Probability Summary

Hasan Baig

Lent 2021

Contents

1 Probability Spaces

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

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

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

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

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

Remark. We talk about probability of events and not outcomes.

1.1 Combinatorial Analysis

Note.

$$\binom{n}{k} \text{ strictly increasing functions from set size } k \text{ to size } n$$

$$\binom{n+k-1}{k} \text{ increasing functions from set size } k \text{ to size } n$$

1.2 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.

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 to reach result

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

1.3 Properties of Probability Measures

1.3.1 Countable subadditivity

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

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

Proof. Define $B_1 = A_1$ and $B_n = A_n \backslash (A_1 \cup \dots \cup A_{n-1}) \, \forall n \ge 2$.
Then $(B_n)_{n \ge 1}$ is a disjoint sequence of events in $\mathcal{F}$ and $\bigcup\limits_{n \ge 1} B_n = \bigcup\limits_{n \ge 1} A_n$.
Then apply properties of probability measure

1.3.2 Continuity of Probability Measures

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

Claim.

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

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 $\bigcup_{k=1}^{n} B_k = A_n$ and $\bigcup_{k=1}^{n} B_k = \bigcup_{k=1}^{n} A_k$
Then use properties of probability measure.

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

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

1.4 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 $A_1, \ldots, A_n \in \mathcal{F}$. then

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

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

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

Notice

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

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

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

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

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

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

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

1.4.1 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 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

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

Proof. trivial

1.6 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. Apply above formula and use countable additivity

1.7 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.

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

Then use countable additivity

1.8 Bayes' Formula

Equation. Let (B_n) be a partition of Ω, 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

1.9 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.

2 Discrete Random Variables

2.1 Definitions and Examples

Definition (Discrete Probability Distribution)**.**

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

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

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

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

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

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.

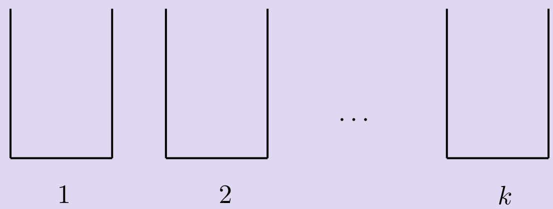
$$\mathbb{P}(\text{we see } k \text{ heads}) = \binom{N}{k} p^k (1 - p)^{n-k}$$

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

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

Example (Multinomial Distribution)**.**

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



k boxes and N balls

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

Throw the balls independently.

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

The set of ordered partitions of N.

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

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

$$\Omega = \{1, 2, \dots\}\ \mathbb{P}(\text{we tossed } k \text{ times until first } H) = (1-p)^{k-1} p = p_k$$

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

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

$$\sum_{k=0}^{\infty} p'_k = 1$$

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 λ.

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

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

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

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 λ.

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

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

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

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

Definition. A random variable X is called **discrete** if it takes values in a countable set.

Notation. Suppose X takes values in the countable set S. For every $x \in S$ we write
$$p_x = \mathbb{P}(X = x) = \mathbb{P}(\{\omega : X(\omega) = x\})$$
We call $(p_x)_{x \in S}$ the probability mass function of X (pmf) or the distribution of X.
If (p_x) is Bernoulli 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
$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \dots \mathbb{P}(X_n = x_n) \; x_n \in S_1, \dots, x_n \in S_n$$

2.2 Expectation

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

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

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

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

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

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

Proof. (Ω countable)

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

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 ≥ 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 $\text{Var}(X)$ is defined to be

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

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 $\sqrt{\text{Var}(X)}$ the standard deviation of X

Properties:

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

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

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

  Proof. Just expand out, use properties of expectation

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

  Proof. Just expand out RHS

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

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

"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 LHS

(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. Expand out

(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. Use remark, $\sum_{(x,y)}$

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.

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

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

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

Then take expectations

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. Use Markov on the random variable $Y = (X - \mathbb{E}[X])^2$ and 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

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

Take expectations and minimise f by taking $t = \mathbb{E}[XY]/\mathbb{E}[Y^2]$. Sub in and result immediate

2.5.4 Cases of Equality

Note. Equality in C-S occurs when

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

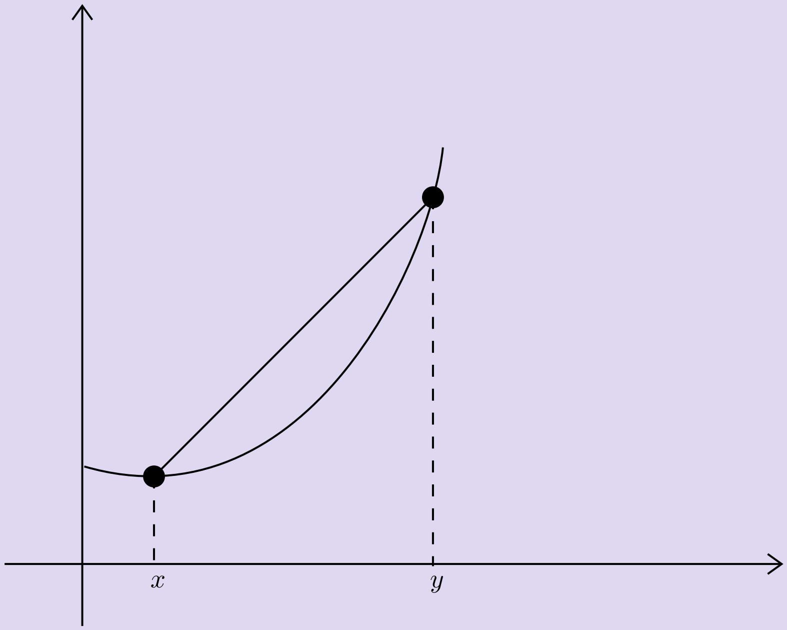
$$\mathbb{E}[(X - tY)^2] = 0 \implies \mathbb{P}(X = tY) = 1$$

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. Let $m \in \mathbb{R}$. Let $x < m < y$. Then $m = tx + (1-t)y$ for some $t \in [0,1]$. Use the definition of convex to get an inequality which leads to

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

Then let

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

and use above to get

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

Set $m = \mathbb{E}[X]$ and apply last inequality to X then take expectation to get result

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

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

where $b = f(\mathbb{E}[X]) - a\mathbb{E}[X]$ so

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

but

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

from before so this forces $f(X) = aX + b$
By assumption $f(\mathbb{E}[X]) = a\mathbb{E}[X] + b$ and $\forall x \neq \mathbb{E}[X]$ $f(x) > ax + b$
So this forces $X = \mathbb{E}[X]$ 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

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

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

Proof. Define X to be the r.v. taking values $\{x_1, \ldots, x_n\}$ all with equal prob
Apply Jensen's with $f(x) = -\log x$

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 (Ω_n) be a partition of Ω 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

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

and take expectations

2.6.2 Joint Distributions

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

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

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

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

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

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

$$\mathbb{P}(X+Y=z)=\sum_{y}\mathbb{P}(X=z-y)\cdot\mathbb{P}(Y=y)$$

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. If $X\sim \text{Poi}(\lambda)$ and $Y\sim \text{Poi}(\mu)$ independent then $X+Y\sim \text{Poi}(\lambda+\mu)$

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

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

$$\mathbb{E}[X|Y=y]=\sum_{x}x\mathbb{P}(X=x|Y=y)$$

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** for 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.

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

2.6.4 Properties of Conditional Expectation

Claim.

- $$\forall c \in \mathbb{R}\ \mathbb{E}[cX|Y] = c \cdot \mathbb{E}[X|Y] \text{ and } \mathbb{E}[c|Y] = c$$
- $X_1, \dots, X_n$ 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:

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

Taking expectation of both sides gives result

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.

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

Expanding the expectation gives result

Claim. Suppose Y and Z 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

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

2.7 Random Walks

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

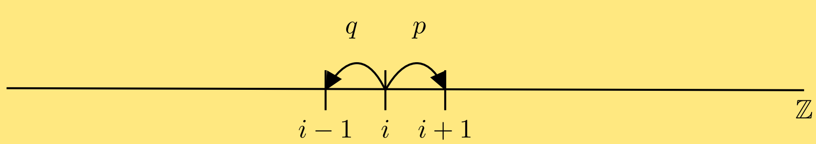
Definition. A **random walk** is a random process that can be expressed in the following way

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

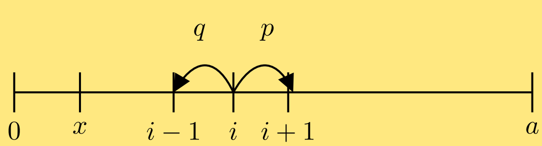
where (Y_i) are independent and identically distributed (iid) r.v.'s and x is a deterministic number (fixed).

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.

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

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,

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

- $p \neq q$:

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

  Solving this recurrence relation with boundary conditions yields:

  Equation.

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

  This is the Gambler's Ruin estimate.

2.7.1 Expected time to absorption

Equation. 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 to find

$$\mathbb{E}_x[T] = \tau_x$$

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

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

$$\tau_0 = \tau_a = 0$$

- Case $p = \frac{1}{2}$. Guessing quadratic solution and applying boundary conditions gives:

$$\tau_x = x(a - x)$$

- Case $p \neq \frac{1}{2}$. Guessing Cx particular integral and solving recurrence relation gives:

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

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 (trivial check)

Theorem. The pgf uniquely determines the distribution of X

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

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

Show $p_r = q_r \; \forall r$ by applying induction: cancelling same terms, dividing by power of z and taking limit to zero

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 $z < 1$)
Then just do analysis, considering the following:
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

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

Follow appropriate similar reasoning for $\mathbb{E}[X] = \infty$.

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

Theorem.

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

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

In particular

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

Moreover

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

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

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

Let

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

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)

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

$$p(z) = (pz + 1 - p)^n$$

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

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

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

$$\mathbb{E}[z^X] = e^{\lambda(z-1)}$$

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 $X_1, X_2, \dots$ 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

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

by working through the algebra

2.9.1 Another Proof Using Conditional Expectation

Method.

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

which leads to

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

So

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

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

Subbing in $z = 1-$ yields

Equation.

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

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 of offspring with distribution

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

Every individual in gen. 1 produces an indep. number of offspring with the same distribution.
Let $Y_{k,n} : k \geq 1, n \geq 0)$ be an iid sequence with distribution $(g_k)_{k \in \mathbb{N}}$
$Y_{k,n}$ 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.

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

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

(trivial to show)
So

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

Taking expectation and 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
$$\begin{aligned} G_{n+1}(z) &= G(G_n(z)) \\ &= G(G(\dots(G(z))\dots)) \\ &= G_n(G(z)) \end{aligned}$$

Proof. Condition on X_n as one would expect and we get:
$$\mathbb{E}[\mathbb{E}[z^{X_{n+1}}|X_n]] = \mathbb{E}[(G(z))^{X_n}] = G_n(G(z))$$

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

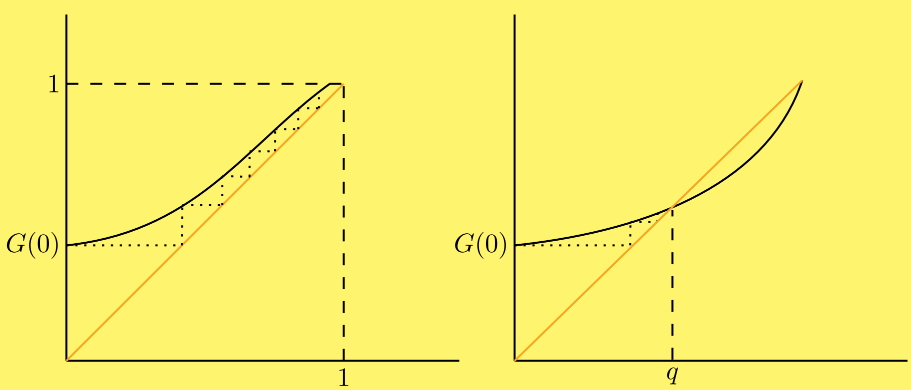
$$X_{n+1} = X_n^{(1)} + \cdots + X_n^{(m)}$$

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

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

Remark. So we have proved

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



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

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

In 2^{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

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

Proof (2^{nd} part)**.** Consider the function $H(z) = G(z) - z$
Have cases $\mathbb{P}(X_1 \leq 1) = 1$ or $\mathbb{P}(X_1 \leq 1) < 1$. The first is trivial. For the second case, think about the diagrams previous and how to use Rolle's theorem on H to show what we desire.

3 Continuous Random Variables

3.1 Definitions and Properties

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

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

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

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.

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

(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

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

and use that $\bigcap_n A_n = \varnothing$.
Left limits exist by the increasing property of F

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

Proof.

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

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)

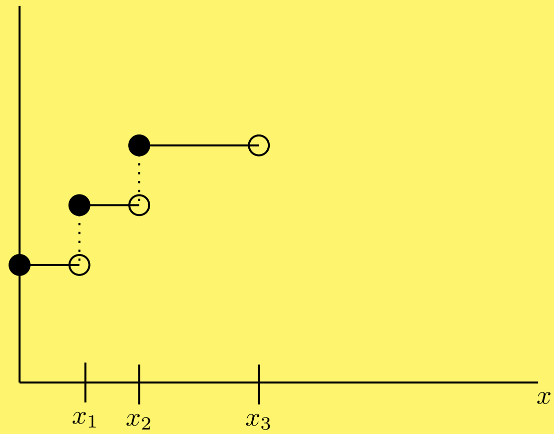
$$\lim_{x \to \infty} F(x) = 1$$

and

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

Proof. Exercise

Remark. 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, $\mathbb{P}(X = x) = 0\ \forall x \in \mathbb{R}$

Equation.

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

F differentiable so say it is absolutely continuous

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^{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^{nd}).

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

Taking expectation, we get

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

Example. Uniform distribution is defined as you expect, write $X \sim U[a,b]$

Example. Exponential distribution

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

$$F(x) = 1 - e^{-\lambda x}$$

and

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

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

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

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 ∞.
Then T has the memoryless property iff T is exponential

Proof. $\implies$:

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

Sub $t = 1$, then $t = m/n$. Then let $\mathbb{P}(t = 1) = e^{-\lambda}$ so we have proved that

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

And for $t \in \mathbb{R}^+$. We can bound $r \leq t < s$ with $r, s \in \mathbb{Q}^+$ and $|r - s| \leq \varepsilon$ then take limit

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. Treat increasing and decreasing cases separately

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

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

Can show expectation and variance are what we expect.
When X has density f, we write $X \sim N(\mu, \sigma^2)$
(X is normal with parameters μ and σ^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 $Y = g(X)$. We can show that $Y \sim N(a\mu + b, a^2\sigma^2)$ by considering density of Y
σ is the 'standard deviation'.
Suppose $X \sim N(\mu, \sigma^2)$, then

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

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

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

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 $X = (X_1, \dots, X_n)$ have 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) Apply definitions

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

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

Factorise this appropriately and let $B_j = \mathbb{R}$ for $j \neq i$ to get:

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

This shows that the density of X_i is

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

Then we can check independence

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

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

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.

$$\mathbb{P}(X + Y \leq z) = \int_{-\infty}^{z} \mathrm{d}y \left(\int_{-\infty}^{\infty} f_Y(y - x) f_X(x) \, \mathrm{d}x \right)$$

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

$$\det J_{ij} = \det \left(\frac{\partial x_i}{\partial y_j} \right)$$

Proof. We do not prove it here.

3.9 Order Statistics for a Random Sample

Equation. 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. We can show:

$$\mathbb{P}(Y_n \leq x) = (F(x))^n$$

$$f_{Y_n}(x) = n(F(x))^{n-1} \cdot f(x)$$

We can show the density of $Y_1, \dots, Y_n$ is:

$$f_{Y_1,\dots,Y_n}(x_1, \dots, x_n) = \begin{cases} n! f(x_1) \dots f(x_n) & \text{when } X_1 < X_2 < \dots X_n \\ 0 & \text{otherwise} \end{cases}$$

Equation. If $X_1, \dots, X_n$ are independent with $X_i \sim \text{Exp}(\lambda_i)$ then

$$\min(X_1, \dots, X_n) \sim \text{ Exp} \left(\sum_{i=1}^{n} \lambda_i \right)$$

Example. Let $X_1, \dots, X_n$ be iid $\text{Exp}(\lambda)$ and let Y_i be their order statistics

$$Z_1 = Y_1, \ Z_2 = Y_2 - Y_1, \dots, Z_n = Y_n - Y_{n-1}$$

So $Z_1, \dots, Z_n$ are independent and $Z_i \sim \text{ Exp}(\lambda(n - i + 1))$. We can show this by considering the bijection with the values of Y_i and applying a previous equation.

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 θ.

Theorem. Suppose the mgf is defined for an open interval of values of θ. 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 $X \sim \Gamma(n, \lambda)$
Check f is a density by showing integral over $\mathbb{R}$ is 1 (can use reduction $I_n = I_{n-1}$)

$$m(\theta) = \left(\frac{\lambda}{\lambda - \theta}\right)^n \text{ for } \lambda > 0$$

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 we can show

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

So by the uniqueness theorem we get $X + Y \sim \Gamma(n + m, \lambda)$.

Equation. In particular, this implies that if $X_1, \ldots, X_n$ 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 $X \sim N(\mu, \sigma^2)$

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

We can show that

$$m(\theta) = e^{\theta\mu + \theta^2\sigma^2/2}$$

by rewriting the integral in the form of constant times integral over a normal distribution.
If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$
So

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

Suppose $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\mu, \tau^2)$ and $X \perp\!\!\!\perp Y$
Then $X + Y \sim N(\mu + \nu, \sigma^2 + \tau^2)$ (we can show this by considering the mgfs)

Example. Cauchy distribution

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

$$m(\theta) = \infty \ \forall \theta \neq 0, \ (m(0) = 1)$$

Moral. 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 θ 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 θ, it uniquely determines the distribution of X. Also

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

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

$$m(\theta) = \prod_{i=1}^{n} \mathbb{E}[e^{\theta_i X_i}] \text{ iff } X_1, \dots, X_n \text{ are indep.}$$

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

$$m_n(\theta) \to m(\theta) \ \forall \theta \in \mathbb{R} \text{ where } m_n(\theta) = \mathbb{E}[e^{\theta X_n}] \text{ and } m(\theta) = \mathbb{E}[e^{\theta X}]$$

Then X_n converges to X in distribution

Note. This is just saying if the mgf's of the X_n converge to some mgf then $X_n \xrightarrow{d} X$

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 $(\sigma^2 = \text{Var}(X_1))$)**.**

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

then apply Chebyshev's inequality

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

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

Note.

$$\mathbb{P}(\forall \varepsilon > 0\ \exists n_0 : |X_n - X| < \varepsilon\ \forall n > n_0) = 1$$

Intuitively, 'pretty much all' events have $|X_n(\omega) - X(\omega)| < \varepsilon$ happening after a certain point. E.g. We can take X_n to be 1 if we have had a head after n tosses with our sample space being the set of sequences of tosses. $X(\omega) = 1$.

Claim. Suppose $X_n \to 0$ almost surely as $n \to \infty$. Then $X_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$

Proof. NTS:

$$\forall \varepsilon > 0\ \mathbb{P}(|X_n| > \varepsilon) \to 0 \text{ as } n \to \infty$$

We do this by considering

$$A_n = \bigcap_{m=n}^{\infty} \{|X_m| \leq \varepsilon\}$$

and then considering $\bigcup A_n$

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

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

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

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 enough:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \approx Z\ Z \sim N(0,1)$$

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

Proof. Consider $Y_i = (X_i - \mu)/\sigma$. 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.

$$\mathbb{E}[e^{\delta X_1}] < \infty \text{ and } \mathbb{E}[e^{-\delta X_1}] < \infty$$

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

Bound the series appropriately to show that it is $o(|\theta|^2)$ by showing it is $O(|\theta|^3)$
Then

$$m\left(\frac{\theta}{\sqrt{n}}\right) = 1 + \frac{\theta^2}{2n} + o\left(\frac{|\theta|^2}{n}\right)$$

and hence

$$\left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n \to e^{\theta^2/2} \text{ as } n \to \infty$$

3.14 Applications

Example. Normal approximation to the Binomial distribution:
Let $S_n \sim \text{ Bin}(n, p)$

$$S_n = \sum_{i=1}^{n} X_i,\ (X_i) \text{ iid } \sim \text{ Ber}(p)\ \mathbb{E}[S_n] = np, \text{Var}(S_n) = np(1-p)$$

and apply CLT to get

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

$$\text{Bin}\left(n, \frac{\lambda}{n}\right) \to \text{ Poi}(\lambda)\ \lambda > 0$$

Example. Normal approximation 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

$$\hat{p}_N = \frac{S_N}{N}$$

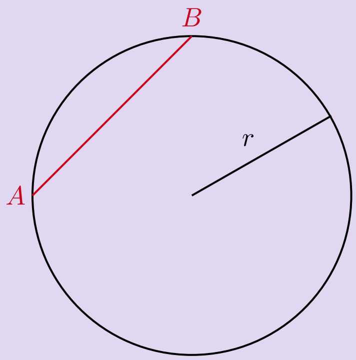
where S_N is the number of yes voters.
How large should N be so that

$$|\hat{p}_N - p| \leq \frac{4}{100} \text{ w.p. } \geq 0.99?$$

Apply CLT to get an approximate normal for S_N and use that

3.16 Bertrand's Paradox

Example.



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

Different interpretations of random lead to different answers

3.17 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 $X = (X_1, \ldots, X_n)^T$ with values in $\mathbb{R}^n$. Then X is a **Gaussian vector** or is just called **Gaussian** if $\forall u = (u_1, \ldots, u_n)^T \in \mathbb{R}^n$

$$u^T X = \sum_{i=1}^{n} u_i X_i \text{ is a Gaussian r.v. in } \mathbb{R}$$

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 Gaussian in $\mathbb{R}^m$.

Proof. Work with definition and set $v = A^T u$

Definition.

$$\mu = \mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix} \quad \mu_i = \mathbb{E}[X_i]$$

$$V = \mathrm{Var}(X) = \mathbb{E}[\underset{n\times 1}{(X-\mu)} \cdot \underset{1\times n}{(X-\mu)^T}] = \begin{bmatrix} \ddots & \vdots & \\ & \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)] & \\ & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} \ddots & \vdots & \\ & \mathrm{Cov}(X_i,X_j)] & \\ & \vdots & \ddots \end{bmatrix}$$

$$V_{ij} = \mathrm{Cov}(X_i, X_j)$$

Equation. We can show that:

$$\mathbb{E}[u^T X] = u^T \mu$$

$$\mathrm{Var}(u^T X) = u^T V u$$

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

$$\mathrm{Var}(u^T X) = u^T V u$$

Since $\mathrm{Var}(u^T X) \geq 0$, we have

$$u^T V u \geq 0 \ \square$$

Method. Finding mgf of X:

$$m(\lambda) = \mathbb{E}[e^{\lambda^T X}] \ \forall \lambda \in \mathbb{R}^n,\ \lambda = (\lambda_1, \dots, \lambda_n)^T$$

We know

$$\lambda^T X \sim N(\lambda^T \mu, \lambda^T V \lambda)$$

So $m(\lambda)$ is characterised by μ and V. Since the mgf uniquely characterises the distribution, we see that a Gaussian vector is uniquely characterised by its mean μ and variance V.

$$m(\lambda) = \mathbb{E}[e^{\lambda^T X}] = e^{\lambda^T \mu + \lambda^T V \lambda/2}$$

In this case we write $X \sim N(\mu, V)$

Claim. Let $Z_1, \ldots, Z_n$ iid $N(0,1)$ r.v.'s .
Set $Z = (Z_1, \ldots, Z_n)^T$. Then Z is a Gaussian vector.

Proof. We can show that $u^T Z \sim N(0, |u|^2)$ by considering the moment generating of Z.

$$\mathbb{E}[Z] = 0 \ \mathrm{Var}(Z) = I_n = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$$

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 μ and variance V using Z.
Let $V = U^T D U$ where D diagonal (possible as V symmetric). Then we set $\sigma = U^T \sqrt{D} U$ (diagonal entries in $\sqrt{D}$ are the root of those in D).
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.
Then we can easily check mean and variance are as desired

Method. Finding density of $X \sim N(\mu, V)$
In the case that V is positive definite:

$$f_X(x) = f_Z(z) \cdot |J| = \prod_{i=1}^{n} \left(\frac{e^{-z_i^2/2}}{\sqrt{2\pi}} \right) \cdot |\det \sigma^{-1}|$$

$$\implies f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det V}} e^{z^T z/2}$$

Subbing in for $z^t \cdot z$ gives:

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

In the case V is non-negative definite, some eigenvalues could be 0.
By an orthogonal change of basis, we can assume that

$$V = \begin{bmatrix} U & 0 \\ 0 & 0 \end{bmatrix} \text{ where } U \text{ is an } m \times m \ (m < n) \text{ positive definite matrix}$$

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^{st})**.** If V is diagonal, then the density $f_X(x)$ factorises into a product. Indeed,

$$(x-\mu)^T V^{-1}(x-\mu) = \sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{\lambda_i}$$

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^{nd})**.**

$$m(\theta) = \mathbb{E}[e^{\theta^T X}] = e^{\theta^T \mu + \theta^T V \theta/2} = e^{\sum \theta_i \mu_i} \cdot e^{\sum \theta_i^2 \lambda_i/2}$$

So $m(\theta)$ factorises into the mgf's of Gaussian r.v.'s in $\mathbb{R}$ $\square$

Moral. So for Gaussian vectors we have

$$(X_1, \ldots, X_n) \text{ are independent iff } \mathrm{Cov}(X_i, X_j) = 0 \text{ whenever } i \neq j$$

3.18 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. (Consider definition of Cov)$\square$

$$V = \mathrm{Var}(X) = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

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. Show $u^T V u \geq 0$ for all $u \in \mathbb{R}^2$

Method. Suppose (X_1, X_2) is a Gaussian vector. We want to find $\mathbb{E}[X_2|X_1]$.
Let $a \in \mathbb{R}$. Consider $X_2 - aX_1$.

$$\begin{aligned} \mathrm{Cov}(X_2 - aX_1, X_1) &= \mathrm{Cov}(X_2, X_1) - a\mathrm{Cov}(X_1, X_1) \\ &= \mathrm{Cov}(X_1, X_2) - a\mathrm{Var}(X_1) \\ &= \rho\sigma_1\sigma_2 - a\sigma_1^2 \end{aligned}$$

Take $a = (\rho\sigma_2)/\sigma_1$. Then $\mathrm{Cov}(X_2 - aX_1, X_1) = 0$.
Set

$$Y = X_2 - aX_1$$

$\begin{bmatrix} X_1 \\ Y \end{bmatrix}$ is a Gaussian vector as it is of the form $A\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$

From the criterion of independence, we get X_1 is independent of Y, since (X_1, Y) is Gaussian and $\mathrm{Cov}(X_1, Y) = 0$.

$$\mathbb{E}[X_2|X_1] = \mathbb{E}[Y + aX_1|X_1] = \mathbb{E}[Y] + aX_1$$

as $X_2 = X_2 - aX_1 + aX_1$. So given X_1,

$$X_2 \sim N(aX_1 + \mu_2 - a\mu_1, \mathrm{Var}(X_2 - aX_1))$$

where

$$\mathrm{Var}(X_2 - aX_1) = \mathrm{Var}(X_2) + a^2\mathrm{Var}(X_1) - 2a\mathrm{Cov}(X_1, X_2)$$

3.19 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 $\forall B \subseteq [0,1]^d$

$$\mathbb{P}(U_N \in B) = \int_B f(X) \, \mathrm{d}x$$

$$\begin{aligned} \mathbb{P}(U_N \in B) &= \sum_{n=1}^{\infty} \mathbb{P}(U_N \in B, N = n) \\ &= \frac{|A \cap B|}{|A|} \end{aligned}$$

by working out sum

$$\frac{|A \cap B|}{|A|} = \int_A \frac{1(x \in B)}{|A|} \, \mathrm{d}x = \int_B f(x) \, \mathrm{d}x$$

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

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

Moral. In the case $d = 3$, imagine surface in 3-D where the z value is the probability. We are using uniform distributions to sample uniformly within a volume bounded by our surface which, in turn, gives (x, y) with desired probability.