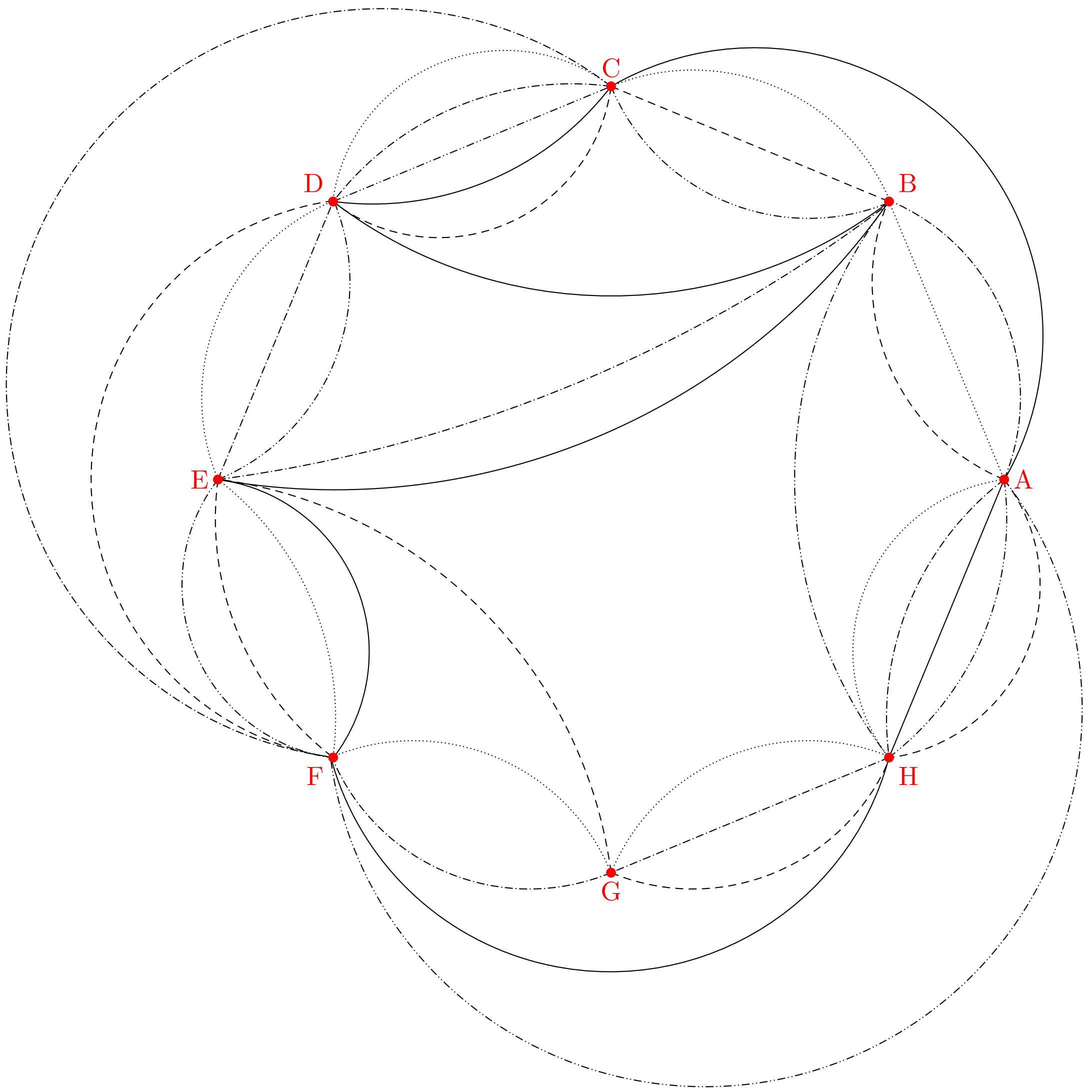
MATH40005A Probability

Lectured by Professor Almut Veraart
Typed by Yu Coughlin

Autumn 2023



A minimum vertex 5-Venn diagram.

Syllabus

Basic set theory and combinatorics, Kolmogorov probability axioms, Conditional probability and independence, DRVs, CRVs, Transformations, Expectaction, Variance, Multivariate calculus, Joint Distributions, LOTUS, pgfs, mgfs, conditional distribution

Contents

0 Introduction

Lecture 1
Monday
30/10/2023

The following are complementary reading for the course.

- G. Grimmett and D. J. A. Welsh, Probability: An Introduction, 1986
- J. K. Blitzstein and J. Hwang, Introduction to Probability, 2019
- D. F. Anderson et al, Introduction to Probability, 2018
- S. M. Ross, Introduction to Pro ability Models, 2014
- G. Grimmett and D. Stirzaker, Probability and Random Processes, 2001
- G. Grimmett and D. Stirzaker, One Thousand Exercises in Probability, 2009

These notes are written assuming all material covered in ICL's **MATH40001**, **MATH40002**, **MATH40003A** and **MATH40004**.

Notation. Common notation is all defined precisely in the aforementioned. The controversial and additional things are defined as such: $\mathbb{N} = \{1, 2, 3, ...\}$, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, $\mathbb{R}^{>0} := (0, \infty)$.

0.1 Sample spaces and set theory

Definition 1. The **sample space** Ω is the set of all possible outcomes of an experiment. An element of the sample space $\omega \in \Omega$ is a **sample point**.

Examples 2. When flipping a coin $\Omega = \{H, T\}$. When rolling a standard die $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Definition 3. Subsets of Ω are collections of sample points and called **events**.

Suppose events $A, B \subseteq \Omega$:

- $A \cup B$ is the event that A or B or both occur,
- $A \cap B$ is the event that A and B both occur,
- $A^c = \bar{A}$ is the event that occurs iff A does not occur.

Let $\mathcal{I}$ be a general index set with $A_i \subseteq \Omega,\ \forall i \in \mathcal{I}$ and $B \subseteq \Omega$. The following identities hold.

$$\left(\bigcup_{i\in\mathcal{I}} A_i\right)^c = \bigcap_{i\in\mathcal{I}} A_i^c, \quad \left(\bigcap_{i\in\mathcal{I}} A_i\right)^c = \bigcup_{i\in\mathcal{I}} A_i^c, \qquad B \cap \left(\bigcup_{i\in\mathcal{I}} A_i\right) = \bigcap_{i\in\mathcal{I}} (A_i \cup B), \quad B \cup \left(\bigcap_{i\in\mathcal{I}} A_i\right) = \bigcup_{i\in\mathcal{I}} (A_i \cap B).$$

These are **De Morgan's Laws** and **Distributivity** respectively.

Lecture 2
Tuesday
31/10/2023

0.2 Interpretation of probability

Definition 4. The **Cardinality** of a set, denoted $\text{card}(A)$ or $|A|$ is the number of elements in the set A.

Definition 5. Two sets have the same cardinality iff there exists a bijection between the them.

Definition 6. A is **finite** if it has as finite numbers of elements, A is **countably infinite** if there exists a bijection $f : A \to \mathbb{N}$, A if **countable** if it is finite or countable infinite, A is **uncountable** or **uncountable infinite** if it isn't countable.

Samples spaces can be countable or uncountable.

Definition 7 (Naive probability)**.** Suppose $|A| < \infty$ and we want to assign a probability to $A \subseteq \Omega$.

$$\mathrm{P}_{Naive}(A) := \frac{|A|}{|\Omega|} \implies \mathrm{P}(A^c) = 1 - \mathrm{P}(A).$$

This Naive example does not consider when $|A|$ is infinite but of finite area.

Example 8. Let $\Omega = \{(x, y) \in \mathbb{R}^2, x^2 + y^2 = 1\}$ and $A \subseteq \Omega$. Define:

$$\mathrm{P}(A) := \frac{\text{area of } A}{\text{area of } \Omega}$$

In the case where $A = \{(x, y) \in \mathbb{R}^2, x^2 + y^2 = 0.5^2\}$ we have $\mathrm{P}(A) = 0.25$

Remark 9. For classical / naive probability we require $|\Omega| < \infty$ or the "area" of Ω be finite.

Definition 10 (Limiting frequency)**.** Consider n_{total} repetitions of an experiment and n_A the number of time A occurs.

$$\mathrm{P}(A) := \lim_{n_{total} \to \infty} \frac{n_A}{n_{total}}$$

Unfortunately, $n_{total} \to \infty$ is often hard to conceive with finite representations not necessarily being representative.

Definition 11 (Subjective probability)**.** For an event A assign the probability $\mathrm{P}(A)$ based on our own personal beliefs. The subjective probability need not be the same for different individuals, and despite its appearance it remains a valid interpretation of probability.

Remark 12. All three interpretations of probability depend of assumptions about the experiment.

Lecture 3
Friday
03/11/2023

1 Counting

1.1 Multiplication principle

Computing naive probabilities often requires some combinatorics.

Definition 13 (Multiplication principle)**.** If we perform an experiment A that has a possible outcomes and an experiment B with b possible outcomes (in any order) then the number of outcomes of the **compound experiment** will be ab.

Remark 14. When dealing with repetitions of the same experiment (with sample space Ω, the sample space is given by the Cartesian product of the individual samples spaces.

$$\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n := \{(\omega_1, \omega_2, \ldots, \omega_n) : \omega_i \in \Omega_i\}.$$

The cardinality of this samples space follows from the multiplication principle.

1.2 Power sets

Definition 15 (Power Set)**.** Given a set A its **power set** is defined as:

$$\mathcal{P}(A) := \{X : X \subseteq A\}.$$

Theorem 16. If A is a finite set, $|\mathcal{P}(A)| = 2^{|A|}$.

1.3 Combinatorial coefficients

Definition 17 (Factorial)**.** Let $n \in \mathbb{N}$ the **factorial** of n is defined as:

$$n! := \prod_{i=1}^{n} i.$$

Definition 18 (Descending factorial)**.** Let $k, n \in \mathbb{N}$ with $k \leq n$ the **descending factorial** denoted $(n)_k$ is defined as:

$$(n)_k := n(n-1)\ldots(n-k+1) = \prod_{i=0}^{k-1}(n-i) = \prod_{j=n-k+1}^{n} j = \frac{n!}{(n-k)!}.$$

Definition 19 (Binomial coefficient)**.** Let $k, n \in \mathbb{N}_0$ the **binomial coefficient** is the number of subsets of size k of a set n:

$$\binom{n}{k} := \begin{cases} \dfrac{n(n-1)\ldots(n-(k-1))}{k!} = \dfrac{(n)_k}{k!} = \dfrac{n!}{(n-k)!k!} & \text{if } k \leq n \\ 0 & \text{otherwise.} \end{cases}$$

Lecture 4
Monday
06/11/2023

1.4 Sampling with and without replacement

"Definitions" given in the context of drawing balls from an urn, $S = \{1, 2, \ldots, n\}$.

Definition 20 (Ordered sampling with replacement)**.** Take out a ball from S, note its number, put it back; repeat this k times. The sample space for this experiment is $\Omega = S^k$.

Definition 21 (Ordered sampling without replacement)**.** Take out a ball form S, note its number but **do not** put it back; repeat $k < n$ times. There are $|\Omega| = (n)_k$ possible outcomes.

Definition 22 (Unordered sampling without replacement)**.** We take k balls out of the urn, there are $\binom{n}{k}$ possibilities.

Definition 23 (Unordered sampling with replacement)**.** We take k balls out of the urn, with the stars and bars argument: there must be k stars divided by $n-1$ bars giving us:

$$|\Omega| = \binom{n+k-1}{k} = \binom{n+k-1}{n-1}.$$

Lecture 5
Tuesday
07/11/2023

2 Axiomatic probability

2.1 Event space

We do not always want to consider all subsets of Ω so denote $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ the **event space**, which contains the events we are allowed to consider. $\mathcal{F}$ must always be a σ-algebra.

Definition 24 (Algebra)**.** $\mathcal{F}$ is an **algebra** (or a field) on Ω iff:
1. $\emptyset \in \mathcal{F}$, 2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$, 3. $A, B \in \mathcal{F} \implies A \cup B \in \mathcal{F}$.

Definition 25 (σ-algebra)**.** $\mathcal{F}$ is a **σ-algebra** (or a σ-field) on Ω iff:
1. $\emptyset \in \mathcal{F}$; 2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$, 3. For all i in some countable indexing set $\mathcal{I}$, $A_i \in \mathcal{F} \implies \bigcup_{i \in \mathcal{I}} A_i \in \mathcal{F}$.

Remark 26.
1. Any algebra is closed under finite unions and finite intersections,
2. Any σ-algebra is closed under countable intersections,
3. Any (σ-)algebra on Ω contains Ω.

Definition 27 (Trivial sigma algebra)**.** The **trivial sigma algebra** on Ω is defined as $\mathcal{F}_{trivial} := \{\emptyset, \Omega\}$.

Example 28 (Smallest σ-algebra of an element)**.** For some $A \subseteq \Omega$, the sigma algebra $\mathcal{F}_A := \{\emptyset, A, A^c, \Omega\}$ is the smallest σ-algebra on Ω (smallest cardinality) that contains A.

Lecture 6
Friday
10/11/2023

2.2 Probability measure

Definition 29 (Probability measure)**.** A mapping $\mathrm{P} : \mathcal{F} \to \mathbb{R}$ is a **probability measure** on $(\Omega, \mathcal{F})$ iff:
1. $\mathrm{P}(A) \geq 0$ for all $A \in \mathcal{F}$; 2. $\mathrm{P}(\Omega) = 1$; 3. for a countable, disjoint sequence of events $(A_i)_{i \in \mathcal{I}}$ on an indexing set $\mathcal{I}$:

$$\mathrm{P}\left(\bigcup_{i \in \mathcal{I}} A_i\right) = \sum_{i \in \mathcal{I}} \mathrm{P}(A_i).$$

2.3 Probability space

Definition 30 (Probability space)**.** A **probability space** is a triple $(\Omega, \mathcal{F}, \mathrm{P})$, with Ω a sample space, $\mathcal{F}$ a σ-algebra on Ω, and P a probability measure on $(\Omega, \mathcal{F})$.

Corollary 31. For $A, B \in \mathcal{F}$:
1. $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$, 2. $A \subseteq B \implies \mathrm{P}(A) \leq \mathrm{P}(B)$, 3. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

Lecture 7
Monday
13/11/2023

3 Conditional probability

Definition 32 (Conditional probability measure)**.** Consider the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and some event $B \in \mathcal{F}$ with $\mathrm{P}(B) > 0$, we construct the probability measure Q on $(\Omega, \mathcal{F})$ by

$$\mathrm{Q}(A) := \frac{P(A \cap B)}{P(B)}.$$

Denote the **conditional probability** of A given B by $P(A|B) = Q(B)$.

Lecture 8
Tuesday
14/11/2023

3.1 Bayes' rule and total probability

Theorem 33 (Bayes' rule)**.** For $A, B \in \mathcal{F}$ with $\mathrm{P}(A) > 0, \mathrm{P}(B) > 0$ we have,

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}.$$

Definition 34 (Partition of a set)**.** A partition of some set Ω is a collection $\{B_i, i \in \mathcal{I}\}$ for some countable index set $\mathcal{I}$ with $B_i \cap B_j = \emptyset$ for all $i, j \in \mathcal{I}$ with $i \neq j$ and $\bigcup_{i \in \mathcal{I}} B_i = \Omega$.

Theorem 35 (Total probability)**.** Given some partition $\{B_i, i \in \mathcal{I}\}$ of Ω with $\mathrm{P}(B_i) > 0$ for all $i \in \mathcal{I}$ and some event $A \in \mathcal{F}$,

$$\mathrm{P}(A) = \sum_{i \in \mathcal{I}} \mathrm{P}(A \cap B_i) = \sum_{i \in \mathcal{I}} \mathrm{P}(A|B_i)\mathrm{P}(B_i).$$

These two theorems can then be combined to form the following.

Theorem 36 (Bayes' rule with extra conditioning)**.** For events $A, B, E \in \mathcal{F}$ with $\mathrm{P}(A \cap E) > 0, P(B \cap E) > 0$ we have

$$\mathrm{P}(A|B \cap E) = \frac{\mathrm{P}(B|A \cap E)\mathrm{P}(A|E)}{\mathrm{P}(B|E)}.$$

Theorem 37 (Total probability with extra conditioning)**.** Given events $A, E \in \mathcal{I}$ with $\mathrm{P}(E) > 0$ and some partition $\{B_i, i \in \mathcal{I}\}$ of Ω with $\mathrm{P}(B_i \cap E) > 0$ for all $i \in \mathcal{I}$,

$$\mathrm{P}(A|E) = \sum_{i \in \mathcal{I}} \frac{\mathrm{P}(A \cap B_i \cap E)}{\mathrm{P}(E)} = \sum_{i \in \mathcal{I}} \mathrm{P}(A|B_i \cap E)\mathrm{P}(B_i|E).$$

Lecture 9
Friday
17/11/2023

4 Independence

4.1 Event independence

Two events $A, B \in \mathcal{F}$ will be independent iff the occurrence of one does not effect the probability the other occurs, i.e $\mathrm{P}(A|B) = \mathrm{P}(A)$ and vice versa.

Definition 38 (Independent events)**.** Two events $A, B \in \mathcal{F}$ are said to be **independent** iff

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B),$$

and **dependent** otherwise.

Corollary 39. If A and B are independent then so are all pairs of their complements.

Definition 40 (General independence)**.** A finite collection of events $\{A_1, A_2, \ldots, A_n\}$ is independent iff

$$\mathrm{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathrm{P}(A_1)\mathrm{P}(A_2)\ldots\mathrm{P}(A_n),$$

and similarly a countably or uncountably infinite collection of events is independent iff each finite subcollection is independent.

Lecture 10
Monday
20/11/2023

4.2 Conditional independence

Definition 41 (Conditional independence)**.** Given the events $A, B, C \in \mathcal{F}$ with $\mathrm{P}(C) > 0$ we say A and B are **conditional independent** given C iff,

$$\mathrm{P}(A \cap B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C).$$

4.3 Product rule for general independence

The upcoming subsection may seem disparate, they are however necessary parts to the omitted proof of the product rule for general independence and therefore deemed relevant.

Definition 42 (Set difference)**.** Given two set $A, B \in \Omega$ the **set difference** of A and B is defined as, $A \setminus B := A \cap B^c$.

Lemma 43. Any countable union of sets can be written as a countable union of disjoint sets.

Definition 44 (Increasing and decreasing sets)**.** A sequence of sets $(A_i)_{i=1}^{\infty}$ is said to **increase** to A (written $A_i \uparrow A$) iff $A_1 \subseteq A_2 \subseteq \dots$ and $\bigcup\limits_{i=1}^{\infty} = A$. The definition for a sequence of sets $(B_i)_{i=1}^{\infty}$ to **decrease** to a set B ($B_1 \downarrow B$) is defined similarly.

Theorem 45 (Continuity property of probability measures)**.** If $A_1, A_2, \dots \in \mathcal{F}$ with $A_i \uparrow A$ or $A_i \downarrow A$ for some $A \in \mathcal{F}$,

$$\lim_{i \to \infty} \mathrm{P}(A_i) = \mathrm{P}(\lim_{i \to \infty} A_i) = \mathrm{P}(A).$$

Theorem 46 (Product rule for general independence)**.** Given a countably infinite set of independent events $A_1, A_2, \dots \in \mathcal{F}$,

$$\mathrm{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \prod_{i=1}^{\infty} \mathrm{P}(A_i).$$

5 Discrete random variables

Lecture 11
Tuesday
21/11/2023

5.1 Images and their properties

throughout this subsection we will be considering the function $f : \mathcal{X} \to \mathcal{Y}$.

Definition 47 (Image)**.** For some subset $A \subseteq \mathcal{X}$ we define the **image** of A under f by,

$$f(A) := \{y \in \mathcal{Y} : \exists x \in A, y = f(x)\} = \{f(x) : x \in A\}.$$

When $A = \mathcal{X}$, $f(\mathcal{X}) = \operatorname{im} f$.

Definition 48 (Pre-image)**.** For some subset $B \subseteq \mathcal{Y}$ we now define the **pre-image** of B under f by,

$$f^{-1}(B) := \{x \in \mathcal{X} : f(x) \in B\}.$$

Despite the similar notation to the inverse function of f they are not the same thing. Notably, the pre-image under f always exists while the inverse function need not exist.

Lemma 49. For a collection of subsets $B_i \in \mathcal{F}$ for i in some indexing set $\mathcal{I}$ we have,

$$f^{-1}\left(\bigcup_{i \in \mathcal{I}} B_i\right) = \bigcup_{i \in \mathcal{I}} f(B_i).$$

5.2 DRVs and their distributions

Definition 50 (Discrete random variable)**.** A **discrete random variable** (**DRV**) on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is a function $X : \Omega \to \mathbb{R}$ that satisfies the following properties:

- $\operatorname{im} X = \{X(\omega) : \omega \in \Omega\}$ must be a countable subset of $\mathbb{R}$,
- $X^{-1}(x) \in \mathcal{F}$ for all $x \in \mathbb{R}$.

Remark 51. The nomenclature of X being discrete stems from the fact that its image is a countable subset of $\mathbb{R}$ and so can be mapped to $\mathbb{N}$ which we see as being discrete.

Definition 52 (Probability mass function)**.** The **probability mass function** (**pmf**) of a DRV X is defined as a function $p_X : \mathbb{R} \to [0,1]$ such that,

$$p_X(x) := \mathrm{P}(X^{-1}(x)).$$

This is commonly denoted by $p_X(x) = \mathrm{P}(X = x)$.

Remark 53. Some useful propoerties of the pmf extending from the definition are:

- $x \notin \operatorname{im} X \implies p_X(x) = 0$,
- For $x_1, x_2 \in \operatorname{im} X$ with $x_1 \neq x_2$, $X^{-1}(x_1) \cap X^{-1}(x_2) = \emptyset$,
- $\sum_{x \in \operatorname{im} X} p_X(x) = \sum_{x \in \mathbb{R}} p_X(x) = 1$.

Theorem 54. Suppose $\mathcal{I}$ is some indexing set and $S = \{s_i \in \mathbb{R} : i \in \mathcal{I}\}$ is countable and $\{\pi_i : i \in \mathcal{I}\}$ is a collection such that $\pi_i \geq 0$ for all $i \in \mathcal{I}$ and $\sum_{i \in \mathcal{I}} \pi_i = 1$. Then there exists some probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and a DRV X on said probability space such that $p_X(s_i) = \pi_i$ for all $i \in \mathcal{I}$ and $p_X(s) = 0$ otherwise.

6 Common DRVs

All DRVs within this section will be considered over the probability space $(\Omega, \mathcal{F}, \mathrm{P})$.

6.1 Bernoulli distribution

Definition 55 (Bernoulli distribution)**.** A DRV X is said to have **Bernoulli distribtuion** with parameter $p \in (0,1)$ if $\operatorname{im} X = \{0,1\}$ with $p_X(1) = p$. This is denoted by $X \sim \mathrm{Bern}(p)$.

Definition 56 (Indicator variable)**.** Given some event $A \in \mathcal{F}$ the **indicator variable** of the event A is given by,

$$\mathbb{I}_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}.$$

Remark 57. $\mathbb{I}_A \sim \mathrm{Bern}(\mathrm{P}(A))$.

6.2 Binomial distribution

Definition 58 (Binomial distribution)**.** Consider a sequence of $n \in \mathbb{N}$ iid Bernoulli trials with parameter p, count the number of successes and denote this by the random variable X then $\operatorname{im} X = [0, n]$ and,

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x \in [0, n].$$

We say X follows a **binomial distribution** and this is denoted by $X \sim \mathrm{Bin}(n, p)$.

6.3 Hypergeometric distribution

As we have done previously, consider of urn of $N \in \mathbb{N}$ balls with $K \in \mathbb{N}$ of these being white and the remainder being black from which we will draw $n \in \mathbb{N}$ balls and want to consider the DRV (X) for the number of white balls drawn. When drawing with replacement we have $X \sim \mathrm{Bin}(n, K/N)$. However, when drawing without replacement X follows the hypergeometric distribution.

Definition 59 (Hypergeometric distribution)**.** A DRV X follows the **hypergeometric distribution** with three parameters $N \in \mathbb{N}_0, K \in \mathbb{N}, n \in [0, N]$ if $\operatorname{im} X = [0, \min(n, K)]$ and,

$$p_X(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \quad \text{for } x \in [0, K].$$

Lemma 60 (Vandemonde's identity)**.** **Vandemonde's identity** is an important tool in the derivation of the pmf for the hypergeometric distribution and so is included here. The identity is as follows, for $k, m, n \in \mathbb{N}$ with $k \leq m + n$, we have:

$$\binom{m+n}{k} = \sum_{i=0}^{k} \binom{m}{i} \binom{n}{k-i}.$$

6.4 Discrete uniform distribution

Definition 61 (Discrete uniform distribution)**.** A DRV X follows the **discrete uniform distribution** over a nonempty set of numbers C, denoted $X \sim \mathrm{DUnif}(C)$, if $\operatorname{im} X = C$ and,

$$p_X(x) = \begin{cases} \dfrac{1}{\mathrm{card}(C)} & \text{for } x \in C \\ 0 & \text{otherwise} \end{cases}.$$

6.5 Poisson distribution

The poisson distribution is commonly used for modelling the number of events occurring in a certain time period. Its pdf is derived by taking the $\lim_{n\to\infty} p_X(x)$ where $X \sim \mathrm{Bin}(n, \frac{\lambda}{n})$ for some $\lambda \in \mathbb{R}$.

Definition 62 (Poisson distribution)**.** A DRV X follows the **poisson distribution** with parameter $\lambda \in \mathbb{R}^{>0}$, denoted $X \sim \mathrm{Poi}(\lambda)$, if $\operatorname{im} X = \mathbb{N}_0$ and,

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for } x \in \mathbb{N}_0.$$

6.6 Geometric distribution

Definition 63 (Geometric distribution)**.** A DRV X follows the **geometric distribution** with parameter $p \in (0, 1)$, denoted $X \sim \mathrm{Geom}(p)$, if $\operatorname{im} X = \mathbb{N}$ and,

$$p_X(x) = (1-p)^x p \quad \text{for } x \in \mathbb{N}.$$

This can be seen as counting the number of Bernoulli trials with parameter p that occur before a failure.

6.7 Negative binomial distribution

Definition 64 (Generalised binomial coefficient)**.** Let $\alpha \in \mathbb{C}$ and $k \in \mathbb{N}$ and define the **generalised binomial coefficient** by,

$$\binom{\alpha}{k} := \frac{\alpha(\alpha-1)\ldots(\alpha-k+1)}{k!}.$$

Lemma 65. For $x \in \mathbb{N}_0$ and $\mathrm{r} \in \mathbb{N}$ the following identity holds,

$$\binom{x+r-1}{r-1} = (-1)^x \binom{-r}{x}.$$

The generalised binomial coefficient as well as this lemma are necessary to have a well defined and valid pdf for the negative binomial distribution.

Definition 66 (Negative binomial distribution)**.** A DRV X follows the **negative binomial distribution** with parameters $r \in \mathbb{N}$ and $p \in (0, 1)$, denoted $X \sim \mathrm{NBin}(r, p)$, if $\operatorname{im} X = \mathbb{N}_0$ and,

$$p_X(x) = \binom{x+r-1}{r-1} p^r (1-p)^x \quad \text{for } x \in \mathbb{N}_0.$$

This is the distribution of the number of failed ii Bernoulli trials with parameter p before r successes have occurred.

7 Continuous random variables

7.1 General random variables and their distributions

Definition 67 (Random variable)**.** A **random variable** (**RV**) on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is a mapping $X : \Omega \to \mathbb{R}$ such that $X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$. By taking the countable union of pre-images of all $\omega \leq x$ in $\mathcal{F}$, it can be seen that a DRV satisfies this condition.

Definition 68 (Cumulative distribution function)**.** For some RV X on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, the **cumulative distribution function** (**CDF**) of X is defined as the mapping $F_X : \mathbb{R} \to [0, 1]$ given by,

$$F_X(x) = \mathrm{P}(X^{-1}((-\infty, x])),$$

often denoted $F_X(x) = \mathrm{P}(X \leq x)$.

Theorem 69 (cdf properties)**.** For some RV X on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ the following properties hold:

1. F_X is monotonically non-decreasing,
2. F_X is right-continuous ($(x_n) \downarrow x \implies F_X(x_n) \to F_X(x)$ as $n \to \infty$),
3. $\lim\limits_{x \to -\infty} F_X(x) = 0$ and $\lim\limits_{x \to \infty} F_X(x) = 1$.

Theorem 70. For $a, b \in \mathbb{R}$ if $a < b$, then $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$.

Remark 71. In general the cdf of an RV is not left continuous.

7.2 CRVs and pdfs

Definition 72 (Continuous random variable)**.** A random variable X on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is called a **continuous random variable** (**CRV**) iff its cdf can be written as:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \quad \text{for all } x \in \mathbb{R},$$

where $f_X : \mathbb{R} \to \mathbb{R}$ satisfies: $f_X(u) \geq 0$ for all $u \in \mathbb{R}$ and $\displaystyle\int_{-\infty}^{\infty} f_X(u)du = 1$. We call f_X the **probability density function** (**pdf**) of X.

Theorem 73. If X is a CRV on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ with pdf f_X, $\mathrm{P}(X = x) = 0$ for all $x \in \mathbb{R}$.

Theorem 74. With the same conditions, $\mathrm{P}(a \leq X \leq b) = \displaystyle\int_{a}^{b} f_X(u)du$ for all $a, b \in \mathbb{R}$ with $a \leq b$.

Remark 75. Combining the results from this section leads to the conclusion that the cdf or a CRV is continuous.

8 Common CRVs

All CRVs X within this section will be considered over the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ with the natural notation for their pdf and cdf. These distribution will be uniquely identified by their pdfs.

8.1 Uniform distribution

Definition 76 (Uniform distribution)**.** A CRV X follows the **uniform distribution** on the interval (a, b) for $a, b \in \mathbb{R}$ with $a < b$, denoted $X \sim \mathrm{U}(a, b)$ if it satisfies:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}, \qquad F_X(x) = \begin{cases} 0 & \text{if } x \leq a \\ \dfrac{1}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b \end{cases}.$$

8.2 Exponential distribution

Definition 77 (Exponential distribution)**.** A CRV X follows the **exponential distribution** with parameter $\lambda \in \mathbb{R}^{>0}$, denoted $X \sim \mathrm{Exp}(\lambda)$ if it satisfies:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}, \qquad F_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}.$$

8.3 Gamma distribution

Definition 78 (Gamma function)**.** For $t \in \mathbb{R}$ with $t > 0$ we define the **gamma function** by,

$$\Gamma(t) := \int_0^\infty x^{t-1} e^{-x} dx.$$

One of the gamma function's many interesting properties is that $\Gamma(t) = (t-1)\Gamma(t-1)$ for $t > 1$.

Definition 79 (Gamma distribution)**.** A CRV X follows the **gamma distribution** with shape and rate parameter $\alpha, \beta \in \mathbb{R}^{>0}$ respectively, denoted $X \sim \mathrm{Gamma}(\alpha, \beta)$ if it satisfies:

$$f_X(x) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Its cdf cannot be written in a closed form so must be left as an integral of the pdf or approximated.

8.4 Chi-squared distribution

Definition 80 (Chi-squared distribution)**.** A CRV X follows the **chi-squared distribution** with $n \in \mathbb{N}$ degrees of freedom, denoted $X \sim \chi^2(n)$ if it satisfies:

$$f_X(x) = \begin{cases} \dfrac{1}{2\Gamma(n/2)} \left(\dfrac{x}{2}\right)^{n/2-1} e^{-x/2} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Its cdf can also not be written in a closed form. The $\chi^2(n)$ distribution is the same as the $\mathrm{Gamma}(\frac{n}{2}, \frac{1}{2})$ distribution.

8.5 F-distribution

These pdfs are getting tough.

Definition 81 (F-distribution)**.** A CRV X follows the **f-distribution** with $d_1, d_2 \in \mathbb{R}^{>0}$ degrees of freedom, denoted $X \sim \mathrm{F}(d_1, d_2)$ if it satisfies:

$$f_X(x) = \begin{cases} \dfrac{\Gamma\left(\frac{d_1+d_2}{2}\right)\left(\frac{d_1}{d_2}\right)^{d_1/2} x^{d_1/2-1}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)\left(1+\frac{d_1}{d_2}x\right)^{(d_1+d_2)/2}} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Its cdf can also not be written in a closed form. It is important to note that d_1, d_2 are not restricted to integer values, and that $X = \frac{X_1/d_1}{X_2/d_2}$ where $X_1 \sim \chi^2(d_1)$ and $X_2 \sim \chi^2(d_2)$.

8.6 Beta distribution

Definition 82 (Beta function)**.** For $\alpha, \beta \in \mathbb{R}^{>0}$ we define the **beta function** by,

$$B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Definition 83 (Beta distribution)**.** A CRV X follows the **beta distribution** with parameters $\alpha, \beta \in \mathbb{R}^{>0}$, denoted $X \sim \mathrm{Beta}(\alpha, \beta)$ if it satisfies:

$$f_X(x) = \begin{cases} \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

Its cdf can also not be written in a closed form.

8.7 Normal distribution

Definition 84 (Standard normal distribution)**.** A CRV X follows the **standard normal distribution** or **Gaussian distribution**, denoted $X \sim \mathrm{N}(0,1)$ if it satisfies,

$$f_X(x) = \phi(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for } x \in \mathbb{R}, \qquad F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \quad \text{for } x \in \mathbb{R}.$$

Where, once again, there is no explicit formula for the cdf.

Definition 85 (Normal distribution)**.** A CRV X follows the **normal distribution** with mean $\mu \in \mathbb{R}$ and variance σ^2 for $\sigma \in \mathbb{R}^{>0}$ denoted $X \sim \mathrm{N}(\mu, \sigma^2)$ if it satisfies,

$$f_X(x) = \phi(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in \mathbb{R}.$$

8.8 Cauchy distribution

Definition 86 (Cauchy distribution)**.** A CRV X follows the **Cauchy distribution** if it satisfies,

$$f_X(x) = \frac{1}{\pi(1+x^2)} \quad \text{for } x \in \mathbb{R}, \qquad F_X(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2} \quad \text{for } x \in \mathbb{R}.$$

If $X, Y \sim \mathrm{N}(0,1)$, then $Z = X/Y$ follows the Cauchy distribution.

8.9 Student t-distribution

Definition 87 ((Student's) t-distribution)**.** A CRV X follows the **Student t-distribution** with $\nu \in \mathbb{R}^{>0}$ degrees of freedom if it satisfies,

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } x \in \mathbb{R}.$$

Its cdf cannot be written in a closed form.

Remark 88. Not all RVs are either discrete or continuous.

9 Transformations of random variables

9.1 DRVs

Theorem 89. Let X be a DRV on $(\Omega, \mathcal{F}, \mathrm{P})$ and let $g : \mathbb{R} \to \mathbb{R}$ be a deterministic function, then $Y = g(X)$ is a DRV with pmf:

$$p_Y(y) = \sum_{\{x \in \operatorname{im} X : g(x) = y\}} p_X(x) \quad \text{if } y \in \operatorname{im} Y \text{ and } 0 \text{ otherwise.}$$

9.2 CRVs

Theorem 90. Let X be a CRV on $(\Omega, \mathcal{F}, \mathrm{P})$ and let $g : \mathbb{R} \to \mathbb{R}$ be a strictly monotonis and differentiable function with inverse $g^{-1} : \mathbb{R} \to \mathbb{R}$, then $Y = g(X)$ is a CRV with pdf:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} \left[g^{-1}(y)\right] \right| \quad \text{for all } y \in \mathbb{R}\ .$$

Remark 91. The term $\left| \dfrac{d}{dy} \left[g^{-1}(y)\right] \right|$ is often called the **Jacobian** of the transformation.

10 Expectation of random variables

Throughout this section, unless otherwise specified, all infinite sums will be assumed to converge absolutely and all integrals will be assumed to be $< \infty$.

10.1 Definition

Definition 92 (Expectation of a DRV). Let X be a DRV on $(\Omega, \mathcal{F}, \mathrm{P})$ then the **expectation** of X is defined by,

$$\mathrm{E}(X) := \sum_{x \in \operatorname{im} X} x p_X(x).$$

Definition 93 (Expectation of a CRV). Let X be a CRV on $(\Omega, \mathcal{F}, \mathrm{P})$ then the **expectation** of X is defined by,

$$\mathrm{E}(X) := \int_{-\infty}^{\infty} x f_X(x) dx.$$

10.2 LOTUS

Theorem 94 (Discrete LOTUS). Let X be a DRV on $(\Omega, \mathcal{F}, \mathrm{P})$ and $g : \mathbb{R} \to \mathbb{R}$, we have,

$$\mathrm{E}(g(X)) = \sum_{x \in \operatorname{im} X} g(x) p_X(x).$$

Theorem 95 (Continuous LOTUS). Let X be a CRV on $(\Omega, \mathcal{F}, \mathrm{P})$ and $g : \mathbb{R} \to \mathbb{R}$, we have,

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Note that this is one of the few theorems throughout the course given without proof.

Theorem 96. If X is non-negative then $\mathrm{E}(X) \geq 0$.

10.3 Variance

Definition 97 (Variance). Let X be a discrete/continuous random variable, then the **variance** of X is defined by,

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

Theorem 98. For a discrete/continuous random variable with finite variance,

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

11 Multivariate random variables

11.1 Multivariate distributions

Definition 99 (Join distribution function). Consider the sequence of random variables $X_1, X_2, \ldots, X_n$ all on $(\Omega, \mathcal{F}, \mathrm{P})$ and write $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Then the **joint distribution function** of $\mathbf{X}$ is $F_{\mathbf{X}} : \mathbb{R}^n \to [0, 1]$ defined by:

$$F_{\mathbf{X}}(\mathbf{x}) := \mathrm{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \quad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

11.2 Independence

Definition 100 (Pairwise independence for n random variables)**.** We call the sequence of RVs, $X_1, X_2, \ldots, X_n$, **pairwise independent** iff,

$$F_{X_i,X_j}(x_i, x_j) = F_{X_i}(x_i)F_{X_j}(x_j) \quad \text{for all } x_i, x_j \in \mathbb{R} \text{ with } i \neq j.$$

Definition 101 (Independence of a family of random variables)**.** Given some indexing set $\mathcal{I} \subset \mathbb{R}$, a family of random variables $\{X_i : i \in \mathcal{I}\}$ is **independent** iff for all finite $\mathcal{J} \subseteq \mathcal{I}$:

$$\mathrm{P}\left(\bigcap_{j\in\mathcal{J}} \{X_j \leq x_j\}\right) = \prod_{j\in\mathcal{J}} \mathrm{P}(\{X_j \leq x_j\}) \quad \text{for all } (x_j)_{j\in\mathcal{J}} \text{ with } x_j \in \mathbb{R}.$$

(All finite subfamilies of the family of random variables is independent by the natural definition)

11.3 Multivariate DRVs

Definition 102 (Joint probability mass functions)**.** Let $X_1, X_2, \ldots, X_n$ all be DRVs on $(\Omega, \mathcal{F}, \mathrm{P})$ that form the random vector $\mathbf{X}$, their **joint probability mass function**, $p_{\mathbf{X}} : \mathbb{R}^n \to [0, 1]$ is defined as,

$$p_{\mathbf{X}}(x_1, x_2, \ldots, x_n) := \mathrm{P}(\{\omega \in \Omega : X_1(\omega) = x_1, X_2(\omega) = x_2, \ldots, X_n(\omega) = x_n\}) = \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

The **marginal probability mass function** of $X_i \in \mathbf{X}$ is given by,

$$\mathrm{p}_{X_i}(k) = \sum_{(x_1,x_2,\ldots,x_n)\in\mathbb{R}^n} \cdots \sum p_{\mathbf{X}}(x_1, x_2, \ldots, x_{i-1}, k, x_{i+1}, \ldots, x_n).$$

It can be obtained that for any sufficiently "nice" set $A \in \mathbb{R}^n$,

$$\mathrm{P}(\mathbf{X} \in A) = \sum_{(x_1,x_2,\ldots,x_n)\in A} \cdots \sum \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Definition 103 (Independence of DRVs)**.** Given some indexing set $\mathcal{I} \in \mathbb{R}$ a family of DRVs, $\{X_i : i \in \mathcal{I}\}$ with joint pmf $p_{\mathbf{X}}$, is **independent** iff for all finite $\mathcal{J} \in \mathcal{I}$:

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i}(x_i) \quad \text{for all } \mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n.$$

11.4 Multivariate CRVs

Definition 104 (Continuous random vector)**.** The random vector $\mathbf{X} = (X_1, X_2, \ldots X_n)$ is a **continuous random vector** iff,

$$F_{\mathbf{X}}(\mathbf{x}) = \int \cdots \int_B f_{\mathbf{X}}(\mathbf{x}) dx_1 dx_2 \cdots dx_n \quad \text{with } B = (\infty, x_1] \times (\infty, x_2] \times \cdots \times (\infty, x_n], \quad \text{for all } \mathbf{x} \in \mathbb{R}^n;$$

for some $f_{\mathbf{X}} : \mathbb{R}^n \to \mathbb{R}$ satisfying: $f_{\mathbf{X}}(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\int \cdots \int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}) dx_1 dx_2 \cdots dx_n = 1$.

Note that $f_{\mathbf{X}}(\mathbf{x}) = \dfrac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F_{\mathbf{X}}(\mathbf{x})$ and $\mathrm{P}(\mathbf{X} \in A) = \int \cdots \int_A f_{\mathbf{X}}(\mathbf{x}) d^n\mathbf{x}$.

Definition 105 (Independence of CRVs)**.** Given some indexing set $\mathcal{I} \in \mathbb{R}$ a family of CRVs, $\{X_i : i \in \mathcal{I}\}$ with joint pdf $f_{\mathbf{X}}$, is **independent** iff for all finite $\mathcal{J} \in \mathcal{I}$:

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) \quad \text{for all } \mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n.$$

11.5 Transformations of random vector

Definition 106 (Transformation)**.** We are going to **transform** the random vector $\mathbf{X}$ with joint pdf $f_{\mathbf{X}}$ to $\mathbf{U} = (u_1(\mathbf{X}), u_2(\mathbf{X}), \ldots, u_n(\mathbf{X}))$ with $u_i : \mathbb{R}^n \to \mathbb{R}^n$ for all $i \in [1, n]$. Firstly, define $T : \mathbb{R}^n \to \mathbb{R}^n$ by $T(\mathbf{x}) = (u_1(\mathbf{x}), u_2(\mathbf{x}), \ldots, u_n(\mathbf{x}))$ and assume T is bijective on $D = \{\mathbf{x} \in \mathbb{R}^n : f_{\mathbf{X}}(\mathbf{x}) > 0\}$ with range $S \subseteq \mathbb{R}^n$. Secondly, have the Jacobian determinant of $T^{-1} : S \to D$, $J(\mathbf{u}) = \det([a_{ij}]_{m\times n})$ with $a_{ij} = \frac{\partial x_i}{\partial u_j}$. Finally, define:

$$f_{\mathbf{U}}(\mathbf{u}) := \begin{cases} f_{\mathbf{X}}(T^{-1}(\mathbf{u}))|J(\mathbf{u})| & \text{if } \mathbf{u} \in S \\ 0 & \text{otherwise} \end{cases}.$$

11.6 Multivariate LOTUS

Theorem 107 (Discrete multivariate LOTUS)**.** If $X_1, X_2, \ldots, X_n$ are DRVs on $(\Omega, \mathcal{F}, \mathrm{P})$ and form the random vector $\mathbf{X}$ with $g : \mathbb{R}^n \to \mathbb{R}$, then $Y = g(\mathbf{X})$ is a DRV on $(\Omega, \mathcal{F}, \mathrm{P})$ with expectation,

$$\mathrm{E}(g(\mathbf{X})) = \sum_{x_i \in \operatorname{im} X_i} \cdots \sum g(\mathbf{X})\mathrm{P}(\mathbf{X} = \mathbf{x}) \quad \text{for all } \mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n.$$

Theorem 108 (Continuous multivariate LOTUS)**.** If $X_1, X_2, \ldots, X_n$ are DRVs on $(\Omega, \mathcal{F}, \mathrm{P})$ and form the random vector $\mathbf{X}$ with $h : \mathbb{R}^n \to \mathbb{R}$ we have,

$$\mathrm{E}(h(\mathbf{X})) = \int \cdots \int_{\mathbb{R}^n} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) dx_1 dx_2 \cdots dx_n \quad \text{for all } \mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n.$$

11.7 Covariance

Definition 109 (Covariance)**.** Given two random variable X and Y on the same probability space with expectations μ_X and μ_Y respectively. The **covariance** of X and Y is defined as,

$$\mathrm{Cov}(X, Y) := \mathrm{E}\left[(X - \mu_X)(Y - \mu_Y)\right] \quad \text{assuming both expectation take finite values.}$$

Definition 110 (Correlation)**.** Given the same X and Y the **correlation** of X and Y is defined as,

$$\mathrm{Cor}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Theorem 111. For jointly discrete/continuous RVs with finite expectations the following hold:

1. when $X = Y$, $\mathrm{Cov}(X, Y) = \mathrm{E}\left[(X - \mu_X)^2\right] = \mathrm{Var}(X)$,
2. $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$,
3. when X and Y are independent, $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$,
4. when variances are also finite, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$.

12 Generating functions

12.1 Probability generating functions

Definition 112 (Probability generating functions)**.** Given a DRV X with $\mathrm{im}(X) \subseteq \mathbb{N}_0$, denote,

$$\mathcal{S}_X := \left\{ s \in \mathbb{R} : \sum_{x=0}^{\infty} |s|^x \mathrm{P}(X = x) < \infty \right\},$$

and define the **probability generating function (pgf)** of X as the function $G_X : \mathcal{S}_X \to \mathbb{R}$ given by,

$$G_X(s) := \mathrm{E}(s^X) = \sum_{x=0}^{\infty} s^x \mathrm{P}(X = x),$$

noting that the pgf is well defined for $|s| < 1$ and $G_X(0) = \mathrm{P}(X = 0)$ and $G_X(1) = 1$.

Theorem 113 (Uniqueness of pgfs)**.** Given two DRVs X and Y with pgfs G_X and G_Y respectively,

$$G_X(s) = G_Y(s) \quad \text{for all } s \in \mathcal{S}_X \cap \mathcal{S}_Y \iff p_X(x) = p_Y(x) \quad \text{for all } x \in \mathbb{N}_0.$$

Theorem 114. Let X, Y be independent DRVs with $\operatorname{im} X, \operatorname{im} Y \in \mathbb{N}_0$, then

$$G_{X+Y}(s) = G_X(s)G_Y(s) \quad \text{for all } s \in \mathcal{S}_X \cap \mathcal{S}_Y.$$

Theorem 115 (Pgfs of sum of independent DRVs)**.** Given a collection of n independent DRVs $X_1, X_2, \ldots, X_n$,

$$G_{\sum_{i=1}^n X_i}(s) = \prod_{i=1}^{n} G_{X_i}(s) \quad \text{for all } s \in \bigcap_{i=1}^{n} \mathcal{S}_{X_i}.$$

Theorem 116 (Moments)**.** Given a DRV X with $\operatorname{im} X \subseteq \mathbb{N}_0$, t he kth derivative of X, for $k \in \mathbb{N}$ is given by,

$$\frac{d^k}{ds^k} G_X(s)\bigg|_{s=1} = G_X^{(s)}(1) = \mathrm{E}[X(X-1)\ldots(X-k+1)].$$

12.2 Common pgfs

Example 117 (Bernoulli distribution)**.** Let $X \sim \mathrm{Bern}(p)$, then $G_X(s) = 1 - p + sp$ for all $s \in \mathbb{R}$.

Example 118 (Binomial distribution)**.** Let $X \sim \mathrm{Bin}(n, p)$, then $G_X(s) = (1 - p + sp)^n$ for all $s \in \mathbb{R}$.

Example 119 (Poisson distribution)**.** Let $X \sim \mathrm{Poi}(\lambda)$, then $G_X(s) = \exp(\lambda(s - 1))$ for all $s \in \mathbb{R}$.

12.3 Moment generating functions

Definition 120 (Moment generating function)**.** Let X be a RV, its **moment generating function (mgf)** is defined as,

$$M_X(t) = \mathrm{E}(e^{tX}),$$

assuming the expectation exists in some neighbourhood of zero.

Remark 121. If X is a RV with a mgf, $M_X(t) = \mathrm{E}(e^{tX}) = G_X(e^t)$

Theorem 122. If X is a RV with a mgf, the kth moment of X is $\mathrm{E}(X^k) = M_X^{(k)}(0)$.

Theorem 123. If $X_1, X_2, \ldots, X_n$ are a family of independent RVs on the same probability space with mgfs $M_{X_1}, M_{X_2}, \ldots, M_{X_n}$ respectively, we have,

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t).$$

Theorem 124 (Characterisation by mgf)**.** If the RVs X, Y have existent mgfs M_X, M_Y respectively such that $M_X(t) = M_Y(t)$ for all t in some neighbourhood of 0, we have,

$$F_X(u) = F_Y(u) \quad \text{for all } u.$$

13 Conditional distribution and expectation

13.1 Discrete: Conditional expectation and total expectation

Definition 125 (Condition distribution and expectation of a DRV)**.** Given a DRV Y on $(\Omega, \mathcal{F}, \mathrm{P})$ and some event $B \in \mathcal{F}$ with $\mathrm{P}(B) > 0$, the **conditional distribution** of Y given B is defined as,

$$\mathrm{P}(Y = y|B) := \frac{P(\{Y = y\} \cap B)}{\mathrm{P}(B)} \quad \text{for } y \in \mathbb{R};$$

with the **conditional expectation** of Y given B defined as,

$$\mathrm{E}(Y|B) := \sum_{i \in \operatorname{im} Y} e\mathrm{P}(Y = y|B).$$

Theorem 126 (Discrete law of total expectation)**.** Given a DRV Y on $(\Omega, \mathcal{F}, \mathrm{P})$ and some parition $\{B_i : i \in \mathcal{I}\}$ of Ω with $\mathrm{P}(B_1) > 0$ for all $i \in \mathcal{I}$ we have,

$$\mathrm{E}(Y) = \sum_{i \in \mathcal{I}} \mathrm{E}(Y|B_i)\mathrm{P}(B_i).$$

13.2 Conditioning on a DRV

Definition 127 (Conditional probability mass function)**.** Given two joint DRVs (X, Y), the **conditional probability mass function** of Y given $X = x$ is given by,

$$p_{Y|X}(y|x) := \frac{p_{X,Y}(x, y)}{p_X(x)} \quad \text{for } y \in \mathbb{R}.$$

Theorem 128 (Conditional expectation)**.** Given two joint DRVs (X, Y), the **conditional expectation** of Y given $X = x$ is given by,

$$\mathrm{E}(Y|X = x) = \sum_{y \in \operatorname{im} Y} y p_{Y|X}(y|x).$$

Independence, LOTUS and a Bayes' rule formulation all follow naturally from this as they do for the non-distribution settings.

13.3 Continuous: Conditional density, distribution and expectation

Definition 129 (Conditional distribution and conditional density)**.** For two joint CRVs (X, Y) the **conditional density** of Y given $X = x$ is define as,

$$f_{Y|X}(y|x) := \frac{f_{X,Y}(x, y)}{f_X(x)} \quad \text{for all } y, x \in \mathbb{R} \text{ with } f_X(x) > 0;$$

with the corresponding **conditional distribution** of Y given $X = x$ defined as,

$$F_{Y|X=x}(y|x) := \frac{1}{f_X(x)} \int_{-\infty}^{y} f_{X,Y}(x, t)dt \quad \text{for all } y, x \in \mathbb{R} \text{ with } f_X(x) > 0.$$

Where, once again, familiar formulations for independence and Bayes' rule can be easily derived.

Definition 130 (Conditional expectation)**.** Given two joint CRVs (X, Y), the **conditional expectation** of Y given $X = x$ is defined as,

$$\mathrm{E}(Y|X = x) := \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy \quad \text{provided } f_X(x) > 0.$$

Theorem 131 (Continuous law of total expectation)**.** Given two joint CRVs (X, Y) with $\mathrm{E}(|Y|) < \infty$, we have,

$$\mathrm{E}(Y) = \int_{\{x: f_X(x) > 0\}} \mathrm{E}(Y|X = x) f_X(x)dx.$$