B_k|}{|\Omega|}$
Notice $B_k = A_k \backslash A_{k-1}$
$|A_k| = (k+1)^n \implies |B_k| = (k+1)^n - k^n$
So $\mathbb{P}(B_k) = \frac{(k+1)^n - k^n}{10^n}$

**Example.** Birthday problem
There are $n$ people. What is the probability that at least 2 of these share the same birthday.
Assume nobody is born on 29/02. Also assume each birthday is equally likely. So $\Omega = \{1, \dots, 365\}^n$
$\mathcal{F}$ = all subsets.
Since all outcomes are equally likely, we take

$$\underset{\omega \in \Omega}{\mathbb{P}(\{\omega\})} = \frac{1}{365^n}$$

$A = \{\text{at least 2 people share same birthday}\}$
$A^C = \{\text{all } n \text{ birthdays are different}\}$.
Since $\mathbb{P}(A) = 1 - \mathbb{P}(A^C)$, it suffices to calculate $\mathbb{P}(A^C)$

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

And hence:

$$\mathbb{P}(A) = 1 - \frac{365 \times 364 \times \dots \times (365 - n + 1)}{365^n}$$

**Note.** $n = 22 \implies \mathbb{P}(A) \approx 0.476$
$n = 23 \implies \mathbb{P}(A) \approx 0.507$

## 1.2 Combinatorial Analysis

$\Omega$ finite set & suppose $|\Omega| = n$.
Want to partition $\Omega$ into $k$ disjoint subsets $\Omega_1 \dots \Omega_k$ with $|\Omega_i| = n_i$ and $\sum_{i=1}^{k} n_i = n$.
How many ways are there?

$$M = \# \text{ of ways}$$

$$M = \binom{n}{n_1}\binom{n - n_1}{n_2} \dots \binom{n - (n_1 + \dots + n_{k-1})}{n_k} = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}$$

We write

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

Strictly increasing and increasing functions $f = \{1, \ldots, k\} \to \{1, \ldots, n\}$ is strictly increasing if whenever $x < y$, then $f(x) < f(y)$.
$f$ is called increasing if whenever $x < y$, then $f(x) \leq f(y)$.

- How many strictly increasing functions $f = \{1, \ldots, k\} \to \{1, \ldots, n\}$ exist?
  Any such function is uniquely determined by its range which is a subset of $\{1, 2, \ldots, n\}$ of size $k$. There are $\binom{n}{k}$ such subsets, and hence $\binom{n}{k}$ strictly increasing
- How many increasing functions $f = \{1, \ldots, k\} \to \{1, \ldots, n\}$ exist?
  Define a bijection

  $$\{f : \{1, \ldots, k\} \to \{1, \ldots, n\} | \text{ increas.}\} \text{ to } \{g : \{1, \ldots, k\} \to \{1, \ldots, n+k-1\} | \text{strict. increas.}\}$$

  $\forall f$ increasing, define $g(i) = f(i) + i - 1$. Then $g$ is strictly increasing and takes values in $\{1, \ldots, n+k-1\}$
  So total number of increasing functions $f = \{1, \ldots, k\} \to \{1, \ldots, n\}$ is $\binom{n+k-1}{k}$.

## 1.3 Stirling's Formula

**Notation.** Let $(a_n)$ and $(b_n)$ be 2 sequences. We write:

$$a_n \sim b_n \text{ if } \frac{a_n}{b_n} \to 1 \text{ as } n \to \infty$$

**Theorem** (Stirling).

$$n! \sim n^n \sqrt{2\pi n} e^{-n} \text{ as } n \to \infty$$

**Note.** Weaker examinable statement proved below

**Proof** (non-examinable). $\forall f$ : twice differentiable, $\forall a < b$:

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

(can prove by doing IBP twice)
Take $f(x) = \log x,\ a = k$ and $b = k+1$

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

Take the sum for $k = 1, \dots, n-1$ of the equality above to get:

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

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

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

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

Note that $a_k \le \frac{1}{2}\int_0^1 \frac{x(1-x)}{k^2}\,\mathrm{d}x = \frac{1}{12k^2}$
So $\sum a_k < \infty$. We set $A = \exp\left(1 - \sum_{k=1}^{n-1} a_k\right)$
Then:

$$n! = n^n \cdot e^{-n}\sqrt{A} \cdot \underbrace{\exp\left(\sum_{k=1}^{n-1} a_k\right)}_{\to 0}$$

So we proved that

$$\frac{n!}{n^n e^{-n}\sqrt{n}} \to A \text{ as } n \to \infty$$

Which means $n! \sim n^n e^{-n}\sqrt{n}A$ as $n \to \infty$.
To finish the proof, need to show $A = \sqrt{2\pi}$

**Theorem** (Continued)**.**

**Claim.** $A = \sqrt{2\pi}$ knowing that $n! \sim n^n w^{-n}\sqrt{n} \cdot A$ as $n \to \infty$

**Proof.**

$$2^{-2n} \cdot \binom{2n}{n} = 2^{-2n} \cdot \frac{(2n)!}{n! \cdot n!} \sim 2^{-2n} \frac{(2n)^2 n \cdot \sqrt{2n} \cdot A \cdot e^{-2n}}{n^n 2^{-n}\sqrt{n} \cdot A \cdot n^n \cdot e^{-n} \cdot \sqrt{n} \cdot A} = \frac{\sqrt{2}}{A\sqrt{n}}$$

Using a different method we will prove that

$$2^{-2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}$$

Which will then force $A = \sqrt{2\pi}$
Consier $I_n = \int_0^{\frac{pi}{2}} (\cos\theta)^n \, \mathrm{d}\theta, \ n \geq 0.$
So $I_0 = \frac{\pi}{2}$ and $I_1 = 1$.
By integration by parts

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

So

$$I_{2n} = \frac{2n-1}{2n} I_{2n-2} = \frac{(2n-1)\cdot(2n-3)\ldots 3\cdot 1}{2n\cdot(2n-2)\ldots 2} \underbrace{I_0}_{=\frac{\pi}{2}} = \frac{(2n)!}{2^2 n \cdot n! \cdot n!} \cdot \frac{\pi}{2}$$

Achieved by mutiplying numerator and denominator by the even terms $2n(2n-2)\ldots 2$.
So $I_{2n} = 2^{-2n} \cdot \binom{2n}{n}\frac{\pi}{2}$
In the same way we get

$$I_{2n+1} = \frac{2n\ldots 4\cdot 2}{(2n+1)\ldots 3\cdot 1} = \frac{1}{2n+1}\left(2^{-2n}\cdot\binom{2n}{n}\right)^{-1}$$

From $I_n = \frac{n-1}{n} I_{n-2}$ we get:
$\frac{I_n}{I_{n-2}} \to 1$ as $n \to \infty$
Want $\frac{I_{2n}}{I_{2n+1}} \to 1$ as $n \to \infty$
Recall $I_n = \int_0^{\frac{\pi}{2}} (\cos\theta)^n \, \mathrm{d}\theta$. We see that $I_n$ is decreasing as a function of $n$.
Therefore

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

And also

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

So $\frac{I_{2n}}{I_{2n+1}} \to 1$ as $n \to \infty$, which means

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

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

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

**Claim.** Weaker statement of Stirling:

$$\log(n!) \sim n \log n \text{ as } n \to \infty$$

**Proof.** Define $l_n = \log(n!) = \log 2 + \dots \log n$
For $x \in \mathbb{R}$, we write $\lfloor x \rfloor$: integer part of $x$.

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

Integrate from 1 to $n$

$$\int_1^n \log\lfloor x \rfloor \, \mathrm{d}x \le \int_1^n \log x \, \mathrm{d}x \le \int_1^n \log\lfloor x+1 \rfloor \, \mathrm{d}x$$

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

For all $n$ we get

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

Divide through by $n \log n$ to get:

$$\frac{l_n}{n \log n} \to 1 \text{ as } n \to \infty \square$$

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

## 1.4 Properties of Probability Measures

**Note.** Have $(\Omega, \mathcal{F}, \mathbb{P})$ probability space.
$\mathbb{P} : \mathcal{F} \to [0,1]$ is a probability measure if:

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

(ii) $(A_n)_{n \ge 1}$ disjoint
$\mathbb{P}(\bigcup A_n) = \sum \mathbb{P}(A_n)$ countable additivity

### 1.4.1 Countable subadditivity

**Claim.** Let $(A_n)_{n\geq 1}$ be a sequence of events in $\mathcal{F}$ $(A_n \in \mathcal{F}\,\forall n)$
Then

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

**Proof.** Define $B_1 = A_1$ and $B_n = A_n \backslash (A_1 \cup \cdots \cup A_{n-1})\,\forall n \geq 2$.
Then $(B_n)_{n\geq 1}$ is a disjoint sequence of events in $\mathcal{F}$ and $\bigcup\limits_{n\geq 1} B_2 = \bigcup\limits_{n\geq 1} A_n$.

So $\mathbb{P}(\bigcup A_n) = \mathbb{P}(\bigcup B_n)$
By countable additivity for $(B_n)$:

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

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

$$\mathbb{P}\left(\bigcup A_n\right) = \mathbb{P}\left(\bigcup B_n\right) = \sum \mathbb{P}\left(B_n\right) \leq \sum_{n\geq 1} \mathbb{P}\left(A_n\right) \square$$

### 1.4.2 Continuity of Probability Measures

Let $(A_n)_{n\geq 1}$ be an increasing sequence on $\mathcal{F}$, i.e. $\forall n\ A_n \in \mathcal{F}$ and $A_n \subseteq A_{n+1}$. Then $\mathbb{P}(A_n) \leq \mathbb{P}(A_{n+1})$.
So $\mathbb{P}(A_n)$ converges as $n \to \infty$.

**Claim.**

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\bigcup_n A_n)$$

**Proof.** Set $B_1 = A_1$ and $\forall n \geq 2\ B_n = A_n \backslash (A_1 \cup \cdots \cup A_{n-1})$
Then $\bigcup\limits_{k=1}^{n} B_k = A_n$ and $\bigcup\limits_{k=1}^{n} B_k = \bigcup\limits_{k=1}^{n} A_k$
So $\mathbb{P}(A_n) = \mathbb{P}(\bigcup\limits_{k=1}^{n} B_k) = \sum\limits_{k=1}^{n} \mathbb{P}(B_k) \underset{n\to\infty}{\to} \sum\limits_{k=1}^{\infty} \mathbb{P}(B_k)$ as $B_k$ disjoint sequence of events.
Remains to prove that:

$$\sum_{k=1}^{\infty} \mathbb{P}(B_k) = \mathbb{P}\left(\bigcup A_n\right)$$

Since $\bigcup\limits_{k=1}^{\infty} B_k = \bigcup\limits_{k=1}^{\infty} A_k$, we get $\mathbb{P}(\bigcup A_n) = \mathbb{P}(\bigcup B_n) = \sum\limits_{n} \mathbb{P}(B_n) \square$

**Note.** Similarly, if $(A_n)$ is a decreasing sequence in $\mathcal{F}$, i.e. $\forall n\ A_n \in \mathcal{F}$ and $A_{n+1} \subseteq A_n$, then

$$\mathbb{P}(A_n) \to \mathbb{P}\left(\bigcap_n A_n\right) \text{ as } n \to \infty$$

### 1.5 Inclusion-Exclusion Formula

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

**Claim.** Let $A_1, \dots, A_n \in \mathcal{F}$. then

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

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

**Proof.** By induction. For $n = 2$ it holds.
Assume it holds for $n-1$ events. We will prove it for $n$ events.

$$\mathbb{P}(A_1 \cup \dots \cup A_n) = \mathbb{P}((A_1 \cup \dots A_{n-1}) \cup A_n) = \mathbb{P}(A_1 \cup \dots A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}((A_1 \cup \dots A_{n-1}) \cap A_n) \ (*)$$

Notice

$$\mathbb{P}((A_1 \cup \dots A_{n-1}) \cap A_n) = \mathbb{P}((A_1 \cap A_n) \cup \dots \cup (A_{n-1} \cap A_n))$$

Set $B_i = A_i \cap A_n$. By the inductive hypothesis,

$$\mathbb{P}(A_1 \cup \dots \cup A_{n-1}) = \sum_{k=1}^{n-1} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \dots < i_k \le n-1} \mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k})$$

$$\mathbb{P}(B_1 \cup \dots \cup B_{n-1}) = \sum_{k=1}^{n-1} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \dots < i_k \le n-1} \mathbb{P}(B_{i_1} \cap \dots \cap B_{i_k})$$

Plugging these two into back into $(*)$ gives the claim. $\square$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ with $|\Omega| < \infty$ and $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ $\forall A \in \mathcal{F}$.
Let $A_1, \dots, A_n \in \mathcal{F}$. Then

$$|A_1 \cup \dots \cup A_{n-1}| = \sum_{k=1}^{n-1} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \dots < i_k \le n-1} |A_{i_1} \cap \dots \cap A_{i_k}|$$

### 1.5.1 Bonferroni Inequalities

**Claim.** Truncating sum in the inclusion-exclusion formula at the $r$- th term gives an overestimate if $r$ is odd and an underestimate if $r$ is even, i.e.

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

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

**Proof.** By induction. For $n = 2$ $\mathbb{P}(A\cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
Assume the claim holds for $n-1$ events. Will prove for $n$.
Suppose $r$ is odd. Then

$$\mathbb{P}(A_1\cup\cdots\cup A_n) = \mathbb{P}(A_1\cup\cdots\cup A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}(B_1\cup\cdots\cup B_{n-1}), \text{ where } B_i = A_i\cap A_n \ (*)$$

Since $r$ is odd, apply the inductive hypothesis to $\mathbb{P}(A_1\cup\cdots\cup A_n)$ to get:

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

Since $r-1$ is even, apply the inductive hypothesis to $\mathbb{P}(B_1\cup\cdots\cup B_{n-1})$

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} B_i\right) \geq \sum_{k=1}^{r-1}(-1)^{k+1} \sum_{1\leq i_1<i_2<\cdots<i_k\leq n-1} \mathbb{P}(B_{i_1}\cap\cdots\cap B_{i_k})$$

Substitute both bounds in $(*)$ to get an overestimate.
In exactly the same way we prove the result for $r$ even. $\square$

### 1.5.2 Counting using Inclusion-Exclusion

**Example.** Number of surjections $f : \{1, \dots, n\} \to \{1, \dots, m\}$
Let $\Omega = \{f : \{1, \dots, n\} \to \{1, \dots, m\}\}$ and $A = \{f \in \Omega : f \text{ is a surjection }\}$.
$|A| =?$ $\forall i \in \{1, \dots, m\}$ define $A_i = \{f \in \Omega : i \notin \{f(1), \dots, \dots, f(n)\}\}$
Then

$$A = A_1^C \cap A_2^C \cap \dots \cap A_m^C = (A_1 \cup \dots \cup A_m)^C$$

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

$$|A_1 \cup \dots \cup A_m| = \sum_{k=1}^{m} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \dots < i_k \le m} \underbrace{|A_{i_1} \cup \dots \cup A_{i_k}|}_{=(m-k)^n} = \sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} (m-k)^n$$

So

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

### 1.5.3 Counting Derangements

**Note.** A derangement is a permutation with no fixed points

**Example.**

$$A = \{\text{derangements}\} = \{f \in \Omega : f(i) \neq i \, \forall i - 1, \dots, n\}$$

Pick a permutation at random. What is the probability it is in $A$?
Define $A_i = \{f \in \Omega : f(i) = i\}$
Then $A = A_1^C \cap A_2^C \cap \dots \cap A_n^C = \left(\bigcup_{i=1}^{n} A_i\right)^C$
So $\mathbb{P}(A) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right)$
By inclusion-exclusion

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

So

$$\mathbb{P}(A) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \xrightarrow[n \to \infty]{} e^{-1} \approx 0.3678$$

## 1.6 Independence

**Definition.** Let $A, B \in \mathcal{F}$. They are called **independent** ($A \perp\!\!\!\perp B$) if

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

A countable collection of events $(A_n)$ is said to be **independent** if $\forall$ distinct $i_1, i_2, \ldots, i_k$ we have

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

**Remark.** Pairwise independent does not imply independent see example below

**Example.** Toss a fair coin twice

$$\Omega = \{(0,0), (0,1), (1,0), (1,1)\} \; \mathbb{P}(\omega) = \frac{1}{4} \, \forall \omega \in \Omega$$

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

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

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{(0,0)\}) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} - \mathbb{P}(A) \cdot \mathbb{P}(B) \implies A \perp\!\!\!\perp B$$

Similarly

$$\mathbb{P}(B \cap C) = \mathbb{P}(B) \cdot \mathbb{P}(C) \implies B \perp\!\!\!\perp C$$

and

$$\mathbb{P}(A \cap C) = \mathbb{P}(A) \cdot \mathbb{P}(C) \implies A \perp\!\!\!\perp C$$

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

So $A, B$ and $C$ are note independent

**Claim.** If $A$ is independent of $B$, then $A$ s also independent of $B^C$

**Proof.**

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

### 1.7 Conditional Probability

**Definition.** Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$
Let $A \in \mathcal{F}$. We define the **conditional probability** of $A$ given $B$ and write $\mathbb{P}(A|B)$ to be

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

**Note.** If $A$ and $B$ are independent, then $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(a) \cdot \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$
So in this case $\mathbb{P}(A|B) = \mathbb{P}(A)$

**Claim.** Suppose $(A_n)$ is a disjoint sequence in $\mathcal{F}$.
Then $\mathbb{P}(\bigcup A_n|B) = \sum_n \mathbb{P}(A_n|B)$ (countable additivity for conditional probability)

**Proof.**

$$\begin{aligned}
\mathbb{P}(\bigcup A_n|B) &= \frac{\mathbb{P}((\bigcup A_n) \cap B}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P} \bigcup_n (A_n \cap B)}{\mathbb{P}(B)} \\
&= \sum_n \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} \text{ countable additivity of } \mathbb{P} \\
&= \sum_n \mathbb{P}(A_n|B) \square
\end{aligned}$$

### 1.8 Law of Total Probability

**Claim.** Suppose $(B_n)_{n \in \mathbb{N}}$ is a disjoint collection in $\mathcal{F}$ and $\bigcup B_n = \Omega$ and $\forall n \, \mathbb{P}(B_n) > 0$.
Let $A \in \mathcal{F}$. Then $\mathbb{P}(A) = \sum_n \mathbb{P}(A|B_n) \cdot \mathbb{P}(B_n)$

**Proof.**

$$\begin{aligned}
\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) &= \mathbb{P}\left(A \cap \left(\bigcup_n B_n\right)\right) \\
&= \mathbb{P}\left(\bigcup_n (A \cap B_n)\right) \\
&= \sum_n \mathbb{P}(A \cap B_n) \text{ countable additivity of } \mathbb{P} \\
&= \sum_n \mathbb{P}(A|B_n) \cdot \mathbb{P}(B_n) \square
\end{aligned}$$

## 1.9 Bayes' Formula

**Claim.** $(B_n)$ disjoint events, $\bigcup B_n = \Omega,\ \mathbb{P}(B_n) > 0\, \forall n$

$$\mathbb{P}(B_n|A) = \frac{\mathbb{P}(A|B_n) \cdot \mathbb{P}(B_n)}{\sum_k \mathbb{P}(A|B_k) \cdot \mathbb{P}(B_k)}$$

**Proof.**

$$\mathbb{P}(B_n|A) = \frac{\mathbb{P}(B_n \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_n) \cdot \mathbb{P}(B_n)}{\mathbb{P}(A)}$$

and

$$P(A) = \sum_k \mathbb{P}(A|B_k) \cdot \mathbb{P}(B_k) \text{ using law of total prob.}$$

**Note.** This formula is the basis of Bayesian statistics.
We know the probabilities of the events $(B_k)$ and we have a model which gives us $\mathbb{P}(A|B_n)$. Bayes' formula tells us how to calculate the posterior probabilities of $B_n$ given that the event $A$ occurs.

**Equation.** Let $(B_n)$ be a partition of $\Omega$, i.e. $(B_n)$ are disjoint and $\cup B_n = \Omega$

$$\forall A \in \mathcal{F}\ \mathbb{P}(B_n|A) = \frac{\mathbb{P}(A|B_n) \cdot \mathbb{P}(B_n)}{\sum_k \mathbb{P}(A|B_k)\mathbb{P}(B_k)}$$

Baye's formula

**Example** (False positives for a rare condition)**.** Suppose that condition $A$ affects 0.1% of the population. We have medical test which is posititve for 98% of the affected population and 1% of those unaffected by the disease. Pick an individual at random. What is the probability they suffer from $A$ given they tested positive?
Define

$$A = \{\text{individual suffers from } A\}$$
$$P = \{\text{individual tested positive}\}$$

Want $\mathbb{P}(A|P)$

$$\mathbb{P}(A) = 0.001,\ \mathbb{P}(P|A) = 0.98,\ \mathbb{P}(P|A^C) = 0.01$$

$$\mathbb{P}(A|P) = \frac{\mathbb{P}(P|A) \cdot \mathbb{P}(A)}{\mathbb{P}(P|A) \cdot \mathbb{P}(A) + \mathbb{P}(P|A^C) \cdot \mathbb{P}(A^C)} = \frac{0.98 \times 0.001}{0.98 * 0.001 + 0.01 * 0.999} = 0.089 \cdots \simeq 0.09$$

So $\mathbb{P}(A|P) \simeq 0.9$
The reason why this is low is because $\mathbb{P}(P|A^C) >> \mathbb{P}(A)$

$$\mathbb{P}(A|P) = \frac{1}{1 + \frac{\mathbb{P}(P|A^C) \cdot \mathbb{P}(A^C)}{\mathbb{P}(P|A) \cdot \mathbb{P}(A)}}$$

But $\mathbb{P}(A^C) \simeq 1$ and $\mathbb{P}(P|A) \simeq 1$ so can approximate

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

Suppose there is a population of 1000 people and about 1 suffers from the disease. In the 999 not suffering from the disease about 10 will test positive. So in total there will be about 11 people testing positive. Pick an individual at random among these 11 people, then the prob they have the disease will be $\frac{1}{11}$

**Example** (Extra knowledge gives surprising results)**.** 3 Statements:

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

(b) I have 2 children and the eldest one is a boy

(c) I have 2 children one of whom is a boy born on a Thursday

$$\mathbb{P}(\text{I have 2 boys}|a) = ?$$

$$\mathbb{P}(\text{I have 2 boys}|b) = ?$$

$$\mathbb{P}(\text{I have 2 boys}|c) = ?$$

Define

$$BG = \{\text{elder} = \text{boy, younger} = \text{girl}\}$$

$$GB = \{\text{elder} = \text{girl, younger} = \text{boy}\}$$

$BB,\ GG$ defined similarly

(a)

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

(b)

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

(c)

$$GT = \{\text{elder} = \text{girl, younger} = \text{boy born on a Thursday}$$

(d)

$$TN = \{\text{elder} = \text{boy born on a Thursday, younger} = \text{boy not born on a Thursday}$$

$TT,\ TG, NT$ defined similarly

$$\mathbb{P}(TT \cup TN \cup NT|GT \cup TG \cup TT \cup TN \cup NT) = \frac{\mathbb{P}(TT \cup TN \cup NT)}{\mathbb{P}(TT \cup TN \cup NT \cup GT \cup TG)}$$

$$\mathbb{P}(TT \cup TN \cup NT) = \frac{1}{2} \cdot \frac{1}{7} \cdot \frac{1}{2} \cdot \frac{1}{7} + \frac{1}{2} \cdot \frac{1}{7} \cdot \frac{1}{2} \cdot \frac{6}{7} + \frac{1}{2} \cdot \frac{6}{7} \cdot \frac{1}{2} \cdot \frac{1}{7} = \frac{13}{49 \times 4}$$

$$\mathbb{P}(TT \cup TN \cup NT \cup GT \cup TG) = \frac{13}{49 \times 4} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{7} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{7} + \frac{1}{2} \cdot \frac{1}{7} \cdot \frac{1}{2}$$

$$\mathbb{P}(TT \cup TN \cup NT|GT \cup TG \cup TT \cup TN \cup NT) = \frac{13}{27}$$

## 1.10 Simpson's Paradox

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

| Men Only | Admitted | Rejected | % Admitted |
|---|---|---|---|
| State | 15 | 22 | 41% |
| Independent | 5 | 8 | 38% |

| Women Only | Admitted | Rejected | % Admitted |
|---|---|---|---|
| State | 10 | 3 | 77% |
| Independent | 23 | 14 | 62% |

**Remark.** This phenomenon is called confounding in statistics. It arises when we aggregate data from disparate populations.

Let

$$\begin{aligned} A &= \{\text{individual is admitted}\} \\ B &= \{\text{individual is a man}\} \\ B^c &= \{\text{individual is a woman}\} \\ C &= \{\text{individual comes from a state school}\} \\ C^C &= \{\text{individual comes from an independent school}\} \end{aligned}$$

Here we see that

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

However we see that

$$\mathbb{P}(A|C^C) > \mathbb{P}(A|C)$$

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

Assume further that $\mathbb{P}(B|C) = \mathbb{P}(B|C^C)$. Then,

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

So under this extra assumption $\left(\mathbb{P}(B|C) = \mathbb{P}(B|C^C)\right)$ which is not valid here, we would get that indeed $\mathbb{P}(A|C) > \mathbb{P}(A|C^C)$

# 2 Discrete Random Variables

## 2.1 Definitions and Examples

**Definition** (Discrete Probability Distribution)**.**

$$(\Omega, \mathcal{F}, \mathbb{P}) \text{ }\Omega\text{finite or countable}$$

$$\Omega = \{\omega_1, \omega_2, \dots, \}$$

$$\mathcal{F} = \{\text{all subsets of } \Omega\}$$

If we know $\mathbb{P}(\{\omega_i\}) \; \forall i$, then this determines $\mathbb{P}$.
Indeed, let $A \subseteq \Omega$ then

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

We write $p_i = \mathbb{P}(\{\omega_i\})$ and we call it a **discrete probability distribution**

**Note.** Properties:

- $p_i \geq 0 \; \forall i$
- $\sum_i p_i = 1$

**Example** (Bernoulli Distribution)**.** Model the outcome of the toss of a coin.

$$\Omega = \{0, 1\} \; p_1 = \mathbb{P}(\{1\}) = p \text{ and } p_0 = \mathbb{P}(\{0\}) = 1 - p$$

$$\mathbb{P}(\text{we see a } H) = p, \; \mathbb{P}(\text{we see a } T) = 1 - p$$

**Example** (Binomial distribution)**.**

$$B(N, p), \; N \in \mathbb{Z}^+, p \in [0, 1]$$

Toss a $p$-coin (prob of $H$ is $p$) $N$ times independently.

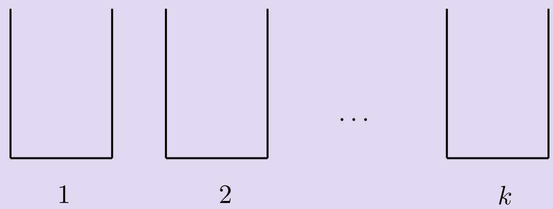
$$\mathbb{P}(\text{we see } k \text{ heads}) = \binom{N}{k} p^k (1-p)^{n-k}$$

$$\Omega = \{0, 1, \dots, N\} \; p_k = \binom{N}{k} \cdot p^k \cdot (1-p)^k$$

$$\sum_{k=0}^{N} p_k = 1$$

**Example** (Multinomial Distribution)**.**

$$M(N, p_1, \ldots, p_k),\ N \in \mathbb{Z}^+,\ p_1, \ldots, p_k \geq 0 \text{ and } \sum_{i=1}^{k} p_i = 1$$




$k$ boxes and $N$ balls

$$\mathbb{P}(\text{pick box } i) = p_i$$

Throw the balls independently.

$$\Omega = \{(n_1, \ldots, n_k) \in N^k : \sum_{i=1}^{k} n_i = N\}$$

The set of ordered partitions of $N$.

$$\mathbb{P}(n_1 \text{ balls fall in box } 1, \ldots, n_k \text{ fell in box } k) = \binom{N}{n_1, \ldots, n_k} \cdot p_1^{n_1} \cdot p_2^{n_1} \ldots p_k^{n_k} \quad \sum n_i = N$$

**Example** (Geometric Distribution)**.** Toss a $p$-coin until the first $H$ appears.

$$\Omega = \{1, 2, \ldots\}$$

$$p_k = \mathbb{P}(\text{we tossed } k \text{ times until first } H) = (1-p)^{k-1} p$$

$$\sum_{k=1}^{\infty} p_k = 1$$

$$\Omega = \{0, 1, \ldots\}\ \mathbb{P}(k \text{ tails before first } H = (1-p)^k \cdot p = p'_k$$

$$\sum_{k=1}^{\infty} p'_k = 1$$

**Example** (Poisson Distribution)**.** This is used to model the number of occurences of an event in a given interval of time. For instance, the number of customers that enter a shop in a day.

$$\Omega = \{1, 2, \dots\} \ \lambda > 0$$

$$p_k = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, \ \forall k \in \Omega$$

We call this the Poisson distribution with parameter $\lambda$.

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

So indeed it is a probability distribution.
Suppose customers arive into a shop during $[0, 1]$. Discretise [0,1] , i.e. subdivide $[0, 1]$ into $N$ intervals $\left[\frac{i-1}{N}, \frac{i}{N}\right], \ i = 1, 2, \dots, N$
In each interval, a customer arrives with probability $p$ (independently of other intervals and with probability (w.p.) $1 - p$ nobody arrives.

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

Take $p = \frac{\lambda}{N}, \ \lambda > 0$:

$$\binom{N}{k} \cdot p^k \cdot (1-p)^{N-k)} = \frac{N!}{k!(N-k)!} \left(\frac{\lambda}{N}\right)^k \cdot \left(1 - \frac{\lambda}{N}\right)^{N-k} = \frac{\lambda^k}{k!} \frac{N!}{N^k (N-k)!} \left(1 - \frac{1}{N}\right)^{N-k}$$

Keep $k$ fixed and send $N \to \infty$
So:

$$\mathbb{P}(k \text{ customers arrived}) \to e^{-\lambda} \cdot \frac{\lambda^k}{k!} \text{ as } N \to \infty$$

This is exactly the Poisson distribution. So we showed that the $B(N, p)$ with $p = \frac{1}{N}$ converges to the Poisson with parameter $\lambda$.

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

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

**Notation.** We will use the shorthand notation: suppose $A \subseteq \mathbb{R}$

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

**Definition.** Given $A \in \mathcal{F}$, define the **indicator** of $A$ to be

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

Because $A \in \mathcal{F}, \ 1_A$ is a random variable.

**Definition.** Suppose $X$ is a random variable. Define the **probability distribution function** of $X$ to be

$$F_X(x) = \mathbb{P}(X \le x),\ \mathcal{F}_X : \mathbb{R} \to [0,1]$$

**Definition.** $(X_1, \dots, X_n)$ is called a **random variable in $\mathbb{R}^n$** if

$$(X_1, \dots, X_n) : \Omega \to \mathbb{R}^n$$

and $\forall x_1, \dots, x_n \in \mathbb{R}$ we have

$$\{X_1 \le x_1, \dots, X_n \le x_n\} \in \mathcal{F}$$

i.e.

$$\{\omega : X_1(\omega) \le x_1, \dots, X_n(\omega) \le x)n\}$$

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

$$\{X_1 \le x_1, \dots, X_n \le x_n\} = \{X_1 \underset{\in\mathcal{F}}{\le} x_n\} \cap \cdots \cap \{X_n \underset{\in\mathcal{F}}{\le} x_n\} \in \mathcal{F}$$

**Definition.** A random variable $X$ is called **discrete** if it takes values in a countable set.
Suppose $X$ takes values in the countable set $S$. For every $x \in D$ we write $p_x = \mathbb{P}(X = x) = \mathbb{P}(\{\omega : X(\omega) = x\})$. We call $(p_x)_{x\in S}$ the probability mass function of $X$ (pmf) or the distribution of $X$.
If $(p_x)$ is Bernoullim then we say that $X$ is a Bernoulli r.v. or that $X$ has the Bernoulli distribution.
If $(p_x)$ is Geometric, similarly say $X$ is a geometric r.v. etc.

**Definition.** Suppose that $X_1, \dots, X_n$ are discrete r.v.s taking values in $S_1, \dots, S_n$. We say $X_1, \dots, X_n$ are **independent** if

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

**Example.** Toss a $p$-biased coin $N$ times independent.
Take
$$\Omega = \{\underset{T}{0}, \underset{H}{1}\}^N$$
$$\omega \in \Omega \; p_\omega = \prod_{k=1}^{N} p^{\omega_k}(1-p)^{1-\omega_k} \; \omega = (\omega_1, \dots, \omega_N) \in \Omega$$
Define
$$X_k(\omega) = \omega_k \; \forall k = 1, \dots, N \; \omega \in \Omega$$
Then the $X_k$ gives the outcome of the $k$-th toss and is a discrete r.v.
$$X_k : \Omega \to \{0, 1\}$$
$$\mathbb{P}(X_k = 1) = \mathbb{P}(w_k = 1) = p \text{ and } \mathbb{P}(X_k = 0) = \mathbb{P}(w_k = 0) = 1 - p$$
So $X_k$ has the Bernoulli distribution with parameter $p$

**Claim.** $X_1, \dots, X_N$ are independent r.v.s

**Proof.** Let $x_1, \dots, x_N \in \{0, 1\}$. Then
$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(\omega = (x_1, \dots, x_n))$$
$$= p_{(x_1, \dots, x_n)} \qquad = \prod_{k=2}^{N} p^{X_k} \cdot (1-p)^{1-X_k}$$
$$= \prod_{k=1}^{n} \mathbb{P}(X - k = x_k) \square$$

Define
$$S_N(\omega_= X_1(\omega) + \dots + X_N(\omega) = \# textof H \text{ in } N \text{ tosses}$$
$$S_N : \Omega \to \{0, \dots, N\}$$
And
$$\mathbb{P}(S_n = k) = \binom{N}{k} = p^k \cdot (1-p)^{N-k}$$
So $S_N$ has the Binomial distribution of parameters $N$ and $p$

## 2.2 Expectation

$(\Omega, \mathcal{F}, \mathbb{P})$. Assume $\Omega$ is finite or countable.
Let $X : \Omega \to \mathbb{R}$ be a r.v. (discrete).
We say $X$ is non-negative if $X \geq 0$.

**Definition** (Expectation of $X \geq 0$)**.**

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

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

So

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

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

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

So the **expectation** of $X$ (mean of $X$, average value) is an average of the values taken by $X$ with weights given by $\mathbb{P}(X = x)$.
So

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

**Example.** Suppose $X$ has the Binomial distribution with $N$ and $p$.

$$(X \sim \mathrm{Bin}(N, p))$$

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

So

$$\mathbb{E}[X] = \sum_{k=1}^{N} k \cdot \mathbb{P}(X = k) = \sum_{k=1}^{N} k \cdot \binom{N}{k} \cdot p^k \cdot (1-p)^{N-k}$$

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

**Example.** Ket $X$ be a Poisson r.v. of parameter $\lambda > 0$, i.e.

$$\mathbb{P}(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!},\ k = 0, 1, \dots\ (X \sim \text{Poi}(\lambda))$$

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

**Definition.** Let $X$ be a general r.v. (discrete). We define $X_+ = \max(X, 0)$ and $X_- = \max(-X, 0)$. Then

$$X = X_+ - X_-$$

$$|X| = X_+ + X_-$$

We can define $\mathbb{E}[X_+]$ and $E[X_i]$ since, they are both non-negative.
If at least one of $\mathbb{E}[X_+]$ or $\mathbb{E}[X_-]$ is finite, then we define

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

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

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

#### 2.2.1 Properties of Expectation

(i) If $X \geq 0$, then $\mathbb{E}[X] \geq 0$
(ii) If $X \geq 0$ and $\mathbb{E}[X] = 0$, then $\mathbb{P}(X = 0) = 1$
(iii) If $x \in \mathbb{R}$, then $\mathbb{E}[cX] = c\mathbb{E}[X]$ and $\mathbb{E}[c + X] = x + \mathbb{E}[X]$
(iv) If $X$ and $Y$ are 2 r.v.s, then ($X$ and $Y$ are both integrable)

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

(v) Let $x_1, \dots, c_n \in \mathbb{R}$ and $X_1, \dots, X_n$ r.v.s Then (all integrable)

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

(vi) If $X = 1(A)$ with $A \in \mathcal{F}$, then $\mathbb{E}[X] = \mathbb{P}(A)$

**Claim.** Suppose $X_1, X_2, \ldots$ are non-negative radom variables. Then

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

**Proof.** ($\Omega$ countable)

$$\mathbb{E}\left[\sum_n X_n\right] = \sum_w \sum_n X_n(\omega)\mathbb{P}(\{\omega\}) = \sum_n \sum_w X_n(\omega)\mathbb{P}(\{\omega\}) = \sum_n \mathbb{E}[X_n]$$

**Claim.** If $g : \mathbb{R} \to \mathbb{R}$, then define $g(X)$ to be the random variable $g(X)(\omega) = g(X(\omega))$
Then $\mathbb{E}[g(X)] = \sum_{x\in\Omega_X} g(x) \cdot \mathbb{P}(X = x)$

**Proof.** Set $Y = g(X)$. Then

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

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

So

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

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

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

**Proof.** We can write since $X$ takes $\geq 0$ integer values

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

Taking $\mathbb{E}$ in (*) and using that $\mathbb{E}[1(A)] = \mathbb{P}(A)$ and countable additivity for $(1(X \geq k))_k$ gives the statement. $\square$

## 2.3 Another proof of the inclusion-exclusion formula

### 2.3.1 Properties of Indicator Random Variables

- $1(A^C) = 1 - 1(A)$
- $1(A \cap B) = 1(A) \cdot 1(B)$
- $1(A \cup B) = 1 - (1 - 1(A))(1 - 1(B))$

More generally

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

Taking $\mathbb{E}$ of both sides we get

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

## 2.4 Terminology

**Definition.** Let $X$ be a r.v. and $r \in \mathbb{N}$. We call $\mathbb{E}[X^r]$ as long as it is well-defined) the **r-th moment** of $X$

**Definition.** The **variance** of $X$ denoted $\mathrm{Var}(X)$ is defined to be

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

The variance is a measure of how concentrated $X$ is around its expectation. The smaller the variance, the more concentrated $X$ is aroudn $\mathbb{E}[X]$.
We call $\sqrt{\mathrm{Var}(X)}$ the standard deviation of $X$

Properties:

- $\mathrm{Var}(X) \geq 0$ and if $\mathrm{Var}(X) = 0$, then

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

- $x \in \mathbb{R}$, then $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$ and $\mathrm{Var}(X + c) = \mathrm{Var}(X)$
- $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

**Proof.**

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

- $\mathrm{Var}(X) = \min_{c \in \mathbb{R}} \mathbb{E}[(X - c)^2]$ and this min is achieved for $c = \mathbb{E}[X]$

**Proof.** Call $f(c) = \mathbb{E}[(X - c)^2] = \mathbb{E}[X^2] - 2c\mathbb{E}[X] + c^2$
Minimise $f$ to get $\min f(c) = f(\mathbb{E}[X]) = \mathrm{Var}(X)\square$

**Example.**

(i) $X \sim \text{Bin}(n, p),\ \mathbb{E}[X] = np$

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

So

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

(ii) $X \sim \text{Poi}(\lambda),\ \lambda > 0,\ \mathbb{E}[X] = \lambda$

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

$$\mathbb{E}[X(X-1)] = \sum_{k=2}^{\infty} k \cdot (k-1) \cdot e^{-\lambda} \cdot \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \cdot \lambda^2 = \lambda^2$$

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

**Definition.** Let $X$ and $Y$ be 2 random variables. Their **covariance** is defined

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

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

Properties

(i)
$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

(ii)
$$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

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

**Proof.** Expand $(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])$ and use properties of expectation

(iv) Let $x \in \mathbb{R}$. Then
$$\mathrm{Cov}(cX,Y) = c\mathrm{Cov}(X,Y)$$
and
$$\mathrm{Cov}(c+X),Y) = \mathrm{Cov}(X,Y)$$

(v)
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

**Proof.**
$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y) - \mathbb{E}[X] - \mathbb{E}[Y])^2] \\ &= \mathbb{E}[((X-\mathbb{E}[X]) + (Y - \mathbb{E}[Y]))^2] \\ &= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] + 2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)\end{aligned}$$

(vi) For all $c \in \mathbb{R}$, $\mathrm{Cov}(c,X) = 0$

(vii) $X, Y, Z$ are random variables, then
$$\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$$

More generally, for $c_1, c_2, \dots, c_n, d_1, \dots, c_n \in \mathbb{R}$ and $X_1, \dots, X_n$ and $Y_1, \dots, Y_N$ r.v's
$$\mathrm{Cov}\left(\sum_{i=1}^n c_i X_i, \sum_{i=1}^n d_i Y_i\right) = \sum_{i=1}^n \sum_{j=1}^n c_i d_j \mathrm{Cov}(X_i, Y_j)$$

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

**Remark.** Recall that $X$ and $Y$ are indep, if for all $x$ and $y$
$$\mathbb{P}(X=x, Y=y) = \mathbb{P}(X=x)\cdot\mathbb{P}(Y=y)$$

**Claim.** Let $X$ and $Y$ be 2 indep. r.v's and let

$$f, g : \mathbb{R} \to \mathbb{R}$$

Then

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

**Proof.**

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

**Equation.** Suppose that $X$ and $Y$ are independent. Then

$$\mathrm{Cov}(X, Y) = 0, \text{ since } \mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) = 0$$

So if $X$ and $Y$ are independent, then

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

**Warning.**

$$\mathrm{Cov}(X, Y) = 0 \not\Longrightarrow \text{ independence}$$

**Example.** Let $X_1, X_2, X_3$ be indep. $\mathrm{Ber}\left(\frac{1}{2}\right)$
Define

$$Y_1 = 2X_1 - 1, \ Y_2 = 2X_2 - 2$$

$$Z_2 = X_3 Y_1, \ Z_2 = X_3 Y_2$$

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

and

$$\mathrm{Cov}(Z_1, Z_2) = \mathbb{E}[Z_1, Z_2] = \mathbb{E}[X_3^2 Y_1 Y_2] \underset{\text{indep.}}{=} 0$$

We will show that $Z_1$ and $Z_2$ are not indep.
Indeed,

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

but

$$\mathbb{P}(Z_1 = 0) \cdot \mathbb{P}(Z_2 = 0) = \mathbb{P}(X_3 = 0)^2 = \frac{1}{4}$$

Hence as not equal, $Z_1$ is not independent of $Z_2$

## 2.5 Inequalities

### 2.5.1 Markov's Inequality

**Claim** (Markov's Inequality)**.** Let $X \geq 0$ be a random variable. Then $\forall a > 0$,

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

**Proof.** Observe that

$$X \geq a \cdot 1(X \geq a)$$

Taking expectations we get

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

So

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

### 2.5.2 Chebyshev's Inequality

**Claim** (Chebyshev's Inequality)**.** Let $X$ be a r.v. with $\mathbb{E}[X] < \infty$. Then $\forall a > 0$

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

**Proof.**

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) = \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq a^2) \underset{\text{markov}}{\leq} \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}$$

### 2.5.3 Cauchy-Schwarz Inequality

**Claim** (Cauchy-Schwarz Inequality)**.** Let $X$ and $Y$ be 2 r.v's. Then

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

**Proof.** Suffices to prove it for $X$ and $Y$ with $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$
Also enough to prove it for $X, Y \geq 0$

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

Assume $\mathbb{E}[X^2] > 0$ and $\mathbb{E}[Y^2] > 0$, otherwise result is trivial.
Let $t \in \mathbb{R}$ and consider

$$\begin{aligned} 0 \leq (X - tY)^2 &= X^2 - 2tXY + t^2Y^2 \\ \implies \underbrace{\mathbb{E}[X^2] - 2t\mathbb{E}[XY] + t^2\mathbb{E}[Y^2]}_{f(t)} &\geq 0 \end{aligned}$$

Minimising $f$ gives that for $t_* = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$, $f$ achieves its minimum.

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

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

### 2.5.4 Cases of Equality

**Note.** Equality in C-S occurs when

$$\mathbb{E}[(X - tY)^2] = 0 \text{ for } t = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$$

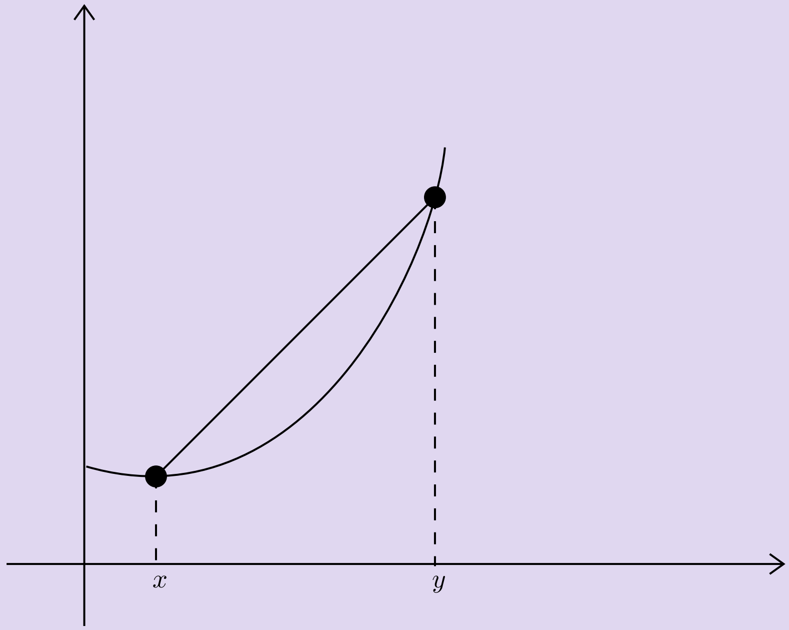
$$\mathbb{E}[(X - tY)^2] = 0 \implies \mathbb{P}(X = tY) = 1$$

### 2.5.5 Jensen's Inequality

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

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

**Example.**



**Claim** (Jensen's Inequality)**.** Let $X$ be a r.v. and let $f$ be a convex function. Then

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

**Proof.**

**Claim.** Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function then $f$ is the supremum of all the lines lying below it. In other words

$$\forall m \in \mathbb{R}\ \exists a, b \in \mathbb{R} \text{ s.t. } f(m) = am + b \text{ and } f(x) \geq ax + b\ \forall x$$

So there exists $a \in \mathbb{R}$

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

s.t.

$$\frac{f(m) - f(x)}{m - x} \leq a \leq \frac{f(y) - f(m)}{y - m} \text{ for all } x < m < y$$

(take tangent)

**Proof.** Let $m \in \mathbb{R}$. Let $x < m < y$. Then $m = tx + (1-t)y$ for some $t \in [0, 1]$. By convexity

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

$$m = tf(m) + (1-t)f(m)$$

So

$$t(f(m) - f(x)) \leq (1-t)(f(y) - f(m)) \implies \frac{f(m) - f(x)}{m - x} \leq \frac{f(y) - f(m)}{y - m}$$

So there exists $a \in \mathbb{R}$

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

s.t.

$$\frac{f(m) - f(x)}{m - x} \leq a \leq \frac{f(y) - f(m)}{y - m} \text{ for all } x < m < y$$

Rearranging this inequality we get

$$f(x) \geq a(x - m) + f(m) \text{ for all } x$$

Set $m = \mathbb{E}[X]$. Then from the claim, we get $\exists a, b \in \mathbb{R}$ s.t $f(\mathbb{E}[X]) = a\mathbb{E}[X] + b$ $(*)$ and $\forall x$ we have $f(x) \geq ax + b$ Apply this last inequality to $X$ to get

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

and taking $\mathbb{E}$

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

**Note.** A rule to remember the direction:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \geq 0$$

implies

$$\mathbb{E}(X^2) \geq (\mathbb{E}[X])^2 \square$$

### 2.5.6 Cases of Equality

Let $X$ be a r.v. and let $f$ be a convex function satisfying if $m = \mathbb{E}[X]$, then $\exists a, b \in \mathbb{R}$ s.t.

$$f(m) = am + b \text{ and } f(x) > ax + b \ \forall x \neq m$$

What is the condition of $X$ in order to have equality in Jensen?
We want

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

Have:

$$\mathbb{E}[X] = m, \ f(m) = am + b \text{ and } f(x) > ax + b \ \forall x \neq m$$

Consider $f(X) \geq aX + b$. Then $f(X) - (aX + b) \geq 0$
So taking expectations

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

But

$$\mathbb{E}[aX + b] = am + b = f(m) = f(\mathbb{E}[X])$$

We assumed

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

which means that

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

But

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

so this forces $f(X) = aX + b$
By assumption $f(m) = am + b$ and $\forall x \neq m$ $f(x) > ax + b$
So this forces $X = m$ with probability 1

### 2.5.7 AM-GM Inequality

**Claim** (AM-GM Inequality)**.** Let $f$ be a convex function and let $x_1, \ldots, x_n \in \mathbb{R}$. Then

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

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

**Proof.** Define $X$ to be the r.v. taking values $\{x_1, \ldots, x_n\}$ all with equal prob

$$\mathbb{P}(X = x_i) = \frac{1}{n} \ \forall i = 1, \ldots, n$$

$$f(\mathbb{E}[X]) = f\left(\sum_{k=1}^{n} x_k \cdot \frac{1}{n}\right)$$

By Jensen's inequality, we get

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

Let $f(x) = -\log x$. This is a convex function and so

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

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

i.e. the geometric mean is $\leq$ the arithmetic mean.

## 2.6 Conditional expectation

**Note.** Recall if $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, we defined

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

**Definition.** Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$ and let $X$ be a r.v.
We define

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

### 2.6.1 Law of Total Expectation

**Claim** (Law of Total Expectation)**.** Suppose $X > 0$ and let $(\Omega_n)$ be a partition of $\Omega$ into disjoint events, i.e.

$$\Omega = \bigcup_n \Omega_n$$

Then

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

**Proof.** Write

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

Taking expectations we get

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

By countable additivity of $\mathbb{E}$
So

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

### 2.6.2 Joint Distributions

**Definition.** Let $X_1, \ldots, X_n$ be r.v.'s (discrete). Their **joint distribution** is defined to be

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)\ \forall x_1 \in \Omega_{X_1}, \ldots, x_n \in \Omega_{X_n}$$

$$\mathbb{P}(X_1 = x_1) = \mathbb{P}(\{X_1 = x_1\} \cap \bigcup_{i=2}^{n} \bigcup_{X_i} \{X_i = x_i\})) = \sum_{X_1, \ldots, X_m} \mathbb{P}(X_1 = x_1, \ldots X_n = x_n)$$

$$\mathbb{P}(X_i = x_i) = \sum_{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

We call $(\mathbb{P}(C_i = x_i))_{x_i}$ the marginal distribution of $X_i$

**Definition.** Let $X$ and $Y$ be 2 r.v.'s
The **conditional distribution** of $X$ giben $Y = y$ $(y \in \Omega_y)$ is defined to be

$$\mathbb{P}(X = x|Y = y),\ x \in \Omega_X$$

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

**Equation.**
$$\mathbb{P}(X=x)=\sum_y \mathbb{P}(X=x,Y=y)=\sum_y \mathbb{P}(X=x|Y=y)\mathbb{P}(Y=y)$$
(law of total probability)

### 2.6.3 Distribution of the sum of independent r.v.'s

**Definition.** Let $X$ and $Y$ be 2 independent r.v.'s (discrete)

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

This last sum is called the **convolution** of the distribution of $X$ and $Y$
Similarly,
$$\mathbb{P}(X+Y=z)=\sum_x \mathbb{P}(X=x)\mathbb{P}(Y=z-x)$$

**Example.** Let $X\sim\text{Poi}(\lambda)$ and $Y\sim\text{Poi}(\mu)$ independent

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

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

**Definition.** Let $X$ and $Y$ be 2 discrete r.v.'s. The **conditional expectation** of $X$ given $Y=y$ is

$$\mathbb{E}[X|Y=y]=\frac{\mathbb{E}[X\cdot 1(Y=y)]}{\mathbb{P}(Y=y)}$$

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

**Note.** We observe that for very $y \in \Omega_Y$, $\mathbb{E}[X|Y=y]$ is a function of $y$ only.
We set

$$g(y) = \mathbb{E}[X|Y=y]$$

**Definition.** We define the **conditional expectation** fo $X$ given $Y$ and write it as $\mathbb{E}[X|Y]$ for the random variable $g(Y)$
We emphasise that $\mathbb{E}[X|Y]$ is a random variable and it depends only on $Y$, because it is a function only of $Y$

**Equation.**

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

**Example.** Toss a $p$-biased coin $n$ times independently. Write

$$X_1 = 1(i - \text{th toss is a } H) \text{ for } i = 1, \dots, n$$

and

$$Y_n = X_1 + \cdots + X_n$$

Want $\mathbb{E}[X_1|Y_n] =$?
Set $g(y) = \mathbb{E}[X_1|Y_n = y]$, then $\mathbb{E}[X_1|Y_n] = g(Y_n)$
Need to find $g$
Let $y \in \{0, \dots, n\}$. Then

$$g(y) = \mathbb{E}[X_2|Y_n = y] = \mathbb{P}(X_1 = 1|Y_n = y)$$

$y = 0$:

$$\mathbb{P}(X_1 = 1|Y_n = 0) = 0$$

$y \neq 0$:

$$\mathbb{P}(X_1 = 1|Y_n = y) = \frac{\mathbb{P}(X_1 = 1, Y_n = y)}{\mathbb{P}(Y_n = y)} = \frac{\mathbb{P}(X_1 = 1, X_2 + \cdots + X_n = y - 1)}{\mathbb{P}(Y_n = y}$$

Since the $(X_i)$ are iid, we get

$$\begin{aligned}\mathbb{P}(X_1 = x_1, X_2 + \cdots + X_n = y - 1) &= \mathbb{P}(X_1 = 1) \cdot \mathbb{P}(X_2 + \cdots + X_n = y - 1) \\ &= p \cdot \binom{n-1}{y-1} \cdot p^{y-1} \cdot (1-p)^{n-y}\end{aligned}$$

$$\mathbb{P}(Y_n = y) = \binom{n}{y} \cdot p^y \cdot (1-p)^{n-y}$$

So

$$\mathbb{P}(X_1 = 1|Y_n = y) = \frac{p \cdot \binom{n-1}{y-1} \cdot p^{y-1} \cdot (1-p)^{n-y}}{\binom{n}{y} \cdot p^y \cdot (1-p)^{n-y}} = \frac{y}{n}$$

So $g(y) = \frac{y}{n}$. Therefore

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

#### 2.6.4 Properties of Conditional Expectation

**Claim.**

- $$\forall c \in \mathbb{R}\ \mathbb{E}[cX|Y = x \cdot \mathbb{E}[X|Y] \text{ and } \mathbb{E}[c|Y] = c$$
- $X_1, \ldots, X_n$ r.v.'s, then
$$\mathbb{E}\left[\sum_{i=1}^{n} X_i|Y\right] = \sum_{i=1}^{n} \mathbb{E}[X_i|Y]$$
- $$\mathbb{E}[\mathbb{E}[X|Y] = \mathbb{E}[X]$$

**Proof.** only prove third:

$$\mathbb{E}[X|Y] = \sum_y 1(Y = y)\mathbb{E}[X|Y = y]$$

By properties of expectation

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

$$\mathbb{E}[X \cdot \sum_y 1(Y = y)]$$

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

**Proof** (Another way)**.**

$$\sum_y \mathbb{E}[X|Y = y] \cdot \mathbb{P}(Y = y) = \sum_x \sum_y x \cdot \mathbb{P}(X = x|Y = y) \cdot \mathbb{P}(Y = y) = \mathbb{E}[X] = 0$$

**Claim.**

- Let $X$ and $Y$ be 2 independent r.v.'s. Then
$$\mathbb{E}[X|Y] = \mathbb{E}[X]$$

**Proof.**

$$\begin{aligned}
\mathbb{E}[X|Y] &= \sum_y 1(Y=y)\cdot \mathbb{E}[X|Y=y] \\
&= \sum_y 1(Y=y)\cdot \sum_x x\cdot \mathbb{P}(X=x|Y=y) \\
&= \underbrace{\sum_y 1(Y=y)}_{1}\underbrace{\sum_x x\cdot \mathbb{P}(X=x)}_{\mathbb{E}[X]} = \mathbb{E}[X] \ \square
\end{aligned}$$

**Claim.** Suppose $Y$ and $X$ are independent r.v.'s. Then

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

**Proof.** We have $\mathbb{E}[X|Y] = g(Y)$ i.e. $\mathbb{E}[X|Y]$ is a function only of $Y$. If $Y$ and $Z$ are indep., then $f(Y)$ is also independent of $Z$ for any function $f$. (can show directly)
So $g(Y)$ is independent of $Z$. By the a previous property, we get

$$\mathbb{E}[g(Y)|Z] = \mathbb{E}[g(Y)] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] \ \square$$

**Claim.** Suppose $h\,\mathbb{R}\to\mathbb{R}$ is a function. Then

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

**Proof.**

$$\begin{aligned}
\mathbb{E}[h(Y)\cdot X|Y=y] &= \mathbb{E}[h(y)\cdot X|Y=y] \\
&= h(y)\cdot \mathbb{E}[X|Y=y]
\end{aligned}$$

So

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

**Corollary.**

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

and

$$\mathbb{E}[X|X] = X$$

**Remark.** Recall $X_i = (i\text{-th toss is } H)$ and $Y_n = X_1 + \cdots + X_n$

$$\mathbb{E}[X_1|Y_n] = \frac{Y_n}{n}$$

By symmetry, for all $i$

$$\mathbb{E}[X_i|Y_n] = \mathbb{E}[X_1|Y_n]$$

$$\mathbb{E}[\underbrace{\sum_{i=1}^{n} X_n}_{Y_n} |Y_n] = \sum_{i=1}^{n} \mathbb{E}[X_i|Y_n] = n \cdot \mathbb{E}[X_1|Y_n]$$

$$\therefore \mathbb{E}[X_1|Y_n] = \frac{1}{n} \cdot \mathbb{E}[Y_n|Y_n] = \frac{Y_n}{n}$$

## 2.7 Random Walks

**Definition.** A **random/ stochastic process** is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$

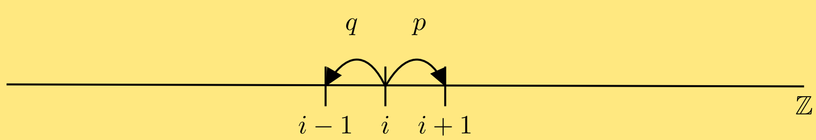
**Definition.** A **random walk** is a random process that can be expressed in the following way

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

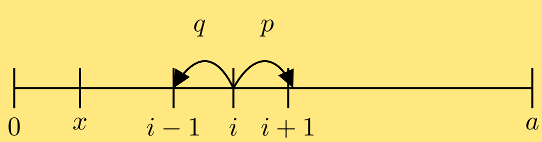
where $(Y_i)$ are independent and identically distributed (iid) r.v.'s and $x$ is a deterministic number (fixed).

**Method.** Let's focus on the *SRW* (simple random walk) on $\mathbb{Z}$ which is defined by taking

$$\mathbb{P}(Y_i = +1) = p \text{ and } \mathbb{P}(Y_i = -1) = q = 1 - p$$




We can think of $X_n$ as the fortune of a gambler who bets 1 at every step and either receives it back doubled it w.p. $p$ or loses it with prob. $q$




Suppose the gambler starts with £$x$ at time 0. What is the prob. he reaches $a$ before going bankrupt?

**Notation.** We write $\mathbb{P}_x$ for the probability measure $\mathbb{P}(\cdot|X_0 = x)$ i.e.

$$\forall A \in \mathcal{F} \ \mathbb{P}_x(A) = \mathbb{P}(A|X_0 = n)$$

**Method.** Define

$$h(x) = \mathbb{P}_x((X_n) \text{ hits } a \text{ before hitting } 0) =?$$

By the law of total probability, we have

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

$$h(x) = p \cdot h(x+1) + q \cdot h(x-1) \ 0 < x < a$$

$$h(0) = 0) \text{ while } h(a) = 1$$

- Case $p = q = \frac{1}{2}$:

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

  In this case,

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

- $p \neq q$:

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

  Try a solution of the form $\lambda^x$
  Substituting gives

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

  So the general solution will be of the form

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

  Using the boundary conditions, $h(a) = 1$ and $h(0) = 0$ yields

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

  This is the Gambler's Ruin estimate.

### 2.7.1 Expected time to absorption

Define

$$T = \min\{n \geq 0 : X_n \in \{0, a\}\}$$

i.e. $T$ is the first time $X$ hits either 0 or $a$.
Want

$$\mathbb{E}_x[T] = \tau_u =?$$

Conditioning on the first step and using the law of total expectation

$$\tau_x = p \cdot \mathbb{E}_x[T|Y_1 = +1] + q \cdot \mathbb{E}_x[T|Y_1 = -1] \; 0 < x < a$$

$$\implies \tau_x = p \cdot (1 + \mathbb{E}_{x+1}[T]) + q \cdot (1 + \mathbb{E}_{x-1}[T])$$

So

$$\tau_x = 1 + p \cdot \tau_{x+1} + q \cdot \tau_{x-1} \; 0 < x < a$$

$$\tau_0 = \tau_a = 0$$

- Case $p = \frac{1}{2}$. Try a solution of the form $Ax^2$.

  $$Ax^2 = 1 + pA(x+1)^2 + qA(x-1)^2 \implies A = -1$$

  General solution will have the form

  $$\tau_x = Ax^2 + Bx + C = -x^2 + Bx + C$$

  $$\tau_0 = \tau_a = 0 \implies \tau_x = x(a - x)$$

- Case $p \neq \frac{1}{2}$. Try $Cx$ as a solution.
  Substituting gives

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

  So the general solution will be of the form

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

  Using $\tau_0 = \tau_a = 0$,

  $$\tau_x = \frac{1}{q-p} x - \frac{q}{q-p} \frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^a - 1}$$

## 2.8 Probability Generating Functions

**Definition.** Let $X$ be a r.v. with values in $\mathbb{N}$. Let

$$p_r = \mathbb{P}(X = r),\ r \in \mathbb{N}$$

be its prob. mass function. The **pgf** of $X$ is defined to be

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

When $|z| \leq 1$, the pgf converges absolutely.
Indeed

$$|\sum_{r=0}^{\infty} p_r z^r| \leq \sum_{r=0}^{\infty} p_r \cdot |z|^r \leq \sum_{r=0}^{\infty} p_r = 1$$

So $p(z)$ is well-defined and has a radius of convergence at least 1

**Theorem.** The pgf uniquely determines the distribution of $X$

**Proof.** Suppose $(p_r)$ and $(q_r)$ are 2 prob. mass functions with

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

We will show that $p_r = q_r \ \forall r$.
Set $z = 0$. Then $p_0 = q_0$.
Suppose $p_r = q_r \ \forall r \leq n$.
RTP:

$$p_{n+1} = q_{n+1}$$

Then

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

Divide through by $z^{n+1}$ and then take the limit as $z \to 0$ gives

$$p_{n+1} = q_{n+1}$$

**Theorem.** we have

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

**Proof.** Assume first that $\mathbb{E}[X] < \infty$.
Let $0 < z < 1$. We can differentiate $p(z)$ term by term and get

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

(because $z < 1$)
Since $0 < z < 1$, we see that $p'(z)$ is an increasing function of $z$.
This implies that

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

Let $\varepsilon > 0$ and $N$ be large enough s.t.

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

Also

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

So

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

This si true for any $\varepsilon > 0$. Therefore

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

Assume $\mathbb{E}[X] = \infty$. For any $M$, take $N$ large enough s.t.

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

We know from above that

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

This is true for all $M > 0$ and hence

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

**Note.** In exactly the same way one can prove the following:

**Theorem.**

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

$$\forall k > 0 \; p^{(k)}(1-) = \lim_{z\to 1} p^{(k)}(z) = \mathbb{E}[X(X-1)\dots(X-k+1)]$$

In particular

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

Moreover

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

**Equation.** Suppose that $X_1, \dots X_n$ are independent r.v.'s with pgf's $q_1, \dots, q_n$ resp., i.e.

$$q_i = \mathbb{E}[z^{X_1}]$$

$$p(z) = \mathbb{E}[z^{X_1+\dots+X_n}] =?$$

So

$$p(z) = \mathbb{E}[z^{X_1} \cdot z^{X_2} \dots z^{X_n} = \mathbb{E}[z^{X_1}] \dots \mathbb{E}[z^{X_n}] == q_1(z) \dots q_n(z)$$

If $X_i$'s are iid, then

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

**Example.**

(i)

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

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

(ii) Let $X \sim \text{ Bin}(n, p)$ and $Y \sim \text{ Bin}(m, p)$ and $X \perp\!\!\!\perp Y$

$$\mathbb{E}[z^{X+Y}] = \mathbb{E}[z^X] \cdot \mathbb{E}[z^Y] = (pz + 1 - p)^n \cdot (pz + 1 - p)^m = (pz + 1 - p)^{n+m}$$

So

$$X + Y \sim \text{ Bin}(n + m, p)$$

(iii) Let $X \sim \text{ Geo}(p)$

$$\mathbb{E}[z^X] = \sum_{r=0}^{\infty} (1-p)^r \cdot p \cdot z^r = \frac{p}{1 - z(1-p)}$$

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

$$\mathbb{E}[z^X] = \sum_{r=0}^{\infty} z^r \cdot e^{-\lambda} \cdot \frac{\lambda^r}{r!} = e^{-\lambda} e^{-\lambda z} = e^{\lambda(z-1)}$$

Let $X \sim \text{ Poi}(\lambda)$, $Y \sim \text{ Poi}(\lambda)$ and $X \perp\!\!\!\perp Y$

$$\mathbb{E}[z^{X+Y}] = e^{\lambda(z-1)} \cdot e^{\mu(z-1)} = e^{(\lambda+\mu)(z-1)} \implies X + Y \sim \text{ Poi}(\lambda + \mu)$$

## 2.9 Sum of a Random Number of r.v.'s

**Method.** Let $X_1, X_2, \ldots$ be iid and let $N$ be an indep r.v. taking values in $\mathbb{N}$.
Define

$$S_n = X_1 + \cdots + X_n \ \forall n \geq 1$$

Then

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

means $\forall \omega \in \Omega$,

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

Let $q$ be the pgf of $N$ and $p$ the pgf of $X_1$.
Then

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

### 2.9.1 Another Proof Using Conditional Expectation

**Method.**

$$\begin{aligned} r(z) &= \mathbb{E}[z^{X_1+\cdots+X_N}] \\ &= \mathbb{E}[\mathbb{E}[z^{X_1+\cdots+X_N}|N]] \end{aligned}$$

$$\begin{aligned} \mathbb{E}[z^{X_1+\cdots+X_N}|N=n] &= \mathbb{E}[z^{X_1+\cdots+X_n}|N=n] \\ &= \left(\mathbb{E}[z^{X_1}]\right)^n \\ &= (p(z))^n \end{aligned}$$

So

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

So

$$\mathbb{E}[S_N] = \lim_{z\to 1} r'(z) = r'(1-)$$

$$r'(Z) = q'(p(z)) \cdot p'(z)$$

So

$$\begin{aligned} \mathbb{E}[S_N] &= q'(\underbrace{p(1-)}_{=1}) \cdot \underbrace{p'(1-)}_{=\mathbb{E}[X_1]} \\ &= \mathbb{E}[N] \cdot \mathbb{E}[X_1] \end{aligned}$$

$$\mathbb{E}[S_N] = \mathbb{E}[N] \cdot \mathbb{E}[X_1]$$

Similarly

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

## 2.10 Branching Processes

From Bienaguie/ Gralton-Watson, 1874.

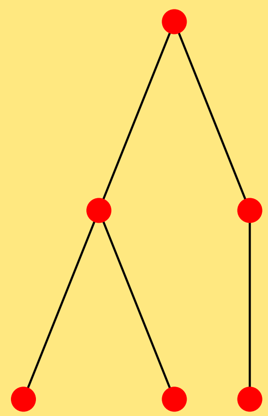
**Method.** $(X_n : n > 0)$ a random process.

$$X_n = \#\text{ of individuals in generation } n$$

$$X_0 = 1$$

The individual in generation 0 produces a random number o offspring with distribution

$$g_k = \underbrace{\mathbb{P}(X_1 = k)}_{\#\text{ children of }1^{\text{st}}\text{ individual}}, \; k = 0, 1, 2, \dots$$



Every individual in gen. 1 produces an indep. number of offspring with the same distribution.
Continue in the same way: every new indiv. produces and indep. number of offspring with the same number of offspring with the same distribution as $X_1$.
Let $Y_{k,n} : k \geq 1, n \geq 0)$ be an iid sequence with distribution $(g_k)_k$
$Y_{k,n}$ is the number of offspring of $k$-th indiv. in gen. $n$

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

**Theorem.**
$$\mathbb{E}[X_n] = (\mathbb{E}[X_1])^n \ \ \forall n \geq 1$$

**Proof.**
$$\mathbb{E}[X_{n+1}] = \mathbb{E}[\mathbb{E}[X_{n+1}|X_n]]$$

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

So
$$\mathbb{E}[X_{n+1}|X_n] = X_n \cdot \mathbb{E}[X_1]$$

So

$$\begin{aligned}\mathbb{E}[X_{n+1}] &= \mathbb{E}[X_n \cdot \mathbb{E}[X_1]] \\ &= \mathbb{E}[X_1] \cdot \mathbb{E}[X_n]\end{aligned}$$

Iterating we get
$$\mathbb{E}[X_{n+1}] = (\mathbb{E}[X_1])^{n+1} \ \ \square$$

**Theorem.** Set
$$G(z) = \mathbb{E}[z^{X_1}]$$

and
$$G_n(z) = \mathbb{E}[z^{X_n}]$$

Then

$$\begin{aligned}G_{n+1}(z) &= G(G_n(z)) \\ &= G(G(\ldots(G(z))\ldots)) \\ &= G_n(G(z))\end{aligned}$$

**Proof.**
$$G_{n+1}(z) = \mathbb{E}[z^{X_{n+1}}] = \mathbb{E}[\mathbb{E}[z^{X_{n+1}}|X_n]]$$

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

So
$$\mathbb{E}[\mathbb{E}[z^{X_{n+1}}|X_n]] = \mathbb{E}[(G(z))^{X_n}] = G_n(G(z))$$

### 2.10.1 Extinction Probability

**Method.**

$$\mathbb{P}(X_n = 0 \text{ for some } n \geq 1) = \text{ extinction prob. } = q$$

$$q_n = \mathbb{P}(X_n = 0)$$

$$A_n = \{X_n = 0\} \subseteq \{X_{n+1} = 0\} = A_{n+1}$$

Then $(A_n)$ is an increasing sequence of events.
So by continuity of prob meas.

$$\mathbb{P}(A_n) \to \mathbb{P}(\bigcup_n A_n) \text{ as } n \to \infty$$

But

$$\bigcup_n A_n = \{X_n = 0 \text{ for some } n \geq 1\}$$

Therefore we get $q_n \to q$ as $n \to \infty$

**Claim.**

$$q_{n+1} = G(q_n) \ (G(z) = \mathbb{E}[z^{X_1}]) \text{ and also } q = G(q)$$

**Proof.**

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

Since $G$ is continuous, taking the limit as $n \to \infty$ and using $q_n \to q$, we get

$$G(q) = q \ \square$$

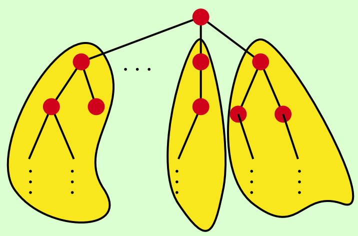
**Claim** (same as previous, different proof).

$$q_{n+1} = G(q_n) \ (G(z) = \mathbb{E}[z^{X_1}]) \text{ and also } q = G(q)$$

**Proof** (Alternative). Conditional on $X_1 = m$, we get $m$ independent branching processes. So we can write

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

where $\left(X_i^{(j)}\right)$ are iid branching processes all with the same offspring distribution.

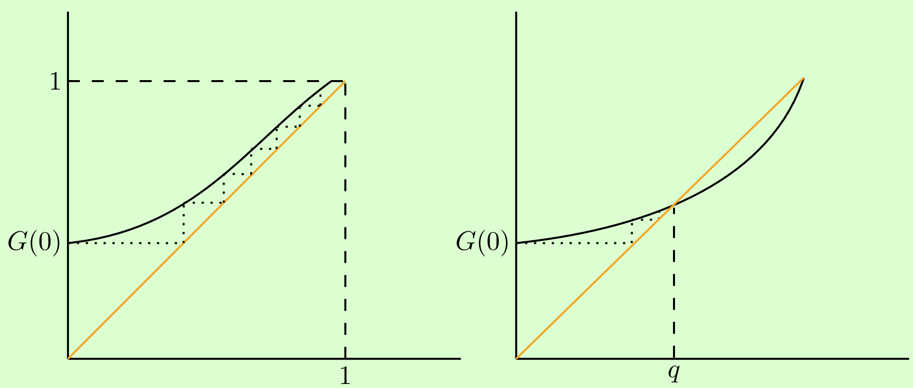


So

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

So we have proved

$$q_{n+1} = G(q_n) \text{ and } q = G(q), \ t = G(t)$$




the tangent to the graph of $G$ at 1 in $1^{\text{st}}$ has slope $< 1$.

$$\text{The slope } = G'(1-) = \mathbb{E}[X_1] < 1$$

In $2^{\text{nd}}$, the slope is

$$G'(1-) = \mathbb{E}[X_1]] > 1$$

and we see that $q < 1$

**Theorem.** Assume $\mathbb{P}(X_1 = 1) < 1$. Then the extinction probability is the minimal non-negative solution to the equation

$$t = G(t)$$

We also have

$$q < 1 \text{ iff } \mathbb{E}[X_1] > 1$$

**Proof** (of minimality)**.** Let $t$ be the smallest non-negative solution to $x = G(x)$. We will show that $q = t$.
We are going to prove by induction that

$$q_n \leq t \ \forall n$$

Then taking the limit as $n \to \infty$ will give us $q \leq t$.
Since we know that $q$ is a solution, this will imply $q = t$.

$$q_0 = \mathbb{P}(X_0 = 0) \leq t$$

Suppose $q_n \leq t$

$$q_{n+1} = G(q_n)$$

$G$ is an increasing function on $[0, 1]$, and since $q_n \leq t$, we get

$$q_{n+1} = G(q_n) \leq G(t) = t \ \square$$

**Theorem.**

**Proof** ($2^{\text{nd}}$ part)**.** Consider the function $H(z) = G(z) - z$
Assume

$$g_0 + g_1 < 1$$

so

$$\mathbb{P}(X_1 \leq 1)$$

since if not, then $\mathbb{P}(X_1 \leq 1) = 1$ which would imply that

$$\mathbb{E}[X_1] = \mathbb{P}(X_1 = 1) < 1$$

In this case we would have

$$G(z) = g_0 + g_1 z = 1 - \mathbb{E}[X_1] + \mathbb{E}[X_1]z$$

Solving $g(z) = z$ we would get $z = 1$, since $\mathbb{E}[X_1] < 1$.

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

This implies that $H'(z)$ is a strictly increasing function in $(0,1)$.
This implies that $H$ can have at most one root different from 1 in $(0,1)$, which follows from Rolle's theorem.
(If it had more, say $z_1 < z_2 < 1$, then $H'$ would be 0 in 2 points inside $(z_1, z_2)$ and $(z_2, 1)$ by Rolle's theorem. Nut this contradicts that $H'$ is strictly increasing)
$1^{\text{st}}$ case: $H$ has no other root apart from 1.
$H(1) = 0$ and $H(0) = g_0 \geq 0 \implies H(z) \geq 0 \;\forall z \in [0,1]$

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

But $H'(1-) = G'(1-) - 1$ and $H'(1-) \leq 0 \implies G'(1-) \leq 1$ and $G'(1-) = \mathbb{E}[X_1]$
So we showed that when $q = 1$, then $\mathbb{E}[X_1] \leq 1$
$2^{\text{nd}}$ case: $H$ has exactly one other root $r < 1$
$H(r) = 0$ and $H(1) = 0 \implies$ by Rolle's theorem $\exists z \in (r, 1)$ s.t. $H'(z) = 0$ but

$$H'(x) = G'(x) - 1 \implies G'(z) = 1$$

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

under the assumption $g_0 + g_1 < 1$, we know

$$G''(x) > 0 \;\forall x \in (0,1) \implies G' \text{ is strictly increasing}$$

Therefore

$$G'(1-) > G'(z) = 1 \implies \mathbb{E}[X_1] > 1$$

So we proved that if $q < 1$ then $\mathbb{E}[X_1] > 1$ □

# 3 Continuous Random Variables

## 3.1 Definitions and Properties

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

$$X : \Omega \to \mathbb{R} \text{ s.t. } \forall x \in \mathbb{R}$$

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

**Definition.** The probability distribution function is defined to be

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

Properties of $F$

(i) if $x < y$ then $F(x) \leq F(y)$

**Proof.**

$$\{X \leq x\} \subseteq \{X \leq y\}$$

(ii)

$$\forall a < b,\ a, b \in \mathbb{R}\ \mathbb{P}(a < X \leq b) = F(b) - F(a)$$

**Proof.**

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

(iii) F is a right continuous function and left limits exists always

$$F(x-) = \lim_{y \to x} F(y) \leq F(x)$$

**Proof.** NTP

$$\lim_{n \to \infty} F\left(x + \frac{1}{n}\right) = F(x)$$

Define

$$A_n = \{x < X \leq x + \frac{1}{n}\}$$

Then $(A_n)$ are decreasing events and $\bigcap_n A_n = \varnothing$

So

$$\mathbb{P}(A_n) \to 0 \text{ as } n \to \infty$$

But

$$\mathbb{P}(A_n) = \mathbb{P}(x < X \leq x + \frac{1}{n}) = F\left(x + \frac{1}{n}\right) - F(x) \to 0 \text{ as } n \to \infty$$

Left limits exist by the increasing property of $F$

(iv) $F(x-) = \mathbb{P}(X < x)$

**Proof.**

$$F(x-) = \lim_{n \to \infty} F\left(x - \frac{1}{n}\right)$$

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

Consider

$$B_n = \left\{X \leq x - \frac{1}{n}\right\}$$

then $(B_n)$ increasing and $\bigcup_n B_n = \{X < x\}$

$$\mathbb{P}(B_n) \to \mathbb{P}(X < n) \implies F(x-) = \mathbb{P}(X < x)$$

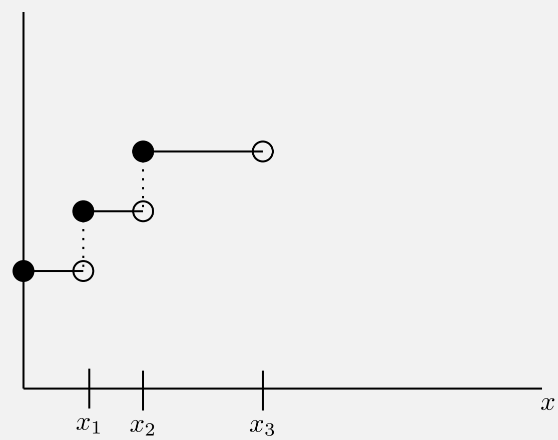
(v)

$$\lim_{x\to\infty} F(x) = 1$$

and

$$\lim_{x\to-\infty} F(x) = 0$$

For a discrete variable, $F(x) = \mathbb{P}(X \leq x)$

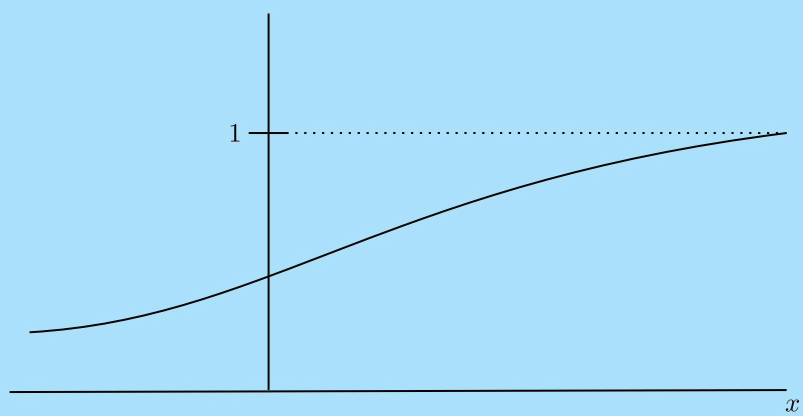



$F$ is a step function (right continuous with left limits)

**Definition.** A r.v. $X$ is called **continuous** if $F$ is a continuous function, which means that

$$F(x) = F(x-)\ \forall x \implies \mathbb{P}(X \leq x) = \mathbb{P}(X < x)\ \forall x$$

In other words, $\mathbb{P}(X = x) = 0\ \forall x \in \mathbb{R}$




**Note.** In this course, we will further restrict to the case where $F$ is not only continuous but also differentiable. (Absolutely continuous)
Set

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

We call $f$ the probability density function of $X$.

Properties of $f$:

(i) $f \geq 0$

(ii)

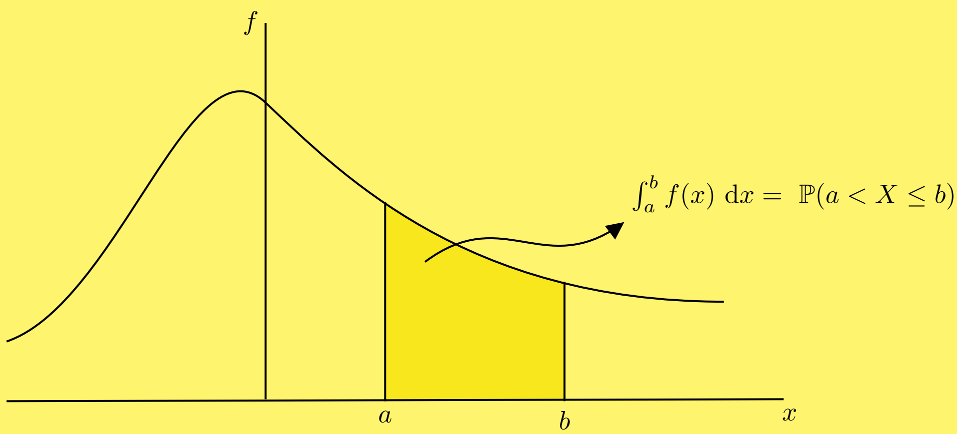
$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

(iii) Also,

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

**Moral.** Intuitive meaning to $f$:
Suppose $\Delta x$ small. Then

$$\mathbb{P}(x < X \leq x + \Delta x) = \int_{x}^{x+\Delta x} f(y)\,\mathrm{d}y \underset{\Delta x \text{ small}}{\simeq} \Delta x \cdot f(x)$$




More generally, for any set $S \subseteq \mathbb{R}$

$$\mathbb{P}(X \in S) = \int_{S} f(x)\,\mathrm{d}x$$

## 3.2 Expectation

**Definition.** Let $X \geq 0$ with density $f$. We define its **expectation**

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

Suppose $g > 0$. Then

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

for any variable $X$
Let $X$ be a general r.v.
Define

$$X_+ = \max(X, 0)$$

and

$$X_- = \max(-X, 0)$$

and if at least one of $\mathbb{E}[X_+]$ or $\mathbb{E}[X_-]$ is finite, then we set

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

since

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

and

$$\mathbb{E}[X_-] = \int_{-\infty}^0 (-x) f(x) \, \mathrm{d}x$$

Easy to check that the expectation is again a linear function

**Claim.** Let $X \geq 0$. Then

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

**Proof** ($1^{\text{st}}$).

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

**Proof** ($2^{\text{nd}}$).

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

Taking expectation, we get

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

**Equation.**

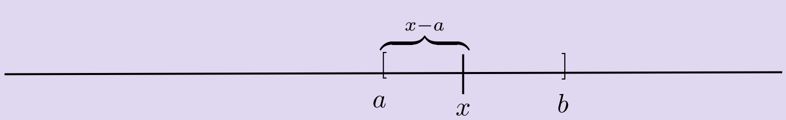
$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

**Example.** Uniform distribution

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

we write $X \sim U[a,b]$

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




$$F(x) = \frac{x-a}{b-a}$$

for $x \in [a,b]$ and $F(x) = 1$ for $x > b$, $F(x) = 0$ otherwise.

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

**Example.** Exponential distribution

$$f(x) = \lambda e^{-\lambda x}, \ \lambda > 0, \ x > 0, \ X \sim \text{ Exp}(\lambda)$$

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

and

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

## 3.3 Exponential as a limit of geometrics

**Equation.** Let $T \sim \text{ Exp}(\lambda)$ and set $T_n = \lfloor nT \rfloor \ \forall n \in \mathbb{N}$

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

So $T_n$ is a geometric of parameter

$$p_n = 1 - e^{-\lambda/n} \sim \frac{\lambda}{n} \text{ as } n \to \infty$$

and

$$\frac{T_n}{n} \to R \text{ as } n \to \infty$$

So the exponential is the limit of a rescaled geometric

**Remark.** Memoryless property:

$$s, t > 0 \ \mathbb{P}(T > t + s | T > s) = e^{-\lambda t} = \mathbb{P}(T > t)$$

$$T \sim \text{ Exp}(\lambda)$$

**Prop.** Let $T$ be a positive r.v. not identically 0 or $\infty$.
Then $T$ has the memoryless property iff $T$ is exponential

**Proof.** $\implies$ :

$$\forall s,t\ \mathbb{P}(T>t+s)=\mathbb{P}(T>s)\mathbb{P}(T>t)$$

Set

$$g(t)=\mathbb{P}(T>t)$$

NTS:

$$g(t)=e^{-\lambda t}\text{ for some }\lambda>0$$

$$g(t+s)=g(t)g(s)\ \forall s,t>0$$

$$t\geq 0,m\in\mathbb{N}\ g(mt)=(g(t))^m$$

$t=1$ gives:

$$\forall m\in\mathbb{N}\ g(m)=g(1)^m$$

$$g\left(\frac{m}{n}\right)^n=g(m)\implies g\left(\frac{m}{n}\right)=g(1)^{m/n},\ \forall m,n\in\mathbb{N}$$

$$g(1)=\mathbb{P}(T>1)\in(0,1)$$

Set

$$\lambda=-\log\mathbb{P}(T>1)>0$$

So we have proved that

$$g(t)=\mathbb{P}(T>t)=e^{-\lambda t}\ \forall t\in\mathbb{Q}_+$$

Let $t\in\mathbb{R}_+$. Then

$$\forall\varepsilon>0\exists r,s\in\mathbb{Q}:r\leq t<s\text{ and }|r-s|<\varepsilon$$

$$e^{-\lambda s}=\mathbb{P}(T>s)\leq\mathbb{P}(T>t)\leq\mathbb{P}(T>r)e^{-\lambda r}$$

Letting $\varepsilon\to 0$ finishes the proof $\square$

**Theorem.** Let $X$ be a continuous r.v. with density $f$. Let $G$ be a continuous function which is either strictly increasing or strictly decreasing and $g^{-1}$ is differentiable.
Then $g(X)$ is a continuous r.v. with density

$$f(g^{-1}(x)) \cdot \left| \frac{\mathrm{d}}{\mathrm{d}x} g^{-1}(x) \right|$$

**Proof.** $g$ increasing:

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

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

$g$ decreasing:

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

since

$$\mathbb{P}(X = g^{-1}(x)) = 0$$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(g(x) \leq x) &= -f(g^{-1}(x)) \cdot \frac{\mathrm{d}}{\mathrm{d}x} g^{-1}(x) \\ &= f(g^{-1}(x)) \cdot \left| \frac{\mathrm{d}}{\mathrm{d}x} g^{-1}(x) \right| \end{aligned}$$

$g^{-1}$ decreasing and so

$$\frac{\mathrm{d}}{\mathrm{d}x} g^{-1}(x) < 0$$

**Example.** Normal distribution:
$-\infty < \mu < \infty,\ \sigma > 0$ are our 2 parameters.

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

Check if $f$ is a density:

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

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{u^2}{2}\right)\mathrm{d}u = 2\int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,\mathrm{d}u = I$$

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

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

$$I^2 = \frac{2}{\pi}\int_0^{\infty}\int_0^{\frac{\pi}{2}} re^{-(r^2)/2}\,\mathrm{d}r\,\mathrm{d}\theta = 1 \implies I = 1$$

as desired
So $f$ is a density
Let $X$ have density $f$

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

first integral is 0 by $u = (x-\mu)/\sigma$
So $\mathbb{E}[X] = \mu$

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X-\mu)^2] \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\ &= \sigma^2 \int_{-\infty}^{\infty} \frac{u^2}{\sqrt{2\pi}} e^{-u^2/2}\,\mathrm{d}u = \sigma^2\end{aligned}$$

So $\mathrm{Var}(X) = \sigma^2$
When $X$ has density $f$, we write $X \sim N(\mu, \sigma^2)$
($X$ is normal with parameters $\mu$ and $\sigma^2$)
When $\mu = 0$ and $\sigma^2 = 1$,, we call $N(0,1)$ the standard normal.
If $X \sim N(0,1)$, we write

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

and

$$\varphi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Have

$$\varphi(x) = \varphi(-x) \implies \Phi(x) + \Phi(-x) = 1 \implies \mathbb{P}(X \le x) = 1 - \mathbb{P}(X \le -x)$$

**Method.** Let $a \neq 0,\ b \in \mathbb{R}$. Set $g(x) = ax + b$
Define $Y = g(X)$. What is the density of $Y$?

$$Y = aX + b$$

$$g^{-1}(x) = \frac{x-b}{a}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}x} g^{-1}(x) = \frac{1}{a}$$

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

So $Y \sim N(a\mu + b, a^2\sigma^2)$
$\sigma$ is the 'standard deviation'.
Suppose $X \sim N(\mu, \sigma^2)$, then

$$\frac{X-\mu}{\sigma} \sim N(0,1)$$

$$\mathbb{P}(-2\sigma < X - \mu < 2\sigma) = \mathbb{P}\left( -2 < \frac{X-\mu}{\sigma} < 2 \right) = \mathbb{P}\left( \left| \frac{X-\mu}{\sigma} \right| < 2 \right) = \Phi(2)$$

and $\Phi(2) \geq 0.95$ (using tables for $\Phi$)
With prob. $\geq 95\%$, the normal is within 2 standard deviations of the mean

**Definition.** Suppose that $X$ is a continuous r.v. The **median** of $X$, denoted by $m$, is the number satisfying

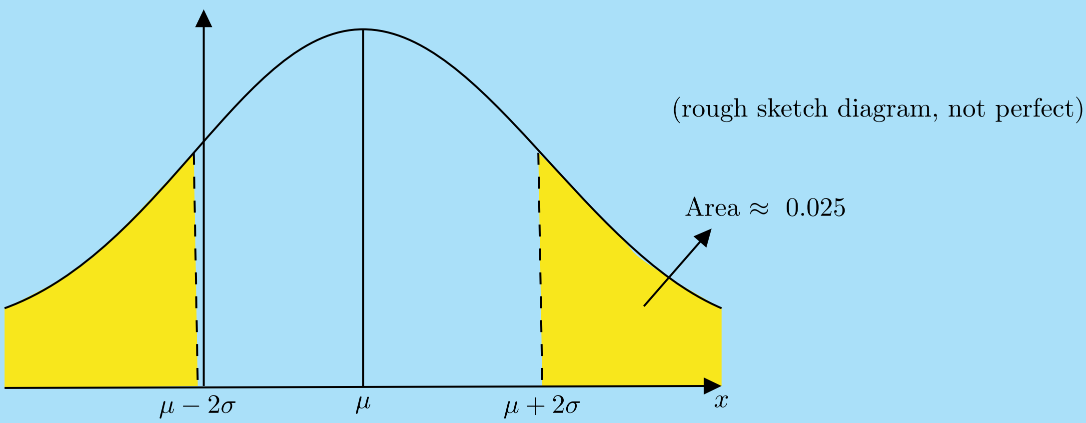
$$\mathbb{P}(X \le m) = \mathbb{P}(X \ge m) = \frac{1}{2}$$

In other words

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

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

$$\mathbb{P}(X \le \mu) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \le 0\right) = \Phi(0) = \frac{1}{2}$$




## 3.4 Multivariate Density Functions

**Equation.** $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ r.v.
We say that $X$ has density $f$ if

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

Then

$$f(X_1, \dots, X_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} F(x_1, \dots, x_n)$$

This generalises: "$\forall$" $B \subseteq \mathbb{R}^n$

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

**Definition.** We say that $X_1, \dots, X_n$ are **independent** if $\forall x_1, \dots, x_n$,

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

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

(i) Suppose $X_1, \dots, X_n$ are independent with densities $f_1, \dots, f_n$. Then

$$f(x_1, \dots, x_n) = f_1(x_1) \dots f_n(x_n) \tag{*}$$

(ii) Suppose that $f$ factorises as in (*) for some non-negative functions $(f_i)$. Then $X_1, \dots, X_n$ are independent and have densities proportional to the $f_i$'s

**Proof.**

(i)

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

So the density of $(X_1, \dots, X_n)$ is $f = \prod f_i$

Let $B_1, \dots, B_n \subseteq \mathbb{R}$ then

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

Take $B_j = \mathbb{R} \, \forall j \neq i$. Then

$$\mathbb{P}(X_i \in B_i) = \mathbb{P}(X_i \in B_i, X_j \in B_j \ \forall j \neq i) = \int_{B_i} f_i(y_i) \, \mathrm{d}y_i \prod_{j \neq i} \int_{\mathbb{R}} f_j(y) \, \mathrm{d}y$$

Since $f$ is a density

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

But $f = \prod f_i$, so

$$\prod_j \int_{-\infty}^{\infty} f_j(y) \, \mathrm{d}y = 1$$

$$\implies \prod_{j \neq i} \int_{\mathbb{R}} f_j(y) \, \mathrm{d}y = \frac{1}{\int_{\mathbb{R}} f_i(y) \, \mathrm{d}y}$$

So

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

This shows that the density of $x_i$ is

$$\frac{f_i}{\int_{\mathbb{R}} f_i(y) \, \mathrm{d}y}$$

The $X_i$'s are independent, since

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \frac{\int_{-\infty}^{x_1} f_1(y_1) \, \mathrm{d}y_1 \cdots \int_{-\infty}^{x_n} f_n(y_1 n \, \mathrm{d}y_n}{\int_{\mathbb{R}} f_1(y_1) \, \mathrm{d}y_1 \cdots \int_{\mathbb{R}} f_n(y_1 n \, \mathrm{d}y_n} = \mathbb{P}(X_1 \le x_1) \dots \mathbb{P}(X_n \le x_n) \square$$

**Equation.** Suppose $(X_1, \dots, X_n)$ has density $f$

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

density of $X_1$ = marginal density of $X_1$

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) \, \mathrm{d}x_2 \dots \mathrm{d}x_n$$

## 3.5 Density of the Sum of Independent r.v.'s

**Equation.** Let $X$ and $Y$ be 2 independent r.v.'s with densities $f_X$ and $f_Y$ respectively.

$$\begin{aligned}
\mathbb{P}(X + Y \le z) &= \iint\limits_{\{x+y \le z\}} f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{z} f_Y(y - x) f_X(x) \, \mathrm{d}y \right) \mathrm{d}x \\
&= \int_{-\infty}^{z} \mathrm{d}y \left( \int_{-\infty}^{\infty} f_Y(y - x) f_X(x) \, \mathrm{d}x \right)
\end{aligned}$$

So the density of $X + Y$ is

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

We call this function the convolution of $f_X$ and $f_Y$

**Definition.** $f, g$: 2 densities

$$f * g(x) = \int_{-\infty}^{\infty} f(x - y) g(y) \, \mathrm{d}y = \textbf{convolution} \text{ of } f \text{ and } g$$

**Moral.** We can non-rigorously show this

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

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

So the density of $X+Y$ is

$$\int_{-\infty}^{\infty} f_X(z-y)F_Y(y)\,\mathrm{d}y$$

### 3.6 Conditional Density

**Definition.** Let $X$ and $Y$ be continuous variables with joint density $f_{X,Y}$ and marginal densities $f_X$ and $f_Y$. Then the **conditional density** of $X$ given $Y=y$ is defined

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

### 3.7 Law of Total Probability

**Equation.**

$$f_X(x) = \int_{\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_{\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)\,\mathrm{d}y$$

**Remark.** Want to define $\mathbb{E}[X|Y]=g(Y)$ for some function $g$.
Define

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

Set $\mathbb{E}[X|Y]=g(Y)=$ conditional expectation of $X$ given $Y$.

## 3.8 Transformation of a multidimensional r.v.

**Theorem.** Let $X$ be a r.v. with values in $D \subseteq \mathbb{R}^d$ and with density $f_X$.
Let $g$ be a bijection from $D$ to $g(D)$ which has a continuous derivative on $D$ and

$$\det g'(x) \neq 0 \ \forall x \in D$$

Then the r.v. $Y = g(X)$ has density

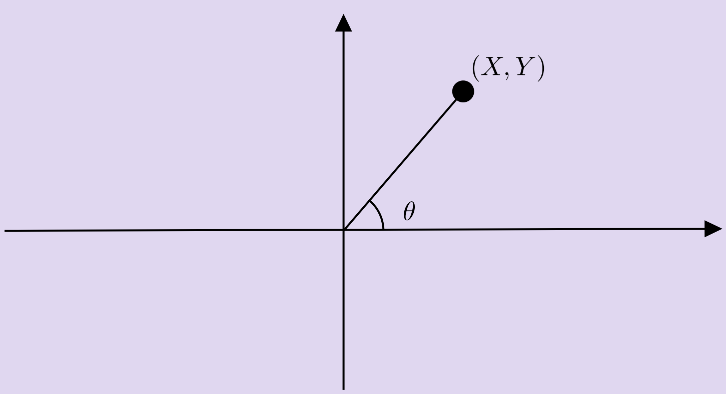
$$f_Y(y) = f_X(x) \cdot |J|$$

where $x = g^{-1}(y)$ and $J$ is the determinant of the Jacobian

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

**Proof.** We do not prove it here.

**Example.** Let $X$ and $Y$ be independent $N(0,1)$ r.v.'s




Want the density of $(R, \theta)$

$$X = R\cos\theta$$

$$Y = R\sin\theta$$

$$f_{R,\theta}(r,\theta) = f_{X,Y}(r\cos\theta, r\sin\theta) \cdot |J|$$

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

So

$$f_{R,\theta}(r,\theta) = f_X(r\cos\theta) \cdot f_Y(r\sin\theta) \cdot r = \frac{1}{\sqrt{2\pi}} e^{-(r^2\cos^2\theta)/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(r^2\sin^2\theta)/2} \cdot r$$

$$\implies f_{R,\theta}(r,\theta) = \frac{1}{2\pi} r e^{\frac{r^2}{2}}$$

for all $r > 0$ and $\theta \in [0, 2\pi]$
This shows that $R$ and $\theta$ are independent with $\theta \sim U[0, 2\pi]$ and $R$ has density $re^{\frac{r^2}{2}}$ on $(0, \infty)$

## 3.9 Order Statistics for a Random Sample

**Method.** Let $X_1, \dots, X_n$ be iid with distr. function $F$ and density $f$.
Put them in increasing order

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

and set

$$Y_i = X_{(i)}$$

Then $(Y_i)$ are the order statistics.

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

$$f_{Y_1}(x) = \frac{\mathrm{d}}{\mathrm{d}x} (1 - (1 - F(x))^n) = n \cdot (1 - F(x))^{n-1} \cdot f(x)$$

$$\mathbb{P}(Y_n \leq x) = (F(x))^n$$

$$f_{Y_n}(x) = n(F(x))^{n-1} \cdot f(x)$$

Density of $Y_1, \dots, Y_n$?

$$\begin{aligned} \mathbb{P}(Y_1 \leq x_1, \dots, Y_n \leq x_n) &= n!\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n,\ X_1 < X_2 < \dots < X_n) \\ &= n! \int_{-\infty}^{X_!} \int_{u_1}^{X_2} \cdots \int_{u_{n-1}}^{x_n} f(u_1) \dots f(u_n) \,\mathrm{d}u_1 \,\dots\, \mathrm{d}u_n \end{aligned}$$

by differentiating we get

$$f_{Y_1, \dots, Y_n}(x_1, \dots, x_n) = \begin{cases} n! f(x_1) \dots f(x_n) & \text{when } X_1 < X_2 < \dots X_n \\ 0 & \text{otherwise} \end{cases}$$

**Example.** Let $X \sim \text{Exp}(\lambda)$ and $X \sim \text{Exp}(\mu)$, $X \perp\!\!\!\perp Y$.
Set $Z = \min(X, Y)$

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

So $Z \sim \text{Exp}(\lambda + \mu)$
Mroe generally, if $X_1, \dots, X_n$ are independent with $X_i \sin \text{Exp}(\lambda_i)$ then

$$\min(X_1, \dots, X_n) \sim \text{ Exp}\left(\sum_{i=1}^{n} \lambda_i\right)$$

Let $X_1, \dots, X_n$ be iid $\text{Exp}(\lambda)$ and let $Y_i$ be their order statistics

$$Z_1 = Y_1,\ Z_2 = Y_2 - Y_1, \dots, Z_n = Y_n - Y_{n-1}$$

Density of $(Z_1, \dots, Z_n) = ?$

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix} = A \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

where

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ & & \vdots & & & \end{bmatrix}$$

have $\det A = 1$ and let $Z = Ay$, then

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

$$\begin{aligned} f_{(Z_1,\dots,Z_n)(z_1,\dots,z_n)} &= f_{(Y_1,\dots,Y_n)}(y_1, \dots, y_n) \cdot |J| \\ &= n! f(y_1) \dots f(y)n) \\ &= n! \lambda e^{-\lambda y_1} \dots \lambda e^{-\lambda y_n} \\ &= n! \lambda^n e^{-\lambda(nz_1 + (n-1)z_2 + \dots + z_n)} \\ &= \prod_{i=1}^{n} (n - i + 1_) \lambda e^{-\lambda(n-i+1} z_i \end{aligned}$$

So $Z_1, \dots, Z_n$ are independent and $Z_i \sim \text{ Exp}(\lambda(n - i + 1))$

## 3.10 Moment Generating Functions (mgfs)

**Definition.** Let $X$ be a r.v. with density $f$. The **mgf** of $X$ is defined to be

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

whenever this integral is finite

$$m(0) = 1$$

**Theorem.** The mgf uniquely determines the distribution of a r.v. provided it is defined for an open interval of values of $\theta$.

**Theorem.** Suppose the mgf is defined for an open interval of values of $\theta$. Then

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

**Example.** Gamma distribution:

$$f(x) = \frac{e^{-\lambda x}\lambda^n x^{n-1}}{(n-1)!},\ \lambda > 0,\ n \in \mathbb{N},\ x \geq 0$$

We denote $X$ with density $f$ as $X \sim \Gamma(m, \lambda)$
Check $f$ is a density:

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

for $n = 1$ $f(x) = \lambda e^{-\lambda x} \implies \mathrm{Exp}(\lambda)$. So $I_1 = 1$

$$\begin{aligned} m(\theta) &= \int_0^\infty e^{\theta x} \cdot e^{-\lambda x} \cdot \frac{\lambda^n x^{m-1}}{(n-1)!} \\ &= \int_0^\infty e^{(\lambda-\theta)x} \frac{\lambda^n x^{n-1}}{(n-1)!}\,\mathrm{d}x \\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^n \text{ for } \lambda > 0 \end{aligned}$$

**Claim.** Suppose that $X_1, \ldots, X_n$ are independent r.v's. Then

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

**Example.** Let $X \sim \Gamma(n, \lambda)$ and $Y \sim \Gamma(m, \lambda)$ and $X \perp\!\!\!\perp Y$. Then

$$\begin{aligned} m(\theta) &= \mathbb{E}\left[e^{\theta(X+Y)}\right] \\ &= \mathbb{E}[e^{\theta X}] \cdot \mathbb{E}[e^{\theta Y}] \\ &= \left(\frac{\lambda}{\lambda - \theta}\right)^n \cdot \left(\frac{\lambda}{\lambda - \theta}\right)^m \\ &= \left(\frac{\lambda}{\lambda - \theta}\right)^{n+m} \quad \text{for } \theta < \lambda \end{aligned}$$

So by the uniqueness theorem we get $X + Y \sim \Gamma(n+m, \lambda)$.
In particular, this implies that if $X_1, \dots, X_n$ are iid Exp(1) ($= \Gamma(1, \lambda)$) then

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

**Remark.** One could also consider $\Gamma(\alpha, \lambda)$ ($\alpha > 0$) by replacing $(n-1)!$ with

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

**Example.** Normal distribution. Let $X \sim N(\mu, \sigma^2)$

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

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

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

So

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

as

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

gives normal distribution
If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$
So

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

Suppose $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\mu, \tau^2)$ and $X \perp\!\!\!\perp Y$
Then

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

So $X + Y \sim N(\mu + \nu, \sigma^2 + \tau^2)$

**Example.** Cauchy distribution
$$f(x) = \frac{1}{\pi(1+x^2)}\ x \in \mathbb{R}$$

$$\begin{aligned} m(\theta) &= \mathbb{E}[e^{\theta X}] \\ &= \int_{-\infty}^{\infty} \frac{e^{\theta x}}{\pi(1+x^2)}\,\mathrm{d}x \\ &= \infty\ \forall \theta \neq 0,\ (m(0) = 1) \end{aligned}$$

Suppose $X \sim f$. Then $X, 2X, 3X, \ldots$ all have the same mgf.
However they do not have the same distribution.
So assumption on $m(\theta)$ being finite for an open interval of values of $\theta$ is essential

### 3.11 Multivariate Moment Generating Function

**Definition.** Let $X = (X_1, \ldots, X_n)$ be a r.v. with values in $\mathbb{R}^n$. Then the **mgf** of $X$ is defined to be
$$m(\theta) = \mathbb{E}[e^{\theta^T X}] = \mathbb{E}[e^{\theta_1 X_1 + \cdots + \theta_n X_n}]$$
where
$$\theta = (\theta_1, \ldots, \theta_n)^T$$

**Theorem.** In this case, provided mgf is finite for a range for values of $\theta$, it uniquely determines the distribution of $X$. also
$$\left.\frac{\partial^r m}{\partial \theta_i^r}\right|_{\theta=0} = \mathbb{E}[X_i^r]$$
$$\left.\frac{\partial^{r+s} m}{\partial \theta_i^r \partial \theta_j^s}\right|_{\theta=0} = \mathbb{E}[X_i^r X_j^s]$$
$$m(\theta) = \prod_{i=1}^{n} \mathbb{E}[e^{\theta_i X_i}] \text{ iff } X_1, \ldots, X_n \text{ are indep.}$$

**Definition.** Let $(X_n : n \in \mathbb{N})$ be a sequence of r.v.'s and let $X$ be another r.v.
We say that $X_n$ **converges to X in distribution** and write $X_n \xrightarrow{d} X$, if
$$F_{X_n}(x) \to F_X(x)\ \forall x \in \mathbb{R} \text{ that are continuity points of } F_X$$

**Theorem** (Continuity Property for mgf's)**.** Let $X$ be a r.v. with $m(\theta) < \infty$ for some $\theta \neq 0$. suppose that
$$m_n(\theta) \to m(\theta)\ \forall \theta \in \mathbb{R} \text{ where } m_n(\theta) = \mathbb{E}[e^{\theta X_m}] \text{ and } m(\theta) = \mathbb{E}[e^{\theta X}]$$
Then $X_n$ converges to $X$ in distribution

## 3.12 Limit Theorems for Sums of iid r.v.'s

**Theorem** (Weak Law of Large Numbers)**.** Let $(X_n : n \in \mathbb{R})$ be a sequence of iid r.v.'s with $\mu = \mathbb{E}[X_1] < \infty$. Set

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

Then $\forall \varepsilon > 0$

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

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

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

$$S_n = X_1 + \cdots + X_n \implies \mathrm{Var}(S_n) = n\sigma^2$$

**Definition.** A sequence $(X_n)$ **converges to X in probability** and we write

$$X_n \xrightarrow{\mathbb{P}} X \text{ as } n \to \infty$$

if $\varepsilon > 0$:

$$\mathbb{P}(|X_n - X > \varepsilon) \to 0 \text{ as } n \to \infty$$

**Definition.** We say $(X_n)$ **converges to X with probability 1** or '**almost surely** (a.s.)' if

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

**Claim.** Suppose $X_n \to 0$ a.s. as $n \to \infty$. Then $X_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$

**Proof.** NTS:

$$\forall \varepsilon > 0 \ \mathbb{P}(|X_n| > \varepsilon) \to 0 \text{ as } n \to \infty$$

or equivalently

$$\mathbb{P}(|X_n| \leq \varepsilon) \to 1 \text{ as } n \to \infty$$

$$\mathbb{P}(|X_n| \leq \varepsilon) \geq \mathbb{P}\left(\underbrace{\bigcap_{m=n}^{\infty} \{|X_m| \leq \varepsilon\}}_{A_n}\right) \ A_n \subseteq A_{n+1}$$

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

So

$$\mathbb{P}(A_n) \to \mathbb{P}\left(\bigcup_n A_n\right) \text{ as } n \to \infty$$

So

$$\lim_{n\to\infty} \mathbb{P}(|X_n| \leq \varepsilon) \geq \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup A_n\right) > \mathbb{P}\left(\lim_{X_n = 0}\right) = 1 \ \square$$

**Theorem** (Strong law of large numbers)**.** Let $(X_n)_{n\in\mathbb{N}}$ be an iid sequence of r.v.'s with $\mu = \mathbb{E}[X_1] < \infty$.
Then setting

$$S_N = X_1 + \cdots + X_n$$

we have

$$\frac{S_n}{n} \to \mu \text{ as } n \to \infty \text{ a.s.}$$

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

**Proof** (non-examinable)**.** Assume further that $\mathbb{E}[X_1^4] < \infty$
Set $Y_i = X_i - \mu$. Then $\mathbb{E}[Y_i] = 0$ and

$$\mathbb{E}[Y_1^4] \le 2^4(\mathbb{E}[X_1]^4 + \mu^4) < \infty$$

It suffices to prove

$$\frac{S_n}{n} \to 0 \text{ where } S_n = \sum_{i=1}^n X_i \text{ with } \mathbb{E}[X_i] = 0 \text{ and } \mathbb{E}[X_i^4] < \infty$$

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

where $R$ is a sum of terms of the form $X_i^2 X_j X_k$ or $X_i^3 X_j$ or $X_i X_j X_k X_l$ for $i, jk, l$ distinct.

$$\mathbb{E}[S_n^4] = n\mathbb{E}[X_1^4] + \binom{4}{2}\frac{n\cdot(n-1)}{2}\mathbb{E}[X_1^2 X_2^2] + \underbrace{\mathbb{E}[R]}_{=0}$$

So

$$\mathbb{E}[S_n^4] \le n \cdot \mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^4]$$
$$\mathbb{E}[S_n^4] \le 3n^2\mathbb{E}[X_1^4]$$

So

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

which implies that

$$\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4 < \infty \text{ w.p. } 1$$

$$\implies \frac{S_n}{n} \to 0 \text{ as } n \to \infty \text{ w.p. } 1 \; \square$$

**Equation.** Suppose $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}(X_1) = \sigma^2 < \infty$

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

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

## 3.13 Central limit theorem

**Theorem.** Let $(X_n)_{n\in\mathbb{N}}$ be an iid sequence of rv.'s with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Set

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

Then

$$\forall x \in \mathbb{R},\ \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) \to \Phi(x) = \int_{-\infty}^{x} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\,\mathrm{d}y \text{ as } n \to \infty$$

In other words,

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

where $Z \sim N(0,1)$
CLT says that for $n$ large enoguh

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \approx Z\ Z \sim N(0,1)$$

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

**Proof.** Consider $Y_1 = (X_i - \mu)/\sigma$. Then $\mathbb{E}[Y_1] = 0$ and $\mathrm{Var}(Y_i) = 1$.
It suffices to prove the CLT when

$$S_n = X_1 + \cdots + X_n \text{ with } \mathbb{E}[X_i] = 0 \text{ and } \mathrm{Var}(X_i) = 1$$

Assume further that $\exists\delta > 0$ s.t.

$$\mathbb{E}[e^{\delta X_1}] < \infty \text{ and } \mathbb{E}[e^{-\delta X_1}] < \infty$$

NTS

$$\frac{S_n}{\sqrt{n}} \to N(0,1) \text{ as } n \to \infty$$

By the continuity property of mgf's, it suffices to show $\forall\theta \in \mathbb{R}$

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

Set

$$m(\theta = \mathbb{E}[e^{\theta X_1}]$$

Then

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

NTS

$$\left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n \to e^{\theta^2/2} \text{ as } n \to \infty$$

$$|\theta| \le \frac{\delta}{2}\ m(\theta) = \mathbb{E}\left[e^{\theta X_1}\right] = \mathbb{E}\left[1 + \theta X_1 + \frac{\theta^2 X_1^2}{2!} + \sum_{k=1}^{\infty} \frac{\theta^k X_1^k}{k!}\right]$$

**Theorem** (cont.).

**Proof** (cont.). So

$$m(\theta) = 1 + \frac{\theta^2}{2} + \mathbb{E}\left[\sum_{k \ge 3} \frac{\theta^k X_1^k}{k!}\right]$$

**Claim.** It suffices to prove that

$$\left|\mathbb{E}\left[\sum_{k \ge 3} \frac{\theta^k X_1^k}{k!}\right]\right| = o(|\theta|^2) \text{ as } \theta \to 0$$

Once we prove this bound, then

$$m\left(\frac{\theta}{\sqrt{n}}\right) = 1 + \frac{\theta^2}{\sqrt{n}} = 1 + \frac{\theta^2}{2n} + o\left(\frac{|\theta|^2}{n}\right)$$

and hence

$$\left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^m \to e^{\theta^2/2} \text{ as } n \to \infty$$

**Proof** (of claim).

$$\begin{aligned}
\left|\mathbb{E}\left[\sum_{k \ge 3} \frac{\theta^k X_1^k}{k!}\right]\right| &\le \mathbb{E}\left[\sum_{k \ge 3} \frac{|\theta|^k X_1^k}{k!}\right] \\
&= \mathbb{E}\left[|\theta X_1|^3 \sum_{k=0}^{\infty} \frac{|\theta X_1|^k}{(k+3)!)}\right] \\
&\le \mathbb{E}\left[|\theta X_1|^3 \sum_{k=0}^{\infty} \frac{|\theta X_1|^k}{k!)}\right] \\
&\le \mathbb{E}\left[|\theta X_1|^3 \cdot e^{\frac{\delta}{2}|X_1|}\right]
\end{aligned}$$

as $|\theta| \le \frac{\delta}{2}$.

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

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

so

$$\left|\mathbb{E}\left[\sum_{k \ge 3} \frac{\theta^k X_1^k}{k!}\right]\right| \le 3! \cdot \left(\frac{2|\theta|}{\delta}\right)^3 \underbrace{\mathbb{E}[e^{\delta|X_1|} + e^{-\delta|X_1|}]}_{<\infty} = o((|\theta|^2) \text{ as } \theta \to 0 \ \square$$

### 3.14 Applications

**Example.** Normal approximation to the binomia distr.
Let $S_n \sim \text{ Bin}(n,p)$

$$S_n = \sum_{i=1}^{n} X_i,\ (X_i) \text{ iid } \sim \text{ Ber}(p)\ \mathbb{E}[S_n] = np, \text{Var}(S_n) = np(1-p)$$

So by the CLT

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty$$

So

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

$$\text{Bin}\left(n, \frac{\lambda}{n}\right) \to \text{ Poi}(\lambda)\ \lambda > 0$$

**Example.** Normal approx. to the Poisson distribution:
Let $S_n \sim \text{ Poi}(n)$.

$$S_n = \sum_{i=1}^{n} X_i,\ (X_i) \text{ iid } \sim \text{ Poi}(1)$$

$$\frac{S_n - n}{\sqrt{n}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty$$

## 3.15 Sampling Error via the CLT

**Example.** Pick $N$ individuals at random. Let

$$\hat{p}_N = \frac{S_N}{N}$$

where $S_N$ is the number of yes voters.
How large should $N$ be so that

$$|\hat{p}_N - p| \leq \frac{4}{100} \text{ w.p. } \geq 0.99?$$

By the CLT

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

So

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

Find $N$ s.t.

$$\mathbb{P}(|\hat{p}_N - p| \leq 0.04) \geq 0.99$$

or equivalently

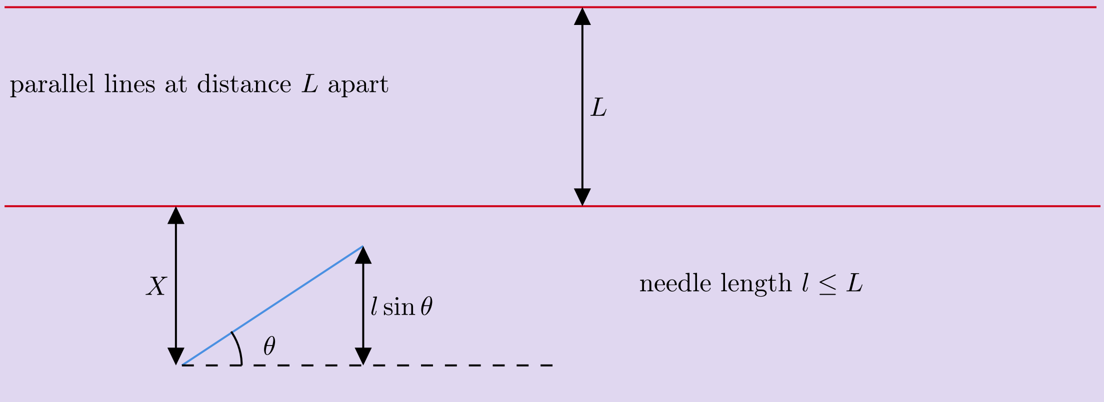
$$\mathbb{P}\left(\sqrt{\frac{p(1-p)}{N}} \cdot |Z| \leq 0.04\right) \geq 0.99$$

$z = 2.58$ $\mathbb{P}(|Z| \geq 2.58) = 0.01$
So we need

$$0.04\sqrt{\frac{N}{p(1-p)}} \geq 2.58 \implies N \geq 1040$$

### 3.16 Buffon's Needle

**Example.**




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 $\pi$. 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

$$\mathbb{P}(|\hat{\pi}_n - \pi| \le 0.001) \ge 0.99$$

How large should $n$ be?

**Example** (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

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

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

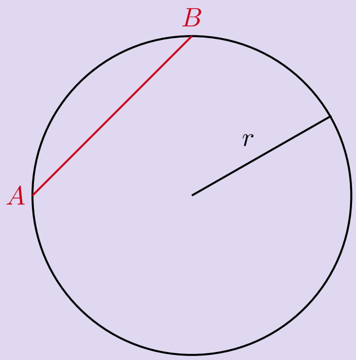
$$\text{Var} = \frac{\pi^2}{n}\left(\frac{\pi}{2} - 1\right)$$

Taking

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

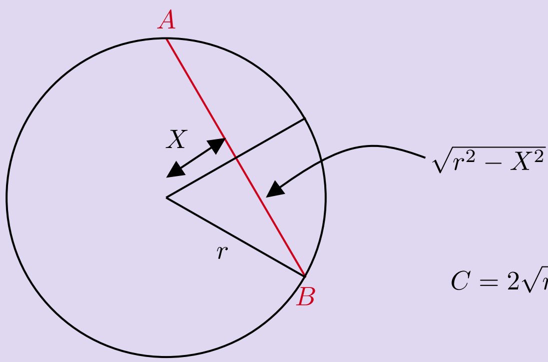
## 3.17 Bertrand's Paradox

**Example.**




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

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



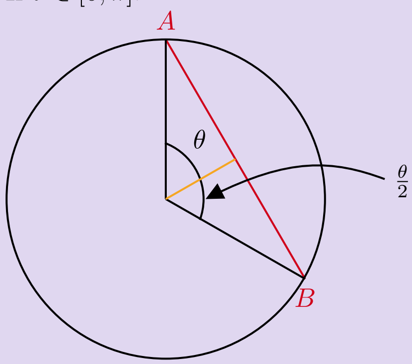

$C = 2\sqrt{r^2 - X^2}$

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

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

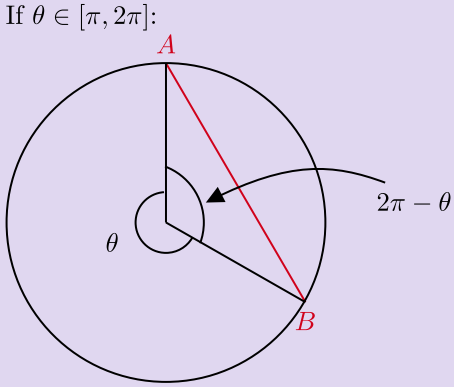
$$\begin{aligned}
\mathbb{P}(C \leq r) &= \mathbb{P}(2\sqrt{r^2 - X^2} \leq r) \\
&= \mathbb{P}(4(r^2 - X^2) \leq r^2) \\
&= \mathbb{P}(4X^2 \geq 3r^2) \\
&= \mathbb{P}(X \geq \sqrt{3}3/2) \\
&= 1 - \frac{\sqrt{3}}{2} \\
&\approx 0.134
\end{aligned}$$

**Example** (cont.)**.** 2nd interpretation: Let $\theta \sim [0, 2\pi]$
Let $C = |AB|$
If $\theta \in [0, \pi]$:




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

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




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

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

## 3.18 Multidimensional Gaussian r.v.'s

**Definition.** A r.v. $X$ with values in $\mathbb{R}$ is called **Gaussian/ normal** if

$$X = \mu + \sigma Z,\ \mu \in \mathbb{R},\ \sigma \in [0, \infty] \text{ and } Z \sim N(0,1)$$

The density of $X$ is

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

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

**Definition.** Let $X = (X_1, \dots, X_n)^T$ with values in $\mathbb{R}^n$. Then $X$ is a **Gaussian vector** or is just called **Gaussian** if $\forall u = (u_1, \dots, u_n)^T \in \mathbb{R}^n$

$$u^T X = \sum_{i=1}^n u_i X_i \text{ is a Gaussian r.v. in } \mathbb{R}$$

**Example.** Suppose $X$ is Gaussian in $\mathbb{R}^n$. Suppose $A$ is an $m \times n$ matrix and $b \in \mathbb{R}^m$. Then $AX + b$ is also Guassian in $\mathbb{R}^m$.

**Proof.** Let $u \in \mathbb{R}^m$. THen

$$u^T(AX + b) = (u^T A)X + u^T b$$

Set $v = A^T u$. Then

$$u^T(AX + b) = v^T X + u^T b = v^T X + \sum_{i=1}^m u_i b_i$$

Since $X$ is Gaussian, we get $v^T X$ is Gaussian, and also $v^T X + u^T b$ is Gaussian.

**Definition.**

$$\mu = \mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix} \quad \mu_i = \mathbb{E}[X_i]$$

$$V = \mathrm{Var}(X) = \mathbb{E}[\underset{n\times 1}{(X-\mu)} \cdot \underset{1\times n}{(X-\mu)^T}] = \begin{bmatrix} \ddots & \vdots & \\ & \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] & \\ & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} \ddots & \vdots & \\ & \mathrm{Cov}(X_i, X_j)] & \\ & \vdots & \ddots \end{bmatrix}$$

**Equation.** $V$ is a symmetric matrix

$$\mathbb{E}[u^T X] = \mathbb{E}\left[\sum_{i=1}^n u_i X_i\right] = \sum_{i=1}^n u_i \mu_i = u^T \mu$$

$$\text{Var}(u^T X) = \text{Var}\left(\sum_{i=1}^n u_i X_i\right) = \sum_{i,j=1}^n u_i \text{Cov}(X_i X_j) u_j = u^T V u$$

So $u^T X \sim N(u^T \mu, u^T V u)$

**Claim.** $V$ is a non-negative definite matrix/ $(\forall u \in \mathbb{R}^n,\ u^T V u \geq 0)$

**Proof.** Let $u \in \mathbb{R}^n$. Then

$$\text{Var}(u^T X) = u^T V u$$

Since $\text{Var}(u^T X) \geq 0$, we have

$$u^T V u \geq 0 \; \square$$

**Method.** Finding mgf of $X$:

$$m(\lambda) = \mathbb{E}[e^{\lambda^T X}] \; \forall \lambda \in \mathbb{R}^n,\ \lambda = (\lambda_1, \dots, \lambda_n)^T$$

$$m(\lambda) = \mathbb{E}[e^{\lambda^T X}] = e^{\lambda^T \mu + \lambda^T V \lambda / 2}$$

We know

$$\lambda^T X \sim N(\lambda^T \mu, \lambda^T V \lambda)$$

So $m(\lambda)$ is characterised by $\mu$ and $V$. Since the mgf uniquely characterises the distribution, we see that a Gaussian vector is uniquely characterised by its mean $\mu$ and variance $V$.
In this case we write $X \sim N(\mu, V)$

**Claim.** Let $Z_q, \dots, Z_n$ iid $N(0,1)$ r.v.'s .
Set $Z = (Z_1, \dots, Z_n)^T$. Then $Z$ is a Gaussian vector.

**Proof.** $\forall u \in \mathbb{R}^n$ $u^T Z$ is Gaussian.

$$u^T Z = \sum_{i=1}^{n} u_i Z_i$$

NTS $\sum\limits_{i=1}^{n} u_i Z_i$ is normal.
Let $\lambda \in \mathbb{R}$.

$$\begin{aligned}
\mathbb{E}[e^{\lambda \sum\limits_{i=1}^{n} u_i Z_i}] &= \mathbb{E}[\prod_{i=1}^{n} e^{\lambda u_i Z_i}] \\
&= \prod_{i=1}^{n} \mathbb{E}[e^{\lambda u_i Z_i}] \\
&= \prod_{i=1}^{n} e^{(\lambda u_i)^2/2} \\
&= e^{\lambda^2 |u|^2/2}
\end{aligned}$$

So $u^T Z \sim N(0, |u|^2)\square$

$$\mathbb{E}[Z] = 0 \ \mathrm{Var}(Z) = I_n = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$$

So $Z \sim N(0, I_n)$

**Method.** Let $\mu \in \mathbb{R}^n$ and $V$ a non-negative definite matrix.
We want to construct a Gaussian vector with mean $\mu$ and variance $V$ using $Z$.
$n = 1:\ \mu, \sigma^2$: If $Z \sim N(0,1)$ then $\mu + \sigma Z \sim N(\mu + \sigma^2)$
Since $V$ is non-negative, definite, $V = U^T D U$ with $U^{-1} = U^T$ and

$$D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

and $\lambda_i \geq 0\ \forall i$
We define the square root of $V$ to be the matrix

$$\sigma = U^T \sqrt{D} U$$

where

$$\sqrt{D} = \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n} \end{bmatrix}$$

Indeed

$$\sigma \cdot \sigma = U^T \sqrt{D} U U^T \sqrt{D} U = U^T D U = V$$

Let $Z = (Z_1, \ldots, Z_n)$ with $(Z_i)$ iid $N(0,1)$ r.v.'s
Set $X = \mu + \sigma Z$

**Claim.** $X \sim N(\mu, V)$

**Proof.** $X$ is Gaussian, since it is a linear transformation of the Gaussian vector $Z$.

$$\mathbb{E}[X] = \mu$$

and

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[(X-\mu)(X-\mu)^T \\ &= \mathbb{E}[(\sigma Z) \cdot (\sigma Z)^T] \\ &= \mathbb{E}[\sigma Z \cdot Z^T \cdot \sigma^T] \\ &= \sigma \cdot \mathbb{E}[Z \cdot Z^T] \sigma \\ &= \sigma I_n \sigma \\ &= \sigma\sigma \\ &= V \ \square \end{aligned}$$

**Method.** Finding density of $X \sim N(\mu, V)$

$$n = 1 : X \sim N(\mu, \sigma^2) \ f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Case $V$ is positive definite ($\lambda_i > 0 \ \forall i$):

$$X = \mu + \sigma Z, \ Z \sim N(0, I_n)$$

$$f_X(x) = f_Z(z) \cdot |J| \ x = \mu + \sigma z$$

Since $V$ is positive definite, $\sigma$ is invertible
So

$$x = \mu + \sigma z \implies z = \sigma^{-1}(x - \mu)$$

So

$$f_X(x) = f_Z(z) \cdot |J| = \prod_{i=1}^{n} \left(\frac{e^{-z_i^2/2}}{\sqrt{2\pi}}\right) \cdot |\det \sigma^{-1}|$$

$$\implies f_X(c) = \frac{1}{(2\pi)^{n/2}} e^{-|z|^2/2} \frac{1}{\sqrt{\det V}} = \frac{1}{\sqrt{(2\pi)^n \det V}} e^{z^T z/2}$$

$$\begin{aligned}
z^T \cdot z &= (\sigma^{-1}(x-\mu))^T (\sigma^{-1}(x-\mu)) \\
&= (x-\mu)^T \underbrace{(\sigma^{-1})^T}_{\sigma^{-1}} \sigma^{-1}(x-\mu) \\
&= (x-\mu)^T \sigma^{-1}\sigma^{-1} \cdot (x-\mu) \\
&= (x-\mu)^T \cdot V^{-1} \cdot (x-\mu)
\end{aligned}$$

Therefore

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

Case $V$ is non-negative definite, so some eigenvalues could be 0.
By an orthogonal change of basis, we can assume that

$$V = \begin{bmatrix} U & 0 \\ 0 & 0 \end{bmatrix} \text{ where } U \text{ is an } m \times n \ (m < n) \text{ positive definite matrix}$$

We can write $X = \begin{bmatrix} Y \\ \nu \end{bmatrix}$ where $Y$ has density

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^m \det U}} \exp\left(-\frac{(y-\lambda)^T \cdot U^{-1}(y-\lambda)}{2}\right)$$

**Claim.** If the $X_i$'s are independent, then $V$ is a diagonal matrix

**Proof.** Since the $X_i$'s are independent, it follows that $\mathrm{Cov}(X_i, X_j) = 0$ whenever $i \neq j$. So $V$ is diagonal.

**Lemma.** Suppose that $X$ is a Gaussian vector. Then if $V$ is a diagonal matrix, then the $X_i$'s are independent

**Proof** ($1^{\text{st}}$)**.** If $V$ is diagonal, then the density $f_X(x)$ factorises into a product. Indeed,

$$(x-\mu)^T V^{-1}(x-\mu) = \sum_{i=1}^{n} \frac{(x_i-\mu_i^2)}{\lambda_i} \quad V == \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \quad \lambda_i > 0$$

so

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det V}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{2\lambda_i}\right)$$

Hence the $X_i$'s are indep.

**Proof** ($2^{\text{nd}}$)**.**

$$m(\theta) = \mathbb{E}[e^{\theta^T X}] = e^{\theta^T\mu + \theta^T V\theta/2} = e^{\sum \theta_i \mu_i} \cdot e^{\sum \theta_i \lambda_i/2}$$

So $m(\theta)$ factorises into the mgf's of Gaussian r.v.'s in $\mathbb{R}$ $\square$

**Moral.** So for Gaussian vectors we have

$$(X_1, \dots, X_n) \text{ are independent iff } \mathrm{Cov}(X_i, X_j) = 0 \text{ whenever } i \neq j$$

### 3.19 Bivariate Gaussian

**Definition.** $n = 2$
Let $X = (X_1, X_2)$ be a Gaussian vector in $\mathbb{R}^2$.
Set $\mu_k = \mathbb{E}[X_k],\ k = 1, 2$. Set $\sigma_k^2 = \mathrm{Var}(X_k)$

$$\rho = \mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}}$$

**Claim.** $\rho \in [-1, 1]$

**Proof.** Immediate from the Cauchy-Schwartz ineq. $\square$

$$V = \mathrm{Var}(X) = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

**Claim.** For all $\sigma_k > 0$ and $\rho \in [-1, 1]$

$$V = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \text{ is non-negative definite}$$

**Proof.** Let $u \in \mathbb{R}^2$

$$\begin{aligned} u^T V u &= (1-\rho)(\sigma_1^2 u_1^2 + \sigma_2^2 + \sigma_2^2) + \rho(\sigma_1 u_1 + \sigma_2 u_2)^2 \\ &= \underbrace{(1+\rho)(\sigma_1^2 u_1^2 + \sigma_2^2 u_2^2) - \rho(\sigma_1 u_1 - \sigma_2 u_2)^2}_{\geq 0 \ \forall \rho \in [-1,1]} \ \square \end{aligned}$$

**Equation.** When $\rho = 0$ and $\sigma_1, \sigma_2 > 0$, then

$$f_{X_1,X_2}(x,x_2) = \prod_{k=1}^{2} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_k - \mu_k)^2}{2\sigma_k^2}\right)$$

So $X_1$ and $X_2$ are independent in this case.
More generally, suppose $(X_1, X_2)$ is a Gaussian vector.
want to find $\mathbb{E}[X_2|X_1]$.
Let $a \in \mathbb{R}$. Consider $X_2 - aX_1$.

$$\begin{aligned} \mathrm{Cov}(X_2 - aX_1, X_1) &= \mathrm{Cov}(X_2, X_1) - a\mathrm{Cov}(X_1, X_1) \\ &= \mathrm{Cov}(C_1, X_2) - a\mathrm{Var}(X_1) \\ &= \rho\sigma_1\sigma_2 - a\sigma_1^2 \end{aligned}$$

Take $a = (\rho\sigma_2)/\sigma_1$. Then $\mathrm{Cov}(X_2 - aX_1, X_1) = 0$.
Set $Y = X_2 - aX_1$

**Claim.** $(X_1, Y)$ is a Gaussian vector

**Proof.**

$$\begin{bmatrix} X_1 \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -a & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

So $X_1 Y$ is of the form $A \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ where $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is a Gaussian vector. $\square$

**Equation.** From the criterion of independence, we get $X_1$ is independent of $Y$, since $(X_1, Y)$ is Gaussian and $\mathrm{Cov}(X_1, Y) = 0$. Have $Y = X_2 - aX_1$ so we can express

$$X_2 = X_2 - aX_1 + aX_1 = Y + aX_1$$

and

$$\mathbb{E}[X_2|X_1] = \mathbb{E}[Y + aX_1|X_1] = \underbrace{\mathbb{E}[Y|X_1]}_{=\mathbb{E}[Y]} + \underbrace{a\mathbb{E}[X_1|X_1]}_{=X_1}$$

using $Y \perp\!\!\!\perp X_1$.
So

$$\mathbb{E}[X_2|X_1] = \mathbb{E}[Y] + aX_1$$

$(X_1, X_2)$ Gaussian, $(X_2 - aX_1, X_1)$ is Gaussian and $X_2 - aX_1 \perp\!\!\!\perp X_1$.

$$X_2 = X_2 - aX_1 + X_1$$

So given $X_1$,

$$X_2 \sim N(aX_1 + \mu_2 - a\mu_1, \mathrm{Var}(X_2 - aX_1)$$

where

$$\mathrm{Var}(X_2 - aX_1) = \mathrm{Var}(X_2) + a^2\mathrm{Var}(X_1) - 2a\mathrm{Cov}(X_1, X_2)$$

### 3.20 Multivariate CLT (non-examinable)

**Equation.** Let $X$ be a random vector in $R^k$ with $\mu = \mathbb{E}[X]$ and covariance matrix $\sigma$. Let $X_1, X_2, \ldots$ be iid with same distribution as $X$. Then

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_1 - \mathbb{E}[X_i]) \xrightarrow{(d)} N(\mu, \sigma) \text{ as } n \to \infty$$

Convergence distribution means that "$\forall$" $B \subseteq \mathbb{R}^k$

$$\mathbb{P}(S_n \in B) \underset{n\to\infty}{\to} \mathbb{P}(N(\mu, \Sigma) \in B)$$

**Example.** Let $U \sim U[0, 1]$. Set $X = -\log U$

$$\mathbb{P}(X \leq x) = \mathbb{P}(-\log U \leq x) = \mathbb{P}(U \geq e^{-x}) = 1 - e^{-x}$$

So $X \sim$ Exp(1)

**Theorem.** Let $X$ be a continuous r.v. with distribution function $F$. Then if $U \sim U[0, 1]$ we have that $F^{-1}(U) \sim F$

**Proof.** Set $Y = F^{-1}(U)$

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

### 3.21 Rejection Sampling

**Example.** Suppose $A \subset [0,1]^d$. Define

$$f(x) = \frac{1(x \in A)}{|A|}, \ |A| = \text{ volume of } A$$

Let $X$ have density $f$. How can we simulate $X$?
Let $(U_n)_{n \in N}$ be an iid sequence of $d$-dimensional uniforms, i.e.

$$U_n = (U_{k,n} : k \in \{1, \dots, d\}), \ (U_{k,n})_{(k,n)} \text{ iid } \sim U[0,1]$$

Let $N = \min\{n \geq 1 : U_n \in A\}$

**Claim.** $U_N \sim f$

**Proof.** We want to show that $\forall B \subseteq [0,1]^d$

$$\mathbb{P}(U_N \in B) = \int_B f(X) \, \mathrm{d}x$$

$$\begin{aligned}
\mathbb{P}(U_N \in B) &= \sum_{i=1}^{[n]\infty} \mathbb{P}(U_N \in B, N = n) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(U_n \in A \cap B, U_{n-1} \notin A, \dots, U_1 \notin A) \\
&\underset{U_i s \text{ indep}}{=} \sum_{n=1}^{\infty} \mathbb{P}(U_n \in A \cap B) \cdot \mathbb{P}(U_{n-1} \notin A) \dots \mathbb{P}(U_1 \notin A) \\
&= \sum_{n=1}^{\infty} |A \cap B| (1 - |A|)^{n-1} \\
&= \frac{|A \cap B|}{|A|}
\end{aligned}$$

and we have:

$$\frac{|A \cap B|}{|A|} = \int_A \frac{1(x \in B)}{|A|} \, \mathrm{d}x = \int_B f(x) \, \mathrm{d}x$$

**Example.** Suppose $f$ is a density on $[0,1]^{d-1}$ which is bounded, i.e.

$$\exists \lambda > 0 \text{ s.t. } f(x) \le \lambda \ \forall x \in [0,1]^{d-1}$$

Want to sample $X \sim f$.
Consider

$$A = \{(x_1, \dots, x_d) \in [0,1]^d : x_d \le f(x_1, \dots, x_{d-1})/\lambda\}$$

From the above we know how to generate a uniform random variable on $A$.
Let $Y = (X_1, \dots, X_d)$ be this r.v.
Set $X = (X_1, \dots, X_{d-1})$

**Claim.** $X \sim f$

**Proof.** We need to show that $\forall B \subseteq [0,1]^{d-1}$

$$\mathbb{P}(X \in B) = \int_B f(x) \, \mathrm{d}x$$

Have:

$$\mathbb{P}(X \in B) = \mathbb{P}((X_1, \dots, X_{d-1}) \in B) = \mathbb{P}((X_1, \dots, X_d) \in (B \times [0,1]) \cap A) = \frac{|(B \times [0,1]) \cap A|}{|A|}$$

as $Y$ is uniform on $A$

$$\begin{aligned} |(B \times [0,1]) \cap A| &= \int \cdots \int 1((x_1, \dots, x_d) \in B \times [0,1] \cap A) \, \mathrm{d}x_1 \dots \mathrm{d}x_d \\ &= \int \cdots \int 1((x_1, \dots, x_{d-1}) \in B) 1 \left( x_d \le \frac{f(X_1, \dots, x_{d-1}}{\lambda} \right) \mathrm{d}x_1 \dots \mathrm{d}x_{d-1} \\ &= \frac{1}{\lambda} \int_B f(x) \, \mathrm{d}x \end{aligned}$$

$$\begin{aligned} |A| &= \frac{1}{\lambda} \int_{[0,1]^{d-1}} f(x) \, \mathrm{d}x \\ &= \frac{1}{\lambda} \end{aligned}$$

So

$$\mathbb{P}(X \in B) = \int_B f(X) \, \mathrm{d}x$$