Probability

Shing Tak Lam*

April 13, 2021

This document is intended for revision purposes. As a result, it does not contain any exposition. This is based off lectures given by Dr Perla Sousi in Lent 2021, but the order of content, as well as some of the proofs have been modified after the fact, primarily to provide simpler proofs for theorems. Note that this also contains theorems from examples sheets, as some are useful elsewhere.

Probability is on *Paper 2*.

Contents

*stl45@cam.ac.uk

1 Probability Spaces

Definition (σ-algebra)**.** Let Ω be a set and $\mathcal{F}$ be a collection of subsets of Ω. $\mathcal{F}$ is a σ-algebra if

- $\Omega \in \mathcal{F}$
- If $A \in \mathcal{F}$ then $A^{\complement} \in \mathcal{F}$
- If $(A_n)_{n\in\mathbb{N}} \in \mathcal{F}$, then we must have $\bigcup\limits_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

Remark. When Ω is countable, we take $\mathcal{F} = \mathcal{P}(\Omega)$.

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

- $\mathbb{P}(Omega) = 1$
- If $(A_n)_{n\in\mathbb{N}} \in \mathcal{F}$ are (pairwise) disjoint, then $\mathbb{P}\left(\bigcup\limits_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$

Definition (Probability Space)**.** We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.

Definition (Outcomes)**.** The elements of Ω are called outcomes.

Definition (Events)**.** The elements of $\mathcal{F}$ are called events.

Proposition.

- $\mathbb{P}\left(A^{\complement}\right) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(\emptyset) = 0$
- *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 Properties of Probability Measures

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

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

Proof. Define $B_1 = A_1$, $B_2 = A_2 \backslash A_1$, $B_n = A_n \backslash (A_1 \cup \cdots \cup A_{n-1})$. Then (B_n) is a sequence of disjoint events in $\mathcal{F}$, and $\bigcup\limits_{n\in\mathbb{N}} A_n = \bigcup\limits_{n\in\mathbb{N}} B_n$. By countable additivity, $\mathbb{P}\left(\bigcup\limits_{n\in\mathbb{N}} A_n\right) = \mathbb{P}\left(\bigcup\limits_{n\in\mathbb{N}} B_n\right) \sum\limits_{n\in\mathbb{N}} \mathbb{P}(B_n)$.

But $B_n \subseteq A_n$, so $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$, as a result

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

□

Proposition (Continuity)**.** *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of increasing ($\leq = \subseteq$) events in $\mathcal{F}$. Then $\mathbb{P}(A_n)$ is increasing and bounded above, so it converges. In addition,*

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

Proof. Let B_n be defined as above. Then $\bigcup\limits_{k=1}^{n} B_k = A_n$. Hence

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

as $n \to \infty$. As $\bigcup\limits_{n=1}^{\infty} A_n = \bigcup\limits_{n=1}^{\infty} B_n$, and $\mathbb{P}\left(\bigcup\limits_{n=1}^{\infty} B_n\right) = \sum\limits_{n=1}^{\infty} \mathbb{P}(B_n)$, we get the required result. □

Proposition (Inclusion-Exclusion)**.** *Let $A_1, \ldots, A_n \in \mathcal{F}$. Then*

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

Proof. By induction. $n = 1$ is trivial. In addition, we have already seen the case for $n = 2$. Now suppose it holds for $n - 1$ events. Then

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

Now let $B_i = A_i \cap A_n$. Then

$$\mathbb{P}((A_1 \cup \cdots \cup A_{n-1}) \cap A_n) = \mathbb{P}(B_1 \cup \cdots \cup B_{n-1})$$

By the inductive hypothesis, we have that

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

and

$$\begin{aligned}\mathbb{P}(B_1 \cup \cdots \cup B_{n-1}) &= \sum_{k=1}^{n-1}(-1)^k \sum_{1\le i_1<\cdots<i_k\le n} \mathbb{P}\left(B_{i_1} \cap \cdots \cap B_{i_k}\right) \\ &= \sum_{k=1}^{n-1}(-1)^k \sum_{1\le i_1<\cdots<i_k\le n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k} \cap A_n\right)\end{aligned}$$

Plugging these into the original expression yields the desired result. □

Proposition (Bonferroni Inequalities)**.** *If $r < n$ and r is odd, then*

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) \le \sum_{k=1}^{r}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \mathbb{P}\left(A_{i_1} \cup \cdots \cup A_{i_k}\right)$$

If r is even, then

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) \ge \sum_{k=1}^{r}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \mathbb{P}\left(A_{i_1} \cup \cdots \cup A_{i_k}\right)$$

Proof. By induction. $n = 2$ is trivial. Suppose this holds for $n-1$ events. Suppose further than r is odd. Then

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}(B_1 \cup \cdots \cup B_{n-1}). \tag{$*$}$$

where $B_i = A_i \cap A_n$. By applying the inductive hypothesis and as r is odd,

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

and as $r-1$ is even,

$$\mathbb{P}(B_1 \cup \cdots \cup B_{n-1}) \ge \sum_{k=1}^{r-1}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \mathbb{P}\left(B_{i_1} \cap \cdots \cap B_{i_k}\right)$$

Substitute these into $(*)$ to get the required result. The even case can be proven similarly. □

1.2 Independence

Definition (Independence)**.** Let $A, B \in \mathcal{F}$. We say that A and B are independent if

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

Definition (Independence)**.** A countable collection of events $(A_n)_{n\in\mathbb{N}}$ is said to be countable if for all distinct $i_1, \ldots, i_k$, we have that

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

1.3 Conditional Probability

Definition (Conditional Probability)**.** Let $B \in \mathcal{F}$, $\mathbb{P}(B) > 0$. Let $A \in \mathcal{F}$, we define the conditional probability of A given B as

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

Proposition. *If A and B are independent, then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.*

Proposition. *Suppose (A_n) is a disjoint sequence of events. Then*

$$\mathbb{P}\left(\bigcup_n A_n \mid B\right) = \sum_n \mathbb{P}(A_n \mid B)$$

Proof.

$$\mathbb{P}\left(\bigcup_n A_n \mid B\right) = \frac{\mathbb{P}\left((\bigcup_n A_n) \cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_n (A_n \cap B)\right)}{\mathbb{P}(B)} = \frac{\sum_n \mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_n \mathbb{P}(A_n \mid B)$$

□

Proposition (Law of Total Probability)**.** *Suppose (B_n) is a disjoint sequence of events such that $\bigcup_n B_n = \Omega$ and for all n, $\mathbb{P}(B_n) > 0$. Let $A \in \mathcal{F}$. Then*

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

Proof.

$$\mathbb{P}(()\,A) = \mathbb{P}(A \cap \Omega) = \mathbb{P}\left(A \cap \left(\bigcup_n B_n\right)\right) = \mathbb{P}\left(\bigcup_n (A \cap B_n)\right) = \sum_n \mathbb{P}(A \cap B_n) = \sum_n \mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)$$

□

Proposition (Bayes' Formula)**.** *Suppose (B_n) is a disjoint sequence of events such that $\bigcup_n B_n = \Omega$ and for all n, $\mathbb{P}(B_n) > 0$. Then*

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

Proof.

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

□

2 Stirling's Formula

Definition (Asymptotic Equivalence)**.** We say $f \sim g$, or f is asymptotically equivalent to g if

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1$$

Theorem (Stirling)**.**

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

Lemma.

$$\log(n!) \sim n \log n$$

Proof of Lemma. Define $l_n = \log(n!) = \log 2 + \cdots + \log n$. We have that $\log \lfloor x \rfloor \leq \log x \leq \log \lfloor x+1 \rfloor$. Integrating from 1 to n,

$$\int_1^n \log \lfloor x \rfloor \, \mathrm{d}x = \sum_{k=1}^{n-1} \log k = l_{n-1}$$

So

$$l_{n-1} \leq \int\limits_1^n \log x \mathrm{d}x \leq l_n$$

Thus $l_{n-1} \leq n \log n - n + 1 \leq l_n$ and $n \log n - n + 1 \leq l_n \leq (n+1)\log(n+1) - (n+1) + 1$. Dividing through by $n \log n$, we get that

$$1 - \frac{n+1}{n \log n} \leq \frac{l_n}{n \log n} \leq \frac{(n+1)\log(n+1) - n}{n \log n}$$

So $\dfrac{l_n}{n \log n} \to 1$ as $n \to \infty$. □

Proof of Stirling. Is non-examinable and omitted. See Lecture Notes or Analysis I Examples Sheet 4. □

3 Discrete Probability Distributions

Definition (Discrete Probability Distribution)**.** Let Ω be finite or countable, $\mathcal{F} = \mathcal{P}(\Omega)$. Let $\Omega = \{\omega_1, \ldots\}$. Then knowing $\mathbb{P}(\{\omega_i\})$ for all i gives us the probability for any event. Let $p_i = \mathbb{P}(\{\omega_i\})$.

Definition (Bernoulli Distribution)**.** For parameter $p \in [0,1]$, we have the Bernoulli Distribution $\mathrm{Ber}(p)$, where:
Let $\Omega = \{0, 1\}$. Then $p_1 = 0$, $p_0 = 1 - p$.

Definition (Binomial Distribution)**.** For parameters $n \in \mathbb{Z}^+$, $p \in [0,1]$, we have the Binomial Distribution $\mathrm{Bin}(n, p)$, where:
Let $\Omega = \{0, \ldots, n\}$. Then $p_k = \binom{n}{k} p^k (1-p)^{n-k}$.

Definition (Multinomial Distribution)**.** For parameters $p_1, \ldots, p_k \in [0,1]$, $n \in \mathbb{Z}^+$, we have the Multinomial Distribution $M(n, p_1, \ldots, p_k)$, where
Let $\Omega = \{(n_1, \ldots, n_k) \in \mathbb{N}^k : n_1 + \cdots + n_k = n\}$. Then

$$\mathbb{P}(\{(n_1, \ldots, n_k)\}) = \binom{n}{n_1, \ldots, n_k} p_1^{n_1} \ldots p_k^{n_k}$$

where $\dbinom{n}{n_1, \ldots, n_k} = \dfrac{n!}{n_1! \ldots n_k!}$

Definition (Geometric Distribution)**.** For parameter p, we have the Geometric Distribution $\mathrm{Geo}(p)$, where:
Let $\Omega = \mathbb{N} = \{1, \ldots\}$. Then $p_k = p(1-p)^{k-1}$.

Definition (Poisson Distribution)**.** For parameter λ, we have the Poisson Distribution $\mathrm{Poi}(p)$, where:
Let $\Omega = \{0, \ldots\}$. Then $p_k = e^{-\lambda} \dfrac{\lambda^k}{k!}$

4 Random Variables

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

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

Remark. We use $\{X \in A\} = \{\omega : X(\omega) \in A\}$ as a shorthand.

Definition (Indicator). For $A \in \mathcal{F}$, define $1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$

Definition (Probability Distribution Function). For a random variable X, define the probability distribution function $F_X : \mathbb{R} \to [0, 1]$ by

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

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

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

5 Discrete Random Variables

Definition (Discrete Random Variable). A random variable X is discrete if it takes values in a countable set.

Definition (Probability Mass Function). For $x \in S$, we define $p_x = \mathbb{P}(X = x)$ to be the probability mass function.

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

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

for all $x_1 \in S_1, \dots, x_n \in S_n$.

5.1 Expectation

Definition (Expectation for Nonnegative Random Variables). For a discrete random variable X, define the expectation

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

Proposition.

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

Proof.

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

□

Definition (Expectation for General Random Variables). Let X be a discrete random variable. Define $X_+ = \max(X, 0)$ and $X_- = \max -X, 0$. Then $X = X_+ - X_-$ and $|X| = X_+ + X_-$. The both $\mathbb{E}[X_+]$ and $\mathbb{E}[X_-]$ are well defined. If at least one is finite, then we define

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

Proposition.

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

Definition (Integrable)**.** If $\mathbb{E}[|X|] < \infty$, then X is integrable.

Proposition. *If $X \geq 0$, then $\mathbb{E}[X] \geq 0$.*

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

Proposition. *For $c \in \mathbb{R}$, $\mathbb{E}[cX] = c\mathbb{E}[X]$ and $\mathbb{E}[X + c] = \mathbb{E}[X] + c$.*

Proposition. *For X, Y integrable, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.*

Proposition. *For $c_1, \ldots, c_n \in \mathbb{R}$, $X_1, \ldots, X_n$ integrable random variables,*

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

Proposition. *Suppose $X_1, \ldots$ are nonnegative random variables. Then*

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

Proof.

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

□

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

Proposition. *For $g : \mathbb{R} \to \mathbb{R}$, we define $g(X)$ to be the random variable such that $g(X)(\omega) = g(X(\omega))$. Then*

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

Proof. Let $Y = g(X)$. Then $\mathbb{E}[Y] = \sum_{y \in Y(\Omega)} y\mathbb{P}(Y = y)$. Now $Y = y \iff x \in g^{-1}(\{y\})$. Hence

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

□

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

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

Definition (Moment)**.** For $r \in \mathbb{N}$, we call $\mathbb{E}[X^r]$ then r-th moment of X.

5.2 Variance

Definition (Variance). We define the variance of X, $\mathrm{Var}(X)$ by

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

Definition (Standard Deviation). $\sigma = \sqrt{\mathrm{Var}(X)}$

Proposition. $\mathrm{Var}(X) \geq 0$, *and* $\mathrm{Var}(X) = 0 \iff \mathbb{P}(X = \mathbb{E}[X]) = 1$.

Proposition. $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$ *and* $\mathrm{Var}(X + c) = \mathrm{Var}(X)$.

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

Proposition. $\mathrm{Var}(X) = \min\{\mathbb{E}\left[(X - c)^2\right] : c \in \mathbb{R}\}$

Definition (Covariance). Let X,Y be random variables, we define the covariance of X and Y as

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

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

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

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

Proposition. $\mathrm{Cov}(cX, Y) = c\,\mathrm{Cov}(X, Y)$ *and* $\mathrm{Cov}(X + c, Y) = \mathrm{Cov}(X, Y)$

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

Proposition. *For* $c_1, \ldots, c_n, d_1, \ldots, d_n \in \mathbb{R}$, *and* $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ *random variables,*

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

Proposition.

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

Proposition. *If* X, Y *are independent random variables, then*

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

Proof.

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

□

Proposition. *If* X *and* Y *are independent, then* $\mathrm{Cov}(X, Y) = 0$.

5.3 Joint Distribution and Convolution

Definition (Joint Distribution)**.** Let $X_1, \ldots, X_n$ be random variables. The joint distribution is defined to be

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

Proposition.

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

Definition (Marginal Distribution)**.** We call $\mathbb{P}(X_i = x_i)$ the marginal distribution of X_i.

Definition (Conditional Distribution)**.** The conditional distribution of X given $Y = y$ is defined to be

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

Proposition. *If X and Y are independent, then*

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

Definition (Convolution)**.** Suppose X, Y are independent random variables. Then

$$\mathbb{P}(X + Y = z) = \sum_y \mathbb{P}(X = z - y)\,\mathbb{P}(Y = y)$$

5.4 Conditional Expectation

Definition (Conditional Expectation (Event))**.** Let $B \in \mathcal{F}$, $\mathbb{P}(B) > 0$ and X be a random variable. Then

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

Proposition (Law of Total Expectation)**.** *Suppose $X \geq 0$, (Ω_n) is a partition of Ω into disjoint events. Then*

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

Proof. $X = X \cdot 1(\Omega) = \sum_n X \cdot 1(\Omega_n)$. Taking expectations yields the required result. □

Definition (Conditional Expectation (Random Variable = Value))**.** Let X, Y be random variables. The conditional expectation of X given $Y = y$ is

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

Definition (Conditional Expectation (Random Variable))**.** Let X, Y be random variables. Let $g(y) = \mathbb{E}[X \mid Y = y]$. We define the conditional expectation of X given Y as

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

Proposition. $\mathbb{E}[cX \mid Y] = c\mathbb{E}[X \mid Y]$

Proposition.

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

Proposition.

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

Proposition. *If X and Y are independent, then $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$.*

Proposition. *Suppose Y and Z are independent. Then $\mathbb{E}[\mathbb{E}[X \mid Y] \mid Z] = \mathbb{E}[X]$.*

Proposition. *Suppose $h : \mathbb{R} \to \mathbb{R}$. Then $\mathbb{E}[h(Y) \cdot X \mid Y] = h(Y)\mathbb{E}[X \mid Y]$.*

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

5.5 Probability Generating Functions

Definition (Probability Generating Function)**.** Let X be a random variable taking values in $\mathbb{N}$. Let $p_r = \mathbb{P}(X = r)$. The probability generating function is defined to be

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

Proposition. *For $|z| < 1$, the pgf is absolutely convergent.*

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

Proof. Suppose (p_r), (q_r) are two pgfs of with

$$\sum_{r=0}^{\infty} p_r z^r = \sum_{r=0}^{\infty} q_r z^r$$

for all $|z| < 1$.
Setting $z = 0$, we get that $p_0 = q_0$. Suppose $p_r = q_r$ for all $r \leq n$. Then

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

Dividing by z^{n+1} and taking $z \to 0$, we get that $p_{n+1} = q_{n+1}$. By strong induction we are done. □

Theorem.

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

Proof. First we assume that $\mathbb{E}[X] < \infty$. In Analysis I, we have seen that within the radius of convergence, we can differentiate a power series term by term. So

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

For $0 < z < 1$, we have that $p'(z)$ is an increasing function. So we have that

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

Given $\varepsilon > 0$, there exists N such that

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

Then we have that as $z > 0$,

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

So for all $\varepsilon > 0$,

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

Now suppose if $\mathbb{E}[X] = \infty$. Then for any M, we have some N such that

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

Then from above,

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

so $\lim\limits_{z \uparrow 1} p'(z) = \infty = \mathbb{E}[x]$. □

Theorem.

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

Proposition.

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

Proposition.

$$\mathbb{P}(X = n) = \frac{1}{n!} p^{(n)}(0)$$

Proposition. *If $X_1, \ldots, X_n$ are independent random variables with pgfs $q_1, \ldots, q_n$, then if $X = X_1 + \cdots + X_n$ and the pgf of X is p, we have that*

$$p(z) = q_1(z) \ldots q_n(z)$$

Proposition. *If $X \sim \mathrm{Bin}(n, p)$, then*

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

Proposition. *If $X \sim \mathrm{Geo}(p)$, then*

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

Remark. We are using $\mathrm{Geo}(p)$ to represent the number of trials including the success.

Proposition. *If $X \sim \mathrm{Poi}(\lambda)$, then*

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

Example. Let (X_i) are iid with pgf p, $S_n = X_1 + \cdots + X_n$, N independent random variable with pgf q. Then

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

We can also use conditional expectation, since

$$\mathbb{E}\left[z^{S_n}\right] = \mathbb{E}\left[\mathbb{E}\left[z^{X_1+\cdots+X_N|N}\right]\right]$$

We have that

$$\mathbb{E}\left[z^{X_1+\cdots+X_N} \mid N = n\right] = (p(z))^n$$

as a result,

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

6 Inequalities

6.1 Markov's Inequality

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

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

Proof. Observe that $X \geq a \cdot 1(X \geq a)$. Taking expectations, we get that

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

□

6.2 Chebyshev's Inequality

Proposition (Chebyshev's Inequality). *If X is a random variable with $\mathbb{E}[X] < \infty$, then for all $a > 0$,*

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

Proof.

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

□

6.3 Cauchy-Schwarz Inequality

Proposition (Cauchy-Schwarz Inequality). *Let X and Y be random variables. Then*

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

Proof. Without loss of generality, we may assume that $\mathbb{E}\left[X^2\right], \mathbb{E}\left[Y^2\right] < \infty$ and $X, Y \geq 0$. As $XY \leq \frac{1}{2}(X^2 + Y^2)$, we must also have that $\mathbb{E}[XY] < \infty$.

We may assume $\mathbb{E}\left[X^2\right], \mathbb{E}\left[Y^2\right] > 0$, as otherwise the result is trivial. Let $t \in \mathbb{R}$, we have that

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

Minimising for t, we find that the minimum occurs when $t = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$. Result follows. □

Proposition. *Equality holds in Cauchy Schwarz if and only if $\mathbb{P}(X = tY) = 1$.*

6.4 Jensen's Inequality

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

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

Lemma. *Let $f : \mathbb{R} \to \mathbb{R}$ be convex. Then f is the supremums of the lines below it. That is,*

$$\forall m \in \mathbb{R}, \exists a, b \in \mathbb{R}, f(m) = am + b \wedge \forall x, f(x) \geq ax + b$$

Proof. Let $m \in \mathbb{R}$, choose $x < m < y$. Then $m = tx + (1-t)y$. Therefore $f(m) \leq tf(x) + (1-t)f(y)$. So $t(f(m) - f(x)) \leq (1-t)(f(y) - f(m))$. This implies that

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

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

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

so $f(x) \geq a(x - m) + f(m)$ for all x. □

Proposition (Jensen's Inequality)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function, let X be a random variable, then*

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

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

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

□

Proposition. *Equality holds if and only if $\mathbb{P}(X = \mathbb{E}[X]) = 1$.*

Proposition. *Let f be a convex function and $x_1, \dots, x_n \in \mathbb{R}$. Then*

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

Proof. Define random variable X taking values $x_1, \dots, x_n$ with equal probability. Result follows from Jensen. □

6.5 AM-GM Inequality

Proposition. *For $x_1, \dots, x_n \geq 0$,*

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

Proof. Use $f(x) = \log x$ in proposition above. □

7 Random Walks

Definition (Random Process)**.** A random (stochastic) process is a sequence of random variables (X_n)

Definition (Random Walk)**.** A random walk is a random process where $X_n = x + Y_1 + \cdots + Y_n$, where x is a constant, (Y_i) are iid random variables.

Definition (Simple Random Walk on $\mathbb{Z}$)**.** We define the simple random walk on $\mathbb{Z}$ by $\mathbb{P}(() Y_i = 1) = p$, $\mathbb{P}(Y_i = -1) = 1 - p = q$.

Definition (Conditional Probability Measure)**.** We define $\mathbb{P}_x(\cdot) = \mathbb{P}(\cdot \mid X_0 = x)$.

Definition.

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

Proposition.

- $h(0) = 0$
- $h(a) = 1$
- *For* $0 < x < a$, $h(x) = ph(x+1) + qh(x-1)$

Proposition. *If* $p = q = 0.5$, *then* $h(x) = \frac{x}{a}$.

Proposition (Gabler's Ruin Estimate)**.** *If* $p \neq q$, *then*

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

Definition (Expected Time of Absorption)**.**

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

Definition.

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

Proposition.

- $\tau_0 = \tau_a = 0$
- *For* $0 < x < a$, $\tau_x = p\tau_{x+1} + q\tau_{x-1} + 1$

Proposition. *If* $p = q = 0.5$, *then*

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

Proposition. *If* $p \neq q$, *then*

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

8 Branching Processes

Let X_n represent the number of individuals in generation n. We take $X_0 = 1$. The individual in generation 0 produces a random number of offspring, with distribution $g_k = \mathbb{P}(X_1 = k)$. Each new individual produces offspring with the same distribution.

Let $(Y_{n,k} : n \geq 0, k \geq 1)$ be an iid sequence of random variables, with distribution (g_k). $Y_{n,k}$ represents the number of offspring of the k-th individual in generation n. Then

$$X_{n+1} = \begin{cases} Y_{n,1} + \cdots + Y_{n,X_n} & \text{if } X_n > 0 \\ 0 & \text{if } X_n = 0 \end{cases}$$

Theorem. *For all $n \geq 1$,*

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

Proof.

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

so

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

□

Theorem. *Let $G(z) = \mathbb{E}\left[z^{X_1}\right]$, and $G_n(z) = \mathbb{E}\left[Z^{X_n}\right]$. Then $G_{n+1}(z) = G(G_n(z)) = G_n(G(z))$.*

Proof.

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

So

$$G_{n+1}(z) = \mathbb{E}\left[Z^{X_{n+1}}\right] = \mathbb{E}\left[\mathbb{E}\left[z^{X_{n+1}} \mid X_n\right]\right] = \mathbb{E}\left[(G(z))^{X_n}\right] = G_n(G(z))$$

□

8.1 Extinction Probability

Definition (Exctinction Probability)**.** Define the extinction probability $q = \mathbb{P}(X_n = 0 \text{ for some } n \geq 1)$.

Proposition. *Let $q_n = \mathbb{P}(X_n = 0)$. Then $q_n \to q$.*

Proof. Let $A_n = \{X_n = 0\}$. Then $A_n \subseteq A_{n+1}$. So (A_n) is an increasing sequence. By continuity of the probability measure,

$$q_n = \mathbb{P}(A_n) \to \mathbb{P}\left(\bigcup_n A_n\right) = q$$

□

Proposition. $q_{n+1} = G(q_n)$*, and* $q = G(q)$*.*

Proof.

$$q_{n+1} = G_{n+1}(0) = G(G_n(0)) = G(q_n)$$

From the continuity of G we have that $q = G(q)$. □

Theorem. *Assume $\mathbb{P}(X_1 = 1) < 1$. Then q is the minimum nonnegative solution to $t = G(t)$.*

Proof. Let t be the minimum nonnegative solution to $t = G(t)$. $q_0 = 0 \leq t$. Now suppose $q_n \leq t$. Then as G is increasing, $q_{n+1} = G(q_n) \leq G(t) = t$. So $q_n \leq t$ for all n. Then as $q_n \to q$, $G(q) = q$, we must have that $t = q$. □

Proposition. *$q < 1$ if and only if $\mathbb{E}[X_1] > 1$.*

Proof. Omitted. □

9 Continuous Random Variables

Definition (Probability Distribution Function)**.** For a random variable X, we define the probability distribution function $F : \mathbb{R} \to [0, 1]$ with $F(x) = \mathbb{P}(X \leq x)$.

Proposition. *F is increasing.*

Proposition. *If $a < b$, then $\mathbb{P}(a \leq X \leq b) = F(b) - F(a)$.*

Proposition. *F is right continous. That is, $\lim_{y \downarrow x} F(y) = F(x)$.*

Proposition. *Left limits for F always exist.*

Proposition. $\lim\limits_{x \to -\infty} F(x) = 0$ *and* $\lim\limits_{x \to \infty} F(x) = 1$.

Definition (Continuous Random Variable)**.** A random variable X is continuous if the distribution function F is continuous.

From now on, we will assume that F is differentiable.

Definition (Probability Density Function)**.** For a random variable X with distribution F, we define the probability density function $f = F'$.

Proposition. $\int\limits_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$ *and* $\int\limits_{-\infty}^{x} f(t)\mathrm{d}t = F(x)$.

9.1 Expectation

Definition (Expectation for Nonnegative Random Variable)**.** For a nonnegative random variable X with density f, we define the expectation

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

Proposition. *Suppose $g(x) \geq 0$ for all x. Then*

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

Definition (Expectation of General Random Variables)**.** Let X be a random variable. Define $X_+ = \max(X, 0)$ and $X_- = \max(-X, 0)$. If at least one of $\mathbb{E}[X_+]$ and $\mathbb{E}[X_-]$ are finite, then we define

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

Proposition.

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

Proposition.

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

Proposition. *If $X \geq 0$, then*

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

Proof.

$$\mathbb{E}[X] = \int_0^\infty x f(x)\mathrm{d}x = \int_0^\infty \int_0^x \mathrm{d}y f(x)\mathrm{d}x = \int_0^\infty \mathrm{d}y \int_y^\infty f(x)\mathrm{d}x = \int_0^\infty \mathrm{d}y(1 - F(y)) = \int_0^\infty \mathbb{P}(X \geq y)\, \mathrm{d}y$$

□

9.2 Distributions

Definition (Uniform Distribution)**.** Let $a < b$, we say $X \sim U[a, b]$ if X has density

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

Proposition. *If $X \sim U[a, b]$, then*

$$F(x) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a \leq x \leq b \\ 1 & b \leq x \end{cases}$$

Proposition. *If $X \sim U[a, b]$, then $\mathbb{E}[X] = \dfrac{a+b}{2}$*

Definition (Exponential Distribution)**.** Let $\lambda \in \mathbb{R}$, $\lambda > 0$. We say $X \sim \mathrm{Exp}(\lambda)$ if X has density

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

Proposition. *If $X \sim \mathrm{Exp}(\lambda)$, then*

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

Proposition. *If $X \sim \mathrm{Exp}(\lambda)$, then $\mathbb{E}[X] = 1 - e^{-\lambda x}$*

Proposition (Memoryless Property)**.** *Let T be a positive random variable, no identically zero or ∞. Then T is memoryless, that is $\forall t, s, \mathbb{P}(T > t + s) = \mathbb{P}(T > t)\mathbb{P}(T > s)$ if and only if T is exponential.*

Proof. If is clear. Suffices to show only if. Suppose $\forall t, s, \mathbb{P}(T > t + s) = \mathbb{P}(T > t)\mathbb{P}(T > s)$. Let $g(t) = \mathbb{P}(T > t)$. Then $g(t + s) = g(t)g(s)$ for all t, s.

Inductively, for $m \in \mathbb{N}$, $g(m) = (g(1))^m$, and $g(\frac{m}{n}) = (g(1))^{\frac{m}{n}}$. As T is not identically zero or ∞, we must have that $g(1) \in (0, 1)$. Let $\lambda = -\log \mathbb{P}(T > 1) > 0$. Then $g(t) = e^{-\lambda t}$ for all $t \in \mathbb{Q}$, $t > 0$.

Now let $t \in \mathbb{R}$. Then there exists $r, s \in \mathbb{Q}$ such that $r < t < s$ and $|r - s| < \varepsilon$. As the distribution function is increasing, we have that

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

Letting $\varepsilon \to 0$ we get the desired result. □

Definition (Normal Distribution)**.** Given $\mu \in \mathbb{R}$, $\sigma > 0$, we say $X \sim N(\mu, \sigma^2)$ if X has density

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

Proposition. *If $X \sim N(\mu, \sigma^2)$, then $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$.*

Definition (Standard Normal)**.** We define the standard normal $Z \sim N(0,1)$ which has density

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

Definition (Φ)**.** Φ is defined to be the distribution function of $Z \sim N(0,1)$.

Proposition. *If $X \sim N(\mu, \sigma)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$*

Proposition. *If $X \sim N(\mu, \sigma)$, then $\dfrac{X-\mu}{\sigma} \sim N(0,1)$*

Definition (*Gamma Distribution*)**.** Given $\alpha, \lambda > 0$, we say $X \sim \Gamma(\alpha, \lambda)$ if X has density

$$f(x) = \frac{e^{-\lambda x}\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)}$$

Proposition. $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$.

9.3 Transformations

Theorem. *Let X be a continuous random variable with density f. Let g be a strictly monotone continuous function with difdferentiable inverse g^{-1}. Then $g(X)$ is a continuous random variable with density*

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

Proof. Suppose g is increasing. Then $\mathbb{P}\left(g(X) \le x\right) = \mathbb{P}\left(X \le g^{-1}(X)\right) = F(g^{-1}(x))$. Now suppose g is decreasing. Then $\mathbb{P}\left(g(X) \le x\right) = \mathbb{P}\left(X \ge g^{-1}X\right) = 1 - F(g^{-1}(x))$. Differentiating both expressions yields the result. □

9.4 Moment Generating Functions

Definition (Moment Generating Function)**.** Let X be a random variable with density f. The mgf of X is

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

Theorem. *The mgf uniquely determines the distribution of a random variable provided it is defined for an open interval of values of θ. (That is, it converges for some $\theta \neq 0$).*

Theorem. *Suppose the mgf is defined on an open interval of θ. Then*

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

Proposition. *If $X \sim \Gamma(n, \lambda)$, then $m(\theta) = \left(\dfrac{\lambda}{\lambda - \theta}\right)^n$ for $\theta < \lambda$.*

Corollary. *If $X \sim \mathrm{Exp}(\lambda)$, then $m(\theta) = \dfrac{\lambda}{\lambda-\theta}$ for $\theta < \lambda$.*

Proposition. *If $X_1, \ldots, X_n$ are independent with mgfs $m_1, \ldots, m_n$, then*

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

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

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

Proof. Note

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

and result follows. □

10 Multivariate Density Functions

Definition (Density)**.** Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random variable. We say X has density f if

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

Proposition.

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

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

10.1 Independence

Definition (Independence)**.** We say $X_1, \dots, X_n$ are independent if for all $x_1, \dots, x_n \in \mathbb{R}$,

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

Theorem. *Let $X = (X_1, \dots, X_n)$ have density f. Suppose $X_1, \dots, X_n$ are independent have have densities $f_1, \dots, f_n$. Then $f(x_1, \dots, x_n) = f_1(x_1) \dots f_n(x_n)$.*

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

$$\int_{-\infty}^{x_1} f_1(y_1) \mathrm{d}y_1 \cdots \int_{-\infty}^{x_n} f_n(y_n) \mathrm{d}y_n = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_1(y_1) \dots f_n(y_n) \mathrm{d}y_n \dots \mathrm{d}y_1$$

we get the result required. ☐

Theorem. *Suppose $X = (X_1, \dots, X_n)$ has density f, and f factorises into $f(x_1, \dots, x_n) = f_1(x_1) \dots f_n(x_n)$, then $X_1, \dots, X_n$ are independent, and have densities proportional to $f_1, \dots, f_n$.*

Proof. As f is a density, we must have that

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

In addition $\mathbb{P}(X_i \le x_i) = \mathbb{P}\left(X_i \le x_i, X_j \in (-\infty, \infty) \text{ for all } j \ne i\right)$, and

$$\mathbb{P}\left(X_i \le x_i, X_j \in (-\infty, \infty) \text{ for all } j \ne i\right) = \int_{-\infty}^{x_i} f_i(y) \mathrm{d}y \prod_{i \ne j} \int_{-\infty}^{\infty} f_j(y) \mathrm{d}y = \frac{\int_{-\infty}^{x_i} f_i(y) \mathrm{d}y}{\int_{-\infty}^{\infty} f_i(y) \mathrm{d}y}$$

Hence the density of X_i is $\dfrac{f_i}{\int_{-\infty}^{\infty} f_i(y) \mathrm{d}y}$. Independence follows from the fact that f factorises. ☐

Definition (Marginal Density)**.** For $X = (X_1, \dots, X_n)$ with density f, we define the marginal density (for X_1) as

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

10.2 Convolution

Definition (Convolution)**.** If f and g are densities, then we define the convolution of f and g as

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

Proposition. $f * g = g * f$

Proposition. *If X, Y are independent random variables with densities f_X, f_Y respectively, then $X+Y$ has density $f_X * f_Y$*

Proof.

$$\begin{aligned}
\mathbb{P}(X+Y \leq z) &= \iint_{\{x+y\leq z\}} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_X(x)f_Y(y)\mathrm{d}y\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_Y(y-x)f_X(x)\mathrm{d}y\mathrm{d}x \\
&= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_Y(y-x)f_X(x)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{z} f_X * f_Y(y)\mathrm{d}y
\end{aligned}$$

□

10.3 Conditional Density

Definition (Conditional Density)**.** Let X, Y be continuous random variables, with joint density $f_{X,Y}$ and marginal densities f_X, f_Y. The conditional density of X given $Y=y$ is

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

Proposition (Law of Total Probability)**.**

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

Definition (Conditional Expectation)**.** Let $g(y) = \int\limits_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\mathrm{d}x$. Then we define the conditional expectation of X given Y to be

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

10.4 Transformations

Theorem. *Let X be a random variable with values in $D \subseteq \mathbb{R}^d$, and with density f_X. Let $g : D \to g(D)$ be a bijection with a continuous derivative, with $\det\big(g'(x)\big) \neq 0$ for all $x \in D$.*

Then $Y = g(X)$ has density

$$f_Y(y) = f_X(x)|J|$$

where $x = g^{-1}(y)$ and $J = \det\left(\frac{\partial \mathbf{x}}{\partial y_1} \;\middle|\; \dots \;\middle|\; \frac{\partial \mathbf{x}}{\partial y_d}\right)$ is the Jacobian.

Proof. Omitted. □

10.5 Order Statistics for a Random Sample

Definition (Order Statistics)**.** Suppose $X_1, \dots, X_n$ are iid random variables with distribution F and density f. Let $Y_1 \leq \dots \leq Y_n$ be X_n in increasing order. Then (Y_i) are the order statistics.

Proposition. $\mathbb{P}(Y_1 \leq x) = 1 - (1 - F(x))^n$

Proof.

$$\mathbb{P}(Y_1 \le x) = 1 - \mathbb{P}(x < Y_1) = 1 - \mathbb{P}(x < \min X_1, \ldots, X_n) = 1 - (1 - F(x))^n$$

□

Proposition. $\mathbb{P}(Y_n \le x) = (F(x))^n$.

Proposition.

$$f_{Y_1,\ldots,Y_n}(x_1, \ldots, x_n) = \begin{cases} n! f(x_1) \ldots f(x_n) & \text{if } x_1 \le \cdots \le x_n \\ 0 & \text{otherwise} \end{cases}$$

Proof.

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

□

Proposition. *The Y_i are not independent.*

Proof. $f_{Y_1,\ldots,Y_n}(x_1, \ldots, x_n) = n! f(x_1) \ldots f(x_n) \cdot 1(x_1 \le \cdots \le x_n)$ so the density does not factorise. □

Example (Order Statistics of iid Exponentially Distributed Random Variables)**.** Let $X_1, \ldots, X_n$ be iid $\mathrm{Exp}(\lambda)$. Let Y_i be the order statistics. Define $Z_1 = Y_1$, $Z_2 = Y_2 - Y_1, \ldots, Z_n = Y_n - Y_{n-1}$. Then

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

and if $z = Ay$, then $y_j = \sum_{i=1}^{j} z_i$. In addition, $|J| = 1$. So

$$\begin{aligned} f_{Z_1,\ldots,Z_n}(z_1, \ldots, z_n) &= f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)|J| \\ &= n! f(y_1) \ldots f(y_n) \\ &= n! \lambda^n \exp(-\lambda(y_1 + \cdots + y_n)) \\ &= n! \lambda^n \exp(-\lambda(n z_1 + \cdots + z_n)) \\ &= \prod_{i=1}^{n} (n - i + 1) \lambda \exp(-\lambda(n - i + 1) z_i) \end{aligned}$$

So the Z_i are independent, and $Z_i \sim \mathrm{Exp}(\lambda(n - i + 1))$. Note this only holds because of the memoryless property of the exponential distribution.

10.6 Multivariate Moment Generating Functions

Definition (Moment Generating Function)**.** Suppose $X = (X_1, \ldots, X_n)$ is a random variable in $\mathbb{R}^n$. Then the mgf of X is defined to be

$$m(\theta) = \mathbb{E}\left[e^{\theta^T X}\right] = \mathbb{E}\left[e^{\theta_1 X_1 + \cdots + \theta_n X_n}\right]$$

where $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$

Theorem. *If the mgf is defined for a range of θ, then it uniquely determines the distribution of X.*

Proposition.

$$\frac{\partial^n m}{\partial \theta_i^n}(0) = \mathbb{E}\left[X_i^r\right]$$

Proposition.

$$\frac{\partial^{r+s} m}{\partial \theta_i^r \partial \theta_j^s}(0) = \mathbb{E}\left[X_i^r X_j^s\right]$$

11 Limit Theorems

11.1 Convergence of Random Variables

Definition (Convergence in Distribution)**.** Let (X_n) be a sequence of random variables. Let X be a random variable. We say that X_n converges to X in distribution, that is $X_n \xrightarrow{d} X$ if for all continuity points x of F_X,

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

Theorem (Convergence of mgfs)**.** *Let (X_n) be a sequence of random variables with mgfs (m_n), and suppose X is a random variable with mgf m. If for all $\theta \in \mathbb{R}$, $m_n(\theta) \to m(\theta)$, then $X_n \xrightarrow{d} X$.*

Definition (Convergence in Probability)**.** Let (X_n) be a sequence of random variables. (X_n) converges to X in probability, that is $X_n \xrightarrow{\mathbb{P}} X$ if for all $\varepsilon > 0$,

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

as $n \to \infty$.

Definition (Almost Sure Convergence)**.** (X_n) converges to X with probability 1, or almost surely (a.s.), that is $X_n \to X$ a.s. if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = \mathbb{P}\left(\forall \varepsilon > 0, \exists n_0, \forall n \geq n_0, |X_n - X| < \varepsilon\right) = 1$$

Proposition.

$$X_n \to X \text{ a.s.} \implies X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{d} X$$

Proposition. *Suppose $X_n \to 0$ a.s.. Then $X_n \xrightarrow{\mathbb{P}} 0$.*

Proof. Suffices to show that $\forall \varepsilon > 0, \mathbb{P}\left(|X_n| \leq \varepsilon\right) \to 1$ as $n \to \infty$. Let $A_n = \bigcap_{m=n}^{\infty} \{|X_m| \leq \varepsilon\}$. Then $\mathbb{P}\left(|X_n| \leq \varepsilon\right) \geq \mathbb{P}\left(A_n\right)$. Then

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n| \leq \varepsilon\right) \geq \lim_{n\to\infty} \mathbb{P}\left(A_n\right) = \mathbb{P}\left(\bigcup_n A_n\right) \geq \mathbb{P}\left(\lim_{n\to\infty} X_n = 0\right) = 1$$

□

11.2 Laws of Large Numbers

Theorem (Weak Law of Large Numbers)**.** *Let (X_n) be a sequence of iid random variables with $\mu = \mathbb{E}[X_1]$. Let $S_n = X_1 + \cdots + X_n$. Then as $n \to \infty$,*

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$$

Proof. Assume further $\sigma^2 = \text{Var}(X) < \infty$. Then using Chebyshev's Inequality we have that

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

□

Theorem (Strong Law of Large Numbers)**.** *Suppose further that* $\mathbb{E}[X_1] = \mu$ *finite. Then as* $n \to \infty$,

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

Proof. Omitted. □

11.3 Central Limit Theorem

Theorem. *Let* (X_n) *be a sequence of iid random variables with* $\mathbb{E}[X_1] = \mu$ *and* $\mathrm{Var}(X_1) = \sigma^2$ *both finite. Let* $S_n = X_1 + \cdots + X_n$, *and* $Z_n = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$, *then*

$$Z_n \xrightarrow{d} Z \sim N(0, 1)$$

Proof. Consider $Y_i = \dfrac{X_i - \mu}{\sigma}$. Thus without loss of generality, we may assume that $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = 1$. Assume further that there exists $\delta > 0$ such that $\mathbb{E}\left[e^{\delta X_i}\right]$ and $\mathbb{E}\left[e^{-\delta X_i}\right]$ are both finite.

By convergence of mgfs, suffices to show that for all $\theta \in \mathbb{R}$, as $n \to \infty$,

$$\mathbb{E}\left[e^{\theta \frac{S_n}{\sqrt{n}}}\right] \to \mathbb{E}\left[e^{\theta Z}\right] = \exp\left(\frac{\theta^2}{2}\right)$$

Let $m(\theta) = \mathbb{E}\left[e^{\theta X_1}\right]$. Then $\mathbb{E}\left[e^{\theta \frac{S_n}{\sqrt{n}}}\right] = \left(\mathbb{E}\left[e^{\frac{\theta}{\sqrt{n}} X_1}\right]\right)^n = \left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n$. Therefore, we need to show that $\left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n \to \exp\left(\frac{\theta^2}{2}\right)$ as $n \to \infty$. Suffices to show that $m(\theta) = 1 + \dfrac{\theta^2}{n} + o(\theta^2)$.

We now let $|\theta| < \dfrac{\delta}{2}$. Then

$$\begin{aligned}
\left|\mathbb{E}\left[\sum_{k\geq 3} \frac{X_1^k \theta^k}{k!}\right]\right| &\leq \mathbb{E}\left[\sum_{k\geq 3} \frac{|X_1|^k |\theta|^k}{k!}\right] \\
&\leq \mathbb{E}\left[|\theta X_1|^3 \exp(|X_1 \theta|)\right] \\
&\leq \mathbb{E}\left[|\theta X_1|^3 \exp\left(\frac{\delta}{2}|X_1|\right)\right]
\end{aligned}$$

Now $|\theta X_1|^3 \exp\left(\frac{\delta}{2}|X_1|\right) = |\theta|^3 \frac{(\frac{\delta}{2}|X_1|)^3}{3!} \frac{3!}{(\frac{\delta}{2})^3} \exp\left(\frac{\delta}{2}|X_1|\right) \leq |\theta|^3 \frac{3!}{(\frac{\delta}{2})^3} \exp(\delta|X_1|) = 3! \left(\frac{2|\theta|}{\delta}\right)^3 \exp(\delta|X_1|)$. Thus

$$\begin{aligned}
\left|\mathbb{E}\left[\sum_{k\geq 3} \frac{X_1^k \theta^k}{k!}\right]\right| &\leq 3! \left(\frac{2|\theta|}{\delta}\right)^3 \mathbb{E}\left[\exp(\delta|X_1|)\right] \\
&\leq 3! \left(\frac{2|\theta|}{\delta}\right)^3 \mathbb{E}\left[\exp(\delta X_1) + \exp(-\delta X_1)\right] \\
&= o(\theta^2)
\end{aligned}$$

□

11.4 Approximations

Proposition (Poisson approximation to Binomial)**.** *As* $n \to \infty$, $\mathrm{Bin}(n, \frac{\lambda}{n}) \to \mathrm{Poi}(\lambda)$.

Proof. Suppose $X \sim \text{Bin}(n, \frac{\lambda}{n})$. Let $p = \lambda n$, then as $n \to \infty$,

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} \frac{n!}{n^k (n-k)!} \left(1 - \frac{\lambda}{n}\right)^{n-k} \to \frac{\lambda^k}{k!} \cdot 1 \cdot e^{-\lambda} = e^{-\lambda} \frac{\lambda^k}{k!}$$

□

Proposition (Normal Approximation to Binomial). *Suppose $S_n \sim \text{Bin}(n, p)$. Then as $n \to \infty$, $S_n \approx N(np, np(1-p))$.*

Proof. If $S_n \sim \text{Bin}(n, p)$, then $S_n = X_1 + \cdots + X_n$, where X_i are iid $\text{Ber}(p)$. So by the CLT as $n \to \infty$,

$$\frac{S_n - np}{np(1-p)} \xrightarrow{d} N(0, 1)$$

□

Proposition (Normal Approximation to Poisson). *If $S_n \sim \text{Poi}(n)$. Then as $n \to \infty$, $S_n \approx N(n, n)$.*

Proof. If $S_n \sum \text{Poi}(n)$, then $S_n = X_1 + \cdots + X_n$, where X_i are iid $\text{Poi}(1)$. □

12 Multidimensional Gaussian Random Variables

Definition (Gaussian Vector). Let $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$. X is a Gaussian Vector if for all $u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$ $u^T X$ is Gaussian.

Definition (Expected Value). If $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$, we define the expected value of X as

$$\mu = \mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

Definition (Variance Matrix). If $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$, we define the variance of X as

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

This is a $n \times n$ matrix.

Proposition. *If X is Gaussian, then $u^T X \sim N(u^T \mu, u^T V u)$*

Proposition. *V is symmetric.*

Proposition. *V is nonnegative definite. That is, for all $u \in \mathbb{R}^n$, $u^T V u \geq 0$.*

Proof. $u^T V u = \text{Var}(u^T X) \geq 0$. □

Proposition (mgf of Gaussian Vector).

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

Remark. As the mgf uniquely characterises the distribution of a random variable, a Gaussian vector X is uniquely characterised by the mean μ and variance V. As a result, we write

$$X \sum N(\mu, V)$$

12.1 Construction of Gaussian Vectors

Proposition. *Let $Z_1, \dot{,} Z_n$ be iid $N(0,1)$. Then $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$ is a Gaussian vector. Furthermore, $Z \sim N(0, I)$.*

Definition (Square Root of Matrix)**.** If V is a nonnegative definite symmetric matrix, then for an orthogonal matrix U and diagonal matrix D, we have that $V = U^T D U$, where $D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$. We define the square root σ of V to be

$$\sigma = U^T \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n} \end{pmatrix} U$$

Proposition. $\sigma^2 = V$.

Proposition. $X = \mu + \sigma Z \sim N(\mu, V)$.

12.2 Density of a Multivariate Gaussian

Proposition. *If V is positive definite, and $X \sim N(\mu, V)$, then*

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

Proof. As V is positive definite, σ is invertible. Let $z = \sigma^{-1}(x-\mu)$. Then

$$\begin{aligned} f_X(x) = f_Z(z)|J| &= \left(\prod_{i=1}^{n} \frac{\exp\left(-\frac{z_i^2}{2}\right)}{\sqrt{2\pi}}\right) \left|\det \sigma^{-1}\right| \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{|z|^2}{2}\right) \frac{1}{\sqrt{\det V}} \\ &= \frac{1}{\sqrt{(2\pi)^n \det V}} \exp\left(-\frac{(x-\mu)^T V^{-1}(x-\mu)}{2}\right) \end{aligned}$$

□

Proposition. *If V is nonnegative definite, then by an orthogonal change of basis, $V = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix}$ and $\mu = \begin{pmatrix} \lambda \\ \nu \end{pmatrix}$, where U is a $m \times m$ matrix, $\lambda \in \mathbb{R}^m$, $\nu \in \mathbb{R}^{n-m}$. Then we can write $X = \begin{pmatrix} Y \\ \nu \end{pmatrix}$ where Y has density*

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^m \det U}} \exp\left(-\frac{(y-\lambda)^T U^{-1}(y-\lambda)}{2}\right)$$

12.3 Independence

Proposition. *If the X_i's are independent, then V is diagonal.*

Proof. For $i \neq j$, $V_{ij} = \mathrm{Cov}(X_i, X_j) = 0$. □

Proposition. *If V is diagonal, then the X_i's are independent.*

Proof. If $V = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$, then $(x-\mu)^T V^{-1}(x-\mu) = \sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{\lambda_i}$ and the density factorises. □

Alternative Proof. Similarly we can show that the mgf factorises. □

12.4 Bivariate Gaussian

This subsection contains information about the special case of $n = 2$. Let (X_1, X_2) be a Gaussian vector in $\mathbb{R}^2$, with mean (μ_1, μ_2) and variance $(\sigma_1^2 \sigma_2^2)$.

Definition (Correlation Coefficient)**.** For random variables X_1, X_2, we define

$$\rho = \operatorname{Corr}(X_1, X_2) = \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}}$$

Proposition. $\rho \in [0, 1]$

Proof. Cauchy-Schwarz □

Proposition. $V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

Proposition. *V is nonnegative definite for $\rho \in [0, 1]$.*

Proposition. *If $\rho = 0$, then X_1 and X_2 are independent.*

Proposition. *Given X_1, $X_2 \sim N(aX_1 + \mu_2 - a\mu_1, \operatorname{Var}(X_2 - aX_1))$, where $\operatorname{Var}(X_2 - aX_1) = \operatorname{Var}(X_2) + a^2 \operatorname{Var}(X_1) - 2a\operatorname{Cov}(X_1, X_2)$ and $a = \frac{\rho\sigma_2}{\sigma_1}$.*

Proof. Let $Y = X_2 - aX_1$. Then $\operatorname{Cov}(Y, X_1) = 0$. Also, (Y, X_1) is a Gaussian vector, so Y and X_1 are independent. Now $X_2 = Y + aX_1$, and $\mathbb{E}[X_2 \mid X_1] = \mathbb{E}[Y] + aX_1$. Result follows. □

13 Sampling

Theorem. *Let X be a continuous random variable with distribution function F. Then if $U \sim U[0, 1]$, we have that $F^{-1}(U) \sim F$.*

Proof. Let $Y = F^{-1}(U)$. Then

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

□

Definition (Box-Muller Transform)**.** Let $X, Y \sim N(0, 1)$ be independent random variables. If we let $X = R\cos\Theta$ and $Y = R\sin\Theta$, then we find that R and Θ are independent, with densities

$$f_R(r) = \begin{cases} re^{-r^2/2} & r \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi} & r \in [0, 2\pi) \\ 0 & \text{otherwise} \end{cases}$$

By computing the distributions, we can find that if $U_1, U_2 \sim U[0, 1]$ and are independent, then setting $\Theta = 2\pi U_1$ and $R = \sqrt{-2\log U_2}$ we can generate a random bivariate Gaussian.

Definition (Rejection Sampling)**.** Let $A \subseteq [0,1]^d$ be a subset with non-zero volume $|A|$. Define $f(x) = \dfrac{1(x \in A)}{|A|}$. Let $X \sim f$. Then X is uniformly distributed on A.

Let (U_n) be an iid sequence of uniform random variables, that is

$$U_n = (U_{k,n} : k = 1, \dots, d)$$

where $(U_{k,n}) \sim U[0,1]$ iid. Let $N = \min\{n \geq 1 : U_n \in A\}$. We claim that U_N has density f.

Proof. Suffices to show that for any $B \subseteq [0,1]^d$, $\mathbb{P}(U_n \in B) = \int_B f(x)\mathrm{d}x$.

$$\begin{aligned}
\mathbb{P}(U_N \in B) &= \sum_{n=1}^{\infty} \mathbb{P}(U_n \in B, N = n) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(U_n \in A \cap B, U_{n-1} \notin A, \dots, U_1 \notin A) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(U_n \in A \cap B)\, \mathbb{P}(U_{n-1} \notin A) \dots \mathbb{P}(U_1 \notin A) \\
&= \sum_{n=1}^{\infty} |A \cap B|(1 - |A|)^{n-1} \\
&= \frac{|A \cap B|}{|A|} \\
&= \int_A \frac{1(x \in B)}{|A|}\mathrm{d}x \\
&= \int_B f(x)\mathrm{d}x
\end{aligned}$$

□

Definition (Rejction Sampling)**.** Now suppose f is a density supported on $[0,1]^{d-1}$ which is bounded. Say $f(x) \leq \lambda$. Then consider

$$A = \left\{ (x_1, \dots, x_d) \in [0,1]^d : x_d \leq \frac{f(x_1, \dots, x_{d-1})}{\lambda} \right\}$$

Let $Y = (X_1, \dots, X_d)$ be a uniform random variable on A, generated as above. Let $X = (X_1, \dots, X_{d-1})$. We claim that X has density f.

Proof. Suffices to show that for any $B \subseteq [0,1]^d$, $\mathbb{P}(X \in B) = \int_B f(x)\mathrm{d}x$.

$$\begin{aligned}
\mathbb{P}(X \in B) &= \mathbb{P}((X_1, \dots, X_{d-1}) \in B) \\
&= \mathbb{P}((X_1, \dots, X_d) \in (B \times [0,1]) \cap A) \\
&= \frac{|(B \times [0,1]) \cap A|}{|A|}
\end{aligned}$$

then

$$\begin{aligned}
|(B \times [0,1]) \cap A| &= \int \cdots \int 1((x_1, \dots, x_d) \in (B \times [0,1] \cap A))\mathrm{d}x_1 \dots \mathrm{d}x_d \\
&= \int \cdots \int 1((x_1, \dots, x_{d-1}) \in B) \cdot 1\left(x_d \leq \frac{f(x_1, \dots, x_{d-1})}{\lambda}\right) \mathrm{d}x_1 \dots \mathrm{d}x_d \\
&= \int \cdots \int 1((x_1 \dots x_{d-1}) \in B) \frac{f(x_1, \dots, x_{d-1})}{\lambda} \mathrm{d}x_1 \dots \mathrm{d}x_{d-1} \\
&= \frac{1}{\lambda} \int_B f(x)\mathrm{d}x
\end{aligned}$$

Furthermore,

$$|A| = \frac{1}{\lambda} \int_{[0,1]^{d-1}} f(x)\mathrm{d}x = \frac{1}{\lambda}$$

and the result follows. □

A Common Distributions

A.1 Discrete Distributions

name	parameters	pmf	mean	variance	pgf
Bernoulli	$p \in [0,1]$	$p^k(1-p)^{1-k}$	p	$p(1-p)$	$q + pz$
Binomial	$n \in \mathbb{N}$, $p \in [0,1]$	$\binom{n}{k} p^k(1-p)^{n-k}$	np	$np(1-p)$	$(q+pz)^n$
Geometric[1]	$p \in [0,1]$	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pz}{1-qz}$
Poisson	$\lambda > 0$	$e^{-\lambda}\frac{\lambda^k}{k!}$	λ	λ	$e^{\lambda(z-1)}$

A.2 Continuous Distributions

name	parameters	pdf	cdf	mean	variance	mgf
Uniform	$a < b$	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{b\theta} - e^{a\theta}}{b-a}$
Normal	$\mu \in \mathbb{R}$, $\sigma > 0$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$	/	μ	σ^2	$\exp\left(\theta\mu + \frac{\theta^2\sigma^2}{2}\right)$
Exponential	$\lambda > 0$	$\lambda e^{-\lambda x}$	$1 - e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - \theta}$
Gamma	$\lambda > 0$, $n \in \mathbb{N}$	$\frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}$	/	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-\theta}\right)^n$

A.3 Multivariate Distributions

name	parameters	pdf	mean	variance
Multivariate Normal	μ, V	$\frac{1}{\sqrt{(2\pi)^n \det V}} \exp\big(-(x-\mu)^T V^{-1}(x-\mu)\big)$	μ	V