MATHS 2103 / MATHS 7103
Probability and Statistics II
Lecture notes

Section 1

Andrew Smith

School of Mathematical Sciences, University of Adelaide

Semester 1, 2017

Course outline

Section 1: General information and introduction to Probability

Section 2: Discrete random variables

Section 3: Continuous random variables

Section 4: Bivariate (multivariate) probability distributions

Section 5: Discrete Time Markov Chains (DTMC)

General Information

MATHS 2103/7056 Probability and Statistics II

Semester 1, 2017.

Lecture times:

- 12:10pm on Monday, Badger, G31, Macbeth Lecture Theatre.
- 1:10pm on Wednesday, Lower Napier, LG29
- 1:10pm on Thursday, Badger, G31, Macbeth Lecture Theatre

Consulting times: 11am on Mondays and Fridays, or by arrangement.

Tutorials will be held in weeks $2, 4, 6, 8, 10, 12$. For each tutorial, a selection of students will be required to present some answers. The schedule is/will be on MyUni. Use the the odd weeks to prepare.

General Information

MATHS 2103/7056 Probability and Statistics II

Semester 1, 2017.

A Small Group Discovery Experience (SGDE) will be conducted over three sessions in IW 235/236 in weeks 3, 5 and 9.[1]

This is a group project. The three practicals will give you a basic understanding of the programming, along with some ideas of what to include in your report. You will then, as a group, decide which project to do and write a report aimed at your specific audience.

The only classes are:
Lectures: every week 12:10pm Mon, 1.10pm Wednesday & Thursday.
Tutorials: even weeks. **SGDE:** weeks 3, 5 & 9.

[1] IW = Ingkarni Wardli

Assessment

Assignments will be due in weeks $3, 5, 7, 9, 11$.
They are to be submitted online in MyUni by Thursday at 4pm.

Please note that late assignments (1 minute or 1 day) will not be marked.

Exemptions can be given on medical or compassionate grounds.

A Small Group Discovery Experience (SGDE) will be conducted, for which a group written report will be due at the end of week 12.

Assessment

Final Exam 70%, SGDE project 15%, assignments 15%.

Grading Scheme

High distinction	85 – 100%
Distinction	75 – 84%
Credit	65 – 74%
Pass	50 – 64%
Fail	0 – 49%

Supplementary Examinations:

Students are advised that the University of Adelaide has an official policy on supplementary, replacement or additional assessment, which they are advised to read at

`http://www.adelaide.edu.au/student/exams/supps.html`

Plagiarism

Policy on plagiarism:

Students are advised that the University of Adelaide has an official policy on academic honesty and plagiarism, which they are advised to read at: `http://www.adelaide.edu.au/policies/230/`

Policy on assignments for this course only:

Students are encouraged to work together in order to enhance their understanding of the subject matter. To this end they may work together on assignments. However, students are required to have, and may be required to demonstrate, a complete understanding of their submitted assignments. Failure to demonstrate a complete understanding of a submitted assignment may be interpreted as evidence of plagiarism.

Hint: One easy way to work together and yet avoid any issues with plagiarism is to plan how to do the assignment together, but then work on the assignment itself separately. This way, you are able to learn together, but are not tempted to **copy** each other's actual assignment.

Course content

The content (all lecture notes, assignments, etc) for this course will be uploaded to MyUni, which now uses the new LMS called Canvas.

There are lecture notes for each of the five sections. In each section, there is also an accompanying tutorial. Solutions for these tutorials will be posted in the same locations.

Useful References

1. *Mathematical Statistics with Applications*

   Wackerly, Mendenhall and Schaeffer (Duxbury – various editions).

2. *Introduction to Probability*

   Blitzstein and Hwang (Chapman and Hall 2014).

3. *Introduction to Stochastic Models*

   Roe Goodman (2nd edition, Dover, 2006).

4. *Introduction to Probability Models*

   Sheldon Ross (Academic Press).

5. *Mathematical Statistics and Data Analysis.*

   John Rice (Duxbury Press).

There are many good books on probability and statistics in the Barr Smith Library. Many texts are also now available as ebooks through

“All possible definitions of probability fall short of the actual practice.”
- William Feller 1906–1970 (Mathematician)

“How dare we speak of the laws of chance? Is not chance the antithesis of all law?.”
- Joseph Bertrand 1822–1900 (Mathematician)

“The probable is what usually happens.”
- Aristotle 384 BC–322 BC (Philosopher)

`http://www.maths.uq.edu.au/probweb/quotes.html`

Example 1.1 Motivating example

You are a medical consultant. A study of the residents of a region showed that 20% are smokers. The probability of death due to lung cancer, given that a person smokes, is ten times the probability of death due to lung cancer, given the person does not smoke. If the probability of death due to lung cancer in the region is 0.006, what is the probability of death due to lung cancer given that a person is a smoker? ◁

Probability notation

- **Sample space:** $\mathcal{S}$ is the set of all possible outcomes.

- **Event:** $A, B, \ldots$ is a combination of outcomes, and a *subset* of the sample space $\mathcal{S}$.

- **Probability:** is a measure, or function on the outcomes and events of the sample space.

  The probability of an event A is denoted $P(A)$. It assigns a numerical value to each outcome and event in the sample space, according to specified rules.

Examples of sample spaces

Annual rainfall for a given city could take any non-negative value:

The number of cars passing a given point on the road in 1 hour could take any non-negative integer:

The outcome of tossing 3 distinct coins:

Examples of events

Rainfall less than 600mm in a year:

Three cars passing a given point:

Obtaining exactly 2 heads:

Set notation

- Universal set: $\mathcal{S}$.

- Empty set: Ø.

- Subset: $A \subset B$.

- Union: $A \cup B$.

- Intersection: $A \cap B$.

- Complement: A^C.

- Disjoint: $A \cap B = \emptyset$.

Probability Axioms

Axiom 1: For any set A,

$$P(A) \geq 0.$$

Axiom 2: $P(\mathcal{S}) = 1$.

Axiom 3: If $A_1, A_2, \ldots$ is any set of disjoint events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Further relationships

If we let $A = \cup_{i=1}^{\infty} A_i$, and $A_1, A_2, \ldots$ are disjoint, then $A_1, A_2, \ldots$ is said to be a *partition* of A.

We can also derive a number of results from these basic axioms:

- **Complements:** $P(A^C) = 1 - P(A)$.
- **Differences:** If A is contained in B (we write $A \subset B$), then

$$P(B \cap A^C) = P(B) - P(A).$$

- **Inclusion-Exclusion:**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Frequency interpretation

Consider a fair coin.
Let A be the event that you get a head on tossing the coin.
Then

$$P(A) = \frac{1}{2}.$$

In the frequentist interpretation of probability, this statement means that if we repeated the experiment many-times, then the long-run average number of heads would tend to $\frac{1}{2}$.

Experiment

	Heads	**Tails**
Total		
Average		

Equally likely outcomes

Sometimes, we can safely assume that outcomes are equally likely, as when we roll a fair dice or toss a fair coin twice.

If all outcomes are equally likely in a finite set $\mathcal{S}$, then the probability that event A occurs is:

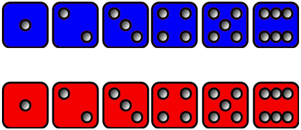
$$P(A) = \frac{n_A}{N},$$

where n_A is the number of sample points in A, and N is the number of sample points in $\mathcal{S}$.

Equally likely outcomes

Example 1.2 Two fair dice are rolled.

What is the probability of the sum of the numbers on the top of each die being less than 4?



◁

Counting method 1: MN Rule

Consider a set with m elements, $a_1, a_2, \ldots, a_m$, and a set with n elements, $b_1, b_2, \ldots, b_n$, then it is possible to form $m \times n$ pairs which contain 1 element from each set. **Example 1.3 Roll a die and select a card from a deck.**

What is the number of sample points in the sample space?

Counting method 2: Permutations

The number of ways of ordering n distinct objects taken r at a time is denoted as P_r^n, and

$$P_r^n = n(n-1)(n-2)\ldots(n-r+1) = \frac{n!}{(n-r)!},$$

where $n! = n(n-1)(n-2)\ldots 2 \times 1$ and $0! = 1$.

Example 1.4

How many ways can I select 3 people from the lecture theatre to stand in front of the board? ◁

Counting method 3: Partitioning

The number of ways of partitioning n distinct objects into k distinct groups containing $n_1, n_2, \ldots, n_k$, where $\sum_{i=1}^{k} n_i = n$, is

$$N = \binom{n}{n_1,\ n_2,\ \ldots,\ n_k} = \frac{n!}{n_1!n_2!,\ldots,n_k!}.$$

Example 1.5 (order doesn't matter here)

What is the number of ways of allocating 60 students into 3 tutorial groups of size 20? ◁

Counting method 4: Combinations

The number of ways of choosing r objects from n available objects without replacement is

$$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Example 1.6

At my favourite restaurant, what is the number of ways I can choose 3 main courses from the menu if there are 20 main dishes? ◁

Conditional probability

Definition 1.1: Conditional Probability

The *conditional probability* of an event A given that an event B has occurred, is equal to

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{provided } P(B) > 0.$$

Conditional probability

Example 1.6 Two fair dice are tossed.

Given the sum is even, what is the probability that both numbers are odd?

◁

Multiplication principle

Rearranging the rule for conditional probability gives

$$\begin{aligned} P(A \cap B) &= P(B|A)P(A) \\ &= P(A|B)P(B). \end{aligned}$$

Tree diagrams: These are a useful way to visualise the multiplication principle, where each branch represents a possible outcome.

Tree diagram

Example 1.7 A system has 2 electrical components.

The first component has a probability of failure of 0.1 (10% of the time).

- If the first component fails, the second fails with probability 0.2 (20% of the time).
- If the first works, then second fails with probability 0.05 (5% of the time).

Find the probability that

- At least one component works.
- Exactly one component works.
- The second component works.

Tree diagram

Tree diagram

Tree diagram

Tree diagram

Independence

Two events A and B are said to be *independent* if any one of the following holds:

- $P(A|B) = P(A)$,
- $P(B|A) = P(B)$,
- $P(A \cap B) = P(A)P(B)$.

Otherwise there are said to be *dependent*.

Example 1.8 Toss a fair coin two times.

Let A be the event that the two coins give the same result, and B be the event that the first coin is a head. Are A and B independent? $\triangleleft$

Independence

Example 1.8 Toss a fair coin two times.

Let A be the event that the two coins give the same result, and B be the event that the first coin is a head. Are A and B independent? $\triangleleft$

Independence

The definition of independence extends to collections of more than two events.

Definition 1.2:

A collection of n events $A_1, \ldots, A_n$ is independent if, for every collection of k events $A_{i_1}, \ldots, A_{i_k}$, for $k \leq n$, we have that

$$\begin{aligned} P\left(\bigcap_{j=1}^{k} A_{i_j}\right) &= P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) \\ &= P(A_{i_1})P(A_{i_2})\ldots P(A_{i_k}). \end{aligned}$$

Independence

Example 1.9 Three events

In the case of $n = 3$ events A, B and C, this condition requires that $\triangleleft$

Independence

Example 1.10

Consider a random experiment for which the sample space consists of four equally likely outcomes

$$\mathcal{S} = \{(1,0,0),\ (0,1,0),\ (0,0,1),\ (1,1,1)\},$$

and so $\mathrm{P}((1,0,0)) = \frac{1}{4},\ \mathrm{P}((0,1,0)) = \frac{1}{4}$, $\mathrm{P}((0,0,1)) = \frac{1}{4}$ and $\mathrm{P}((1,1,1)) = \frac{1}{4}$.
Let the events:

- E be the event that the first coordinate is 1,
- F be the event that the second coordinate is 1,
- G be the event that the third coordinate is 1.

Are E, F, and G independent events ? ◁

Independence

Independence

Law of total probability

Consider the example of the system with 2 electrical components given in example on [S1-28].

Let A be the event that the first component works and B the event that the second component works.

It can easily be seen that

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

Law of total probability

This can be extended and formalised as in the following theorem.

Theorem 1 (Law of Total Probability)

If $B_1, \ldots, B_n$ *is a partition of* S*, such that* $P(B_i) > 0$ *for* $i \in 1, \ldots, n$*, then*

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + \ldots + P(A|B_n)P(B_n) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i). \end{aligned}$$

Bayes' rule

We saw in the definition of conditional probability that

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A).$$

Provided $P(A) > 0$, $\Rightarrow P(B|A) = \dfrac{P(A|B)P(B)}{P(A)}$.

Then using Theorem 1, we have

Theorem 2 (Bayes' Rule)

If $B_1, \ldots, B_n$ is a partition of S, then

$$P(B_j|A) \quad = \quad \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}.$$

Bayes’ rule

Example 1.11 Rare diseases

Consider a rare disease which has a blood test to test for it.

- If a person has the disease then the probability of a positive test is 0.98, while
- if a person does not have the disease, then the probability of a negative test is 0.97.

If only 0.07% of the population has the disease, what is the probability that a randomly selected person who has a positive blood test has the disease? ◁

Bayes’ rule

Example 1.10 Rare diseases

◁

Odds

Another way of representing probabilities are odds. Given an event A with probability $P(A)$, then the *odds* in favour of A are defined as the ratio

$$\begin{aligned}\text{Odds}(A) &= \frac{P(A)}{P(A^C)} \\ &= \frac{P(A)}{1 - P(A)}.\end{aligned}$$

Example 1.11 A fair coin

Given a fair coin what is odds(A), where A is the event: obtain a head?

◁

Bayes' rule for odds

Bayes' rule can also be applied to odds to give

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

This can also be considered as

Posterior odds = Likelihood ratio x Prior odds.

Bayes’ rule for odds

Example 1.12 A lie detector test.

Let *B* be the event that a person is lying and *A* be the event that the lie detector says that the person is lying. Assuming that the probability a person is lying is 0.01,

- what are the prior odds that a person is lying?

It is known that if a person is lying, then the probability the machine will pick it up is 0.88, and if the person is telling the truth the probability that the machine will say they are truthful is 0.86.

- What is the likelihood ratio?
- What are the posterior odds that a person is lying?

◁

Bayes' rule for odds

Example 1.12 A lie detector test.

◁

Example 1.13 Motivating example

You are a medical consultant. A study of the residents of a region showed that 20% are smokers. The probability of death due to lung cancer, given that a person smokes, is ten times the probability of death due to lung cancer, given the person does not smoke. If the probability of death due to lung cancer in the region is 0.006, what is the probability of death due to lung cancer given that a person is a smoker? ◁

MATHS 2103 / MATHS 7103
Probability and Statistics II
Lecture notes

Section 2

Andrew Smith

School of Mathematical Sciences, University of Adelaide

Semester 1, 2017

Course outline

- **Section 01:** General information and introduction to Probability
- **Section 02:** Discrete random variables
- **Section 03:** Continuous random variables
- **Section 04:** Bivariate (multivariate) probability distributions
- **Section 05:** Discrete Time Markov Chains (DTMC)

Motivation example

Example 2.1

The number of bacteria colonies of a certain type in samples of polluted water has a Poisson distribution with a mean of 2 per cm^3.

1. If four 1-cm^3 samples are independently selected from this water, find the probability that at least one sample will contain one or more bacteria colonies.

2. How many 1-cm^3 samples should be selected in order to have a probability of approximately 0.95 of seeing at least one bacteria colony?

◁

Random variables

Recall that $\mathcal{S}$ is used to denote a sample space, whose elements x correspond to the possible outcomes of an experiment, so if we consider a function $Y(x)$ from the sample space $\mathcal{S}$ to the real number line $\mathbb{R}$:

$Y : \mathcal{S} \rightarrow \mathbb{R},$ then such a function is called a *random variable*.

A little more formally.

Definition 2.1: Random Variable

A real-valued **random variable** is a (measurable) function $Y(x)$ which maps $\mathcal{S}$ onto the real number line $\mathbb{R}$.

Random variables

Notation:

We use uppercase letters, e.g. Y, to represent random variables.

We use lowercase letters, e.g. y, to represent a particular value or realisation of Y.

Random variables

Example 2.2

Consider tossing a fair coin three times and let the random variable Y be the number of heads observed.

$\triangleleft$

State space

In Example 2.2 the set $Y(x)$ takes values in the set $\{0, 1, 2, 3\} \subseteq \mathbb{R}$, which leads to the following

Definition 2.2: State Space of a random variable

The **state space** of a random variable $Y(x)$ is the set $\mathcal{S}_Y \subseteq \mathbb{R}$ of all possible values that can be taken by $Y(x)$.

Probability measure

From here on, we abbreviate $Y = Y(x)$, etc. when convenient.

When the outcome of an experiment is known, a random variable Y takes some numerical value, which is often more likely to lie in certain subsets of the state space than others and we wish to describe the distribution of likelihoods of possible values of Y.

Probability mass

The random variable Y, counts the total number of “heads”, and so we can establish a probability measure for Y.

$$\begin{aligned} \mathrm{P}(Y=0) &= \mathrm{P}(TTT) = \frac{1}{8} \\ \mathrm{P}(Y=1) &= \mathrm{P}(HTT \cup THT \cup TTH) \\ &= \mathrm{P}(HTT) + \mathrm{P}(THT) + \mathrm{P}(TTH) = \frac{3}{8} \end{aligned}$$

$$\mathrm{P}(Y=2) = \frac{3}{8} \quad \text{and} \quad \mathrm{P}(Y=3) = \frac{1}{8}$$

Probability mass function

Definition 2.3: Probability mass function (PMF)

A discrete random variable Y has a **probability mass function** $f_Y : \mathcal{S}_Y \to [0, 1]$, given by

$$f_Y(y_i) = \mathrm{P}(Y = y_i) = \mathrm{P}(x \in \mathcal{S} : Y(x) = y_i)\,,$$

for each i such that $\sum_i f_Y(y_i) = 1$.

Properties of the PMF

The *probability mass function* (PMF) is a formula, a table, or a graph, that gives $f(y_i) = \mathrm{P}(Y = y_i)$ for all possible values of $Y = y_i$.

To be a valid PMF, the following properties must hold:

- $0 \leq f(y_i) \leq 1$ for all i,
- $\sum_i f(y_i) = 1$.

Valid or invalid?

Example 2.3

The following table claims to be the PMF for the random variable X.

i	1	2	3	4
x_i	0	1	4	12
$P(X = x_i)$	0.1	0.01	0.88	0.02

Is this PMF valid? ◁

Probability distribution function

An alternative way to characterise a probability distribution is by its distribution function.

Definition 2.4: Probability distribution function

The **distribution function** of a random variable Y is the function $F_Y : \mathbb{R} \to [0, 1]$, given by

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(x \in \mathcal{S} : Y(x) \leq y)\,.$$

Probability distribution function

Distribution functions become particularly important in the study of continuous random variables.

However, for most applications involving discrete random variables, it is more convenient to utilise the probability mass function.

Probability distribution function

Recall in Example 2.2, the random variable Y which counted the number of heads had the following PMF

$$f(0) = \frac{1}{8},\ f(1) = \frac{3}{8},\ f(2) = \frac{3}{8} \text{ and } f(3) = \frac{1}{8}.$$

Therefore the cumulative probability distribution function of the random variable Y is

Expected values

Definition 2.5: Expectation

The **expectation** (or expected value) of a discrete random variable Y with probability mass function $f(y)$ is given by

$$\mathbb{E}[Y] = \sum_i y_i f(y_i)\,, \tag{2.1}$$

provided that this sum is absolutely convergent.

That is, $\sum_i |y_i| f(y_i) < \infty$.

A function of a r.v.

If Y is a discrete random variable with PMF $f(y)$, and $g(Y)$ is a real-valued function of Y, then $g(Y)$ is also a discrete random variable with expected value given by

$$\mathbb{E}[g(Y)] = \sum_i g(y_i) f(y_i) \,,$$

provided the sum is again absolutely convergent.

Example 2.4

Find $\mathbb{E}\left[Y^2\right]$, given that the random variable Y has the PMF

y	-1	0	1
$P(Y = y)$	1/4	1/2	1/4

◁

A function of a discrete r.v.

Moments

Definition 2.6: Moments and central moments

For $k \in \mathbb{Z}^+$, the k^{th} **moment** m_k of Y is given by

$$m_k = \mathbb{E}\left[Y^k \right].$$

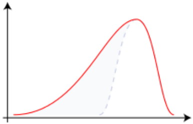
The k^{th} **central moment** σ_k is

$$\sigma_k = \mathbb{E}\left[(Y - m_1)^k \right]. \tag{2.2}$$

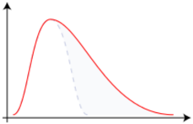
More moments

Sometimes we observe particular types of asymmetry, such as peakedness (kurtosis) and spread (skewedness).

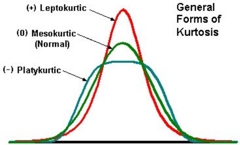
- If the curve has one tail that is longer than the other, it is skewed. If the longer tail is on the left, it is called negatively skewed, whereas if the longer tail is on the right, it is called positively skewed.
- If the curve is “too peaked” to be normal, it is called leptokurtic .
- If it is “too flat”, it is called platykurtic.



Negative Skew



Positive Skew



Variance

When obtaining the variance (or second central moment) of a r.v. Y, it is often useful to use the fact that

$$\sigma_2 = m_2 - (m_1)^2, \; \left(\text{that is, } \operatorname{var}(Y) = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 \; \right).$$

Linear functions

The expectation operator, $\mathbb{E}[\cdot]$ is a linear operator as shown in the following theorem.

Theorem 2.1 (Properties of $\mathbb{E}[\cdot]$)

For random variables X and Y, real constants $a, c_i \in \mathbb{R}$ and real functions $g(\cdot), g_i(\cdot)$ we have that

- $\mathbb{E}[a] = a$,
- $\mathbb{E}[aY] = a\mathbb{E}[Y]$,
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$,
- $\mathbb{E}[cg(Y)] = c\mathbb{E}[g(Y)]$.
- $\mathbb{E}[\sum_i c_i g_i(Y)] = \sum_i c_i \mathbb{E}[g_i(Y)]$.

Linear functions

An alternative way to prove the variance result is by making use of the fact that taking the expectation of a r.v. is a linear operation.

Bernoulli distribution

Definition 2.7: Bernoulli random variable

Consider a single trial with two outcomes, success (1) or failure (0). If the probability of success is p, then the PMF is

y	0	1
$P(Y = y)$	$1 - p$	p

A random variable, Y, with this PMF is said to have a *Bernoulli* distribution. Note that the random variable Y is said to have support on $\{0, 1\}$.

What is the expected value, $\mathbb{E}[Y]$, and variance, $\text{var}(Y)$, of a Bernoulli random variable Y?

Bernoulli distribution

Bernoulli distribution

Bernoulli random variables arise frequently as indicators of events.

The **indicator** of an event A is the random variable

$$1_A \;=\; \begin{cases} 1 & \text{if A occurs} \\ 0 & \text{if A does not occur} \end{cases}$$

Then 1_A is a Bernoulli random variable with parameter p, which has the convenient property that

$$E[1_A] \;=\; 0 \times (1 - p) + 1 \times p \;=\; p \;=\; \mathrm{P}(A)\,.$$

This property is sometimes useful for simplifying certain calculations which would otherwise be quite complicated.

Bernoulli trials

Consider a sequence of trials that are independent and identically distributed (iid) such that each trial has only two possible outcomes: success or failure.

By identically distributed, we mean that each trial has the same probability of success, and independence means that the outcome of one trial does not influence the outcome of the other trials.

Geometric distribution

Definition 2.8: Geometric Distribution

Consider a sequence of independent Bernoulli trials with the same probability of success p.
Let Y be the total number of trials until the first success.
It is easy to see that the PMF of Y is

$$f(y) = (1-p)^{y-1}p, y = 1, 2, \dots$$

A random variable Y with this PMF is said to have a *geometric* distribution. Note that Y has support on $\mathbb{Z}^+$.
We write $Y \sim \text{Geo}(p)$.

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

Geometric distribution

Geometric distribution

Geometric distribution

Binomial distribution

Definition 2.9: Binomial Distribution

Consider a *fixed* number of independent Bernoulli trials, n, each with the same probability of success, p.
Let Y be the number of successes, then the PMF for Y is

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, y = 0, 1, 2, \ldots, n.$$

A random variable with this PMF is said to have a *binomial* distribution. Note that Y has support on $\{0, 1, \ldots, n\}$.

We denote it as

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

Binomial distribution

Example 2.5

Toss a coin 10 times, let Y be the number of heads. What is

- $f(0)$?
- $f(5)$?
- $P(Y \geq 2)$?

$\triangleleft$

Binomial distribution

Binomial distribution

The previous example gives us the key to establishing whether a random variable is binomial or not.

For a discrete random variable to be a Binomial random variable, then the following must be true:

Binomial r.v.

If Y is a binomial random variable with n trials and probability of success p, then

- $\mathbb{E}[Y] = np$, and
- $\text{var}(Y) = np(1 - p)$.

Binomial r.v.

Random sampling

Consider a population of size *N*.

Let each item in the population be classified into one of two classes, success or failure.

If *r* of the population are classified as success, then if we randomly sample an item from this population, the probability of obtaining a success is $p = r/N$.

If we randomly select a second item from the population, then the probability of a success depends on our sampling method.

Random sampling with replacement

In this case, we return the item to the population before sampling the second item.

Thus the original item could be sampled again.

Therefore, the probability of a success is

Random sampling no replacement

In this case, we do not replace the original item after sampling.

The probability of the second item being a success depends on the first item.

Let S_i be the event that the ith item
is a success and F_i be the event that the ith item is a failure, then

Hypergeometric distribution

Definition 2.10: Hypergeometric distribution

Consider a population of size N containing r successes and $N - r$ failures.

If we randomly sample n items without replacement, then the number of successes, Y, has the PMF,

$$f(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}.$$

A random variable with this PMF is said to have a *hypergeometric* distribution.

Hypergeometric distribution

Example 2.6

Draw five cards from a standard deck of cards, what is the probability of 5 clubs? What is the probability of 4 aces?

◁

Hypergeometric distribution

If Y has a hypergeometric distribution, with a sample size n, and r successes in a population of size N, then

- $\mathbb{E}[Y] = \frac{rn}{N}$
- $\text{var}(Y) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$.

Proof: Omitted.

Hypergeometric distribution

Approximating the hypergeometric with the binomial

It can be shown that

$$f_Y(y) = \mathrm{P}(Y = y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} \approx \binom{n}{y} p^y (1-p)^{n-y},$$

where $p = \frac{r}{N}$ and r and N are large in relation to y and n.

Consider the following: $\mathrm{Bin}(n, p)$, where $n = 10$ and $p = \frac{1}{2}$.

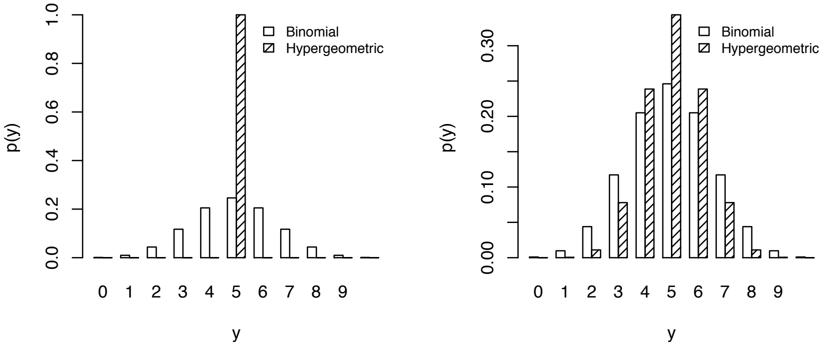


Figure 2.1: (a) $r = 5, N = 10$ (b) $r = 10, N = 20$

Sample size $n = 10$

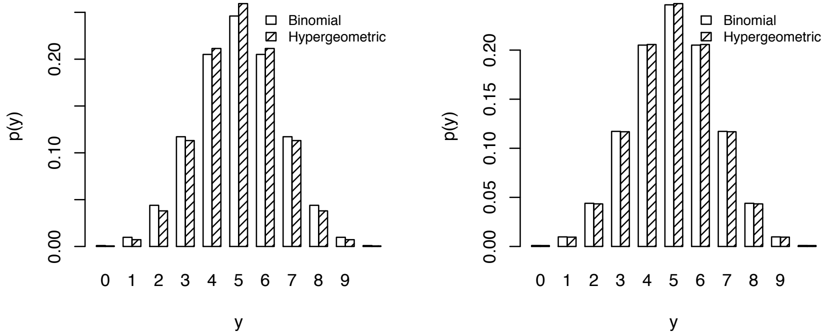


Figure 2.2: (c) $r = 50, N = 100$ (d) $r = 500, N = 1000.$

Sample size $n = 10$

The Poisson distribution

Let the random variable Y be the number of cases of a particular disease in Australia in a year. If the expected number of cases per year is λ, what is the PMF? We could treat this as a binomial, i.e. consider each month as a trial with success being a case and failure no case. This gives

$$Y \sim \text{Bin}(12, \lambda/12), \ \ \text{only has support on } \{0, 1, \ldots, 12\},$$

but there is always the chance of 2 or more cases in a month.

So we could consider each week as a trial giving

$$Y \sim \text{Bin}(52, \lambda/52), \ \ \text{only has support on } \{0, 1, \ldots, 52\}$$

but still two cases or more could occur in a period. So we could try day, hour, minute,. . .

The Poisson distribution

Consider the PMF of the binomial as *n* approaches infinity, *p* approaches zero, and *np* remains equal to λ, then

$$\mathrm{P}(Y = y) = \lim_{n \to \infty} \binom{n}{y} \left(\frac{\lambda}{n} \right)^y \left(1 - \frac{\lambda}{n} \right)^{n-y} = \frac{e^{-\lambda} \lambda^y}{y!}.$$

The Poisson distribution

The Poisson distribution

Definition 2.11: Poisson distribution

Let Y be the number of events occurring in a time period or in a region of space with a rate of λ, then

$$f_Y(y) = \mathrm{P}(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, y = 0, 1, \ldots$$

A random variable with this PMF is said to have a *Poisson* distribution.

The Poisson distribution

An alternative method of derivation is to consider a time interval t broken into small subintervals of length $\triangle t$. If λ is the arrival rate of events per unit time, then assume that

$$\begin{aligned} q_1(\triangle t) &= \lambda \triangle t + o(\triangle t) \\ q_0(\triangle t) &= 1 - \lambda \triangle t + o(\triangle t) \\ q_i(\triangle t) &= o(\triangle t) \quad \text{for all } i \geq 2, \end{aligned}$$

where $q_n(\triangle t) = \mathrm{P}(n \text{ events in } (t, t + \triangle t))$ and the function $o(\triangle t)$ is such that $\lim_{\triangle t \to 0} \frac{o(\triangle t)}{\triangle t} = 0$.

Let $p_n(t) = \mathrm{P}(n \text{ events in } (0, t))$, for $n = 0, 1, 2, 3, \ldots,$ then

$$p_0(t + \triangle t) \;=\; p_0(t) q_0(\triangle t) \;=\; p_0(t)\big(1 - \lambda \triangle t + o(\triangle t)\big).$$

The Poisson distribution

Hence rearranging, we divide by $\triangle t$ to get

$$\frac{p_0(t+\triangle t)-p_0(t)}{\triangle t} \quad = \quad p_0(t)\left(\frac{o(\triangle t)}{\triangle t}-\lambda\right)$$

Then letting $\triangle t \to 0$, we see that

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -\lambda p_0(t) \\ \implies p_0(t) &= Ce^{-\lambda t} \end{aligned}$$

However ,we have that $p_0(0)=1$, which implies that

$$p_0(t) \quad = \quad e^{-\lambda t}.$$

The Poisson distribution

Similarly we have that

$$
\begin{aligned}
p_n(t+\triangle t) &= p_n(t)q_0(\triangle t) + p_{n-1}(t)q_1(t) + \sum_{k=0}^{n-2} p_k(t)q_{n-k}(\triangle t) \\
&= p_n(t)\Big(1 - \lambda \triangle t + o(\triangle t)\Big) \\
&\qquad + p_{n+1}(t)\Big(\lambda \triangle t + o(\triangle t)\Big) + o(\triangle t)\sum_{k=0}^{n-2} p_k(t).
\end{aligned}
$$

Rearranging, we divide by $\triangle t$ to get

$$
\begin{aligned}
\frac{p_n(t+\triangle t) - p_n(t)}{\triangle t} &= p_n(t)\left(\frac{o(\triangle t)}{\triangle t} - \lambda\right) \\
&\qquad + p_{n-1}(t)\left(\lambda + \frac{o(\triangle t)}{\triangle t}\right) + \frac{o(\triangle t)}{\triangle t}\sum_{k=0}^{n-2} p_k(t).
\end{aligned}
$$

The Poisson distribution

Then letting $\triangle t \to 0$, we see that

$$\begin{aligned} \frac{dp_n(t)}{dt} &= -\lambda p_n(t) + \lambda p_{n-1}(t) \\ \Longrightarrow \quad \frac{dp_1(t)}{dt} &= -\lambda p_1(t) + \lambda p_0(t) \ = \ -\lambda p_1(t) + \lambda e^{-\lambda t} \\ \Longrightarrow \quad p_1(t) &= \lambda t e^{-\lambda t} \quad \text{(using integrating factor)} \end{aligned}$$

By induction we then can show that $p_n(t) = \dfrac{(\lambda t)^n e^{-\lambda t}}{n!}$.
Hence per unit time we have

$$p_y(1) = \mathrm{P}(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \ \text{for all } y = 0, 1, 2, \ldots$$

The Poisson distribution

Example 2.7

Let Y be the number of car accidents on a Australian road. It is known that the rate of accidents is 4 per month. What is the probability of no accidents in a month? What is the probability of more than 3 accidents in a month? $\triangleleft$

The Poisson distribution

Mean and variance of Poisson distribution

If Y has a Poisson distribution with a rate of λ, then

- $\mathbb{E}[Y] = \lambda$.
- $\text{var}(Y) = \lambda$.

The Poisson distribution

Lemma 2.2

Bounding Probabilities

Bounding probabilities for discrete random variables

In some cases, we may not know the PMF of the process we are modelling, but we may only know or in some way can get an estimate of the mean, or both the mean and variance of the probability distribution of interest.

Tail sum formula

Theorem 2.2 (Tail Sum Formula)

Let Y be a discrete random variable that takes the values $0, 1, 2, \ldots, n$, then

$$\mathbb{E}[Y] = \sum_{i=1}^{n} P(Y \geq i).$$

Proof:

Markov's inequality

Theorem 2.3 (Markov's Inequality)

If Y is a random variable that takes only nonnegative values, then for any value $a > 0$,

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

Example 2.8

Suppose the average cost to maintain a car for a year is \$1500, what is the upper bound on the probability that the cost in one year is greater than \$7500? ◁

Markov's inequality

Chebyshev's inequality

Theorem 2.4 (Chebyshev's inequality)

Let Y be a random variable with expected value μ and finite variance σ^2. Then for any constant $k > 0$,

$$\begin{array}{lrcl} & P(|Y-\mu| < k\sigma) & \geq & 1 - \dfrac{1}{k^2} \\ \textit{or} & P(|Y-\mu| \geq k\sigma) & \leq & \dfrac{1}{k^2}. \end{array}$$

Example 2.9

The number of customers at a shop in a day has a mean of 20 with a variance of 16. What is a lower bound on the probability that the number of customers will lie between 12 and 28?

Chebyshev’s inequality

Chebyshev’s inequality

Moment generating function

Definition 2.12: Moment generating function (mgf)

The *moment generating function* $m(t)$ of a random variable Y is defined to be

$$m(t) \quad = \quad \mathbb{E}\left[e^{tY}\right]. \tag{2.3}$$

Moment generating function

Example 2.10

Show the moment generating function of Y where Y is a Poisson random variable with rate λ is

$$m(t) \quad = \quad e^{\lambda(e^t - 1)}.$$

$\triangleleft$

Moment generating function

Properties of the moment generating function

Theorem 2.5 (Moments from the mgf)

If the moment generating function, $m(t)$, exists for a random variable Y, then for any positive integer k

$$
\begin{aligned}
m^{(k)}(0) &= \left.\frac{d^k m(t)}{dt^k}\right|_{t=0} \\
&= \left.\frac{d^k \mathbb{E}\left[e^{tY}\right]}{dt^k}\right|_{t=0} = E[Y^k].
\end{aligned}
$$

Moment generating function

Moment generating function

Example:

Let Y be a Poisson random variable with rate λ. Using the moment generating function show that

- $\mathbb{E}[Y] = \lambda$.
- var$(Y) = \lambda$.

Probability generating functions (pgf)

The pgf of a discrete random variable is a power series representation of the probability mass function of the random variable. Probability generating functions are also known as Z-transforms.

Definition 2.13:

If $f_Y(i)$ $i = 0, 1, 2, \ldots$ is a probability distribution for a random variable Y (it's pmf), then its **probability generating function** is

$$P(z) \;=\; \mathbb{E}\Big[z^Y\Big] \;=\; \sum_{i=0}^{\infty} f_Y(i) z^i, \qquad \text{for } z \in [0, 1].$$

Probability generating functions

Probability generating functions are another way of representing the information contained in a probability distribution.

Example 2.11

Find the probability generating function for Y, where Y is a geometric random variables with probability of success p.

◁

Probability generating functions

Properties of the probability generating function

1. $P(1) = 1$.

2. $P(0) = P(Y = 0)$.

3. $\left.\frac{d^k P(z)}{dz^k}\right|_{z=0} = k! P(Y = k)$.

4. $\left.\frac{d^k P(z)}{dz^k}\right|_{z=1} = \mathbb{E}[Y(Y-1)(Y-2)\ldots,(Y-K+1)]$.

Probability generating functions

Example 2.12

Let Y be a Poisson random variable with rate λ. Show the probability generating function is

$$P(z) = e^{\lambda z} e^{-\lambda}.$$

◁

Probability generating functions

Example 2.13

Using the probability generating function for a Poisson random variable with rate λ, show that

$$\mathbb{E}[Y] = \lambda \quad \text{and} \quad \text{var}(Y) = \lambda.$$

$\triangleleft$

Motivation example

Example 2.1

The number of bacteria colonies of a certain type in samples of polluted water has a Poisson distribution with a mean of 2 per cm^3.

1. If four 1-cm^3 samples are independently selected from this water, find the probability that at least one sample will contain one or more bacteria colonies.

2. How many 1-cm^3 samples should be selected in order to have a probability of approximately 0.95 of seeing at least one bacteria colony?

◁

Motivation example

MATHS 2103 / MATHS 7103
Probability and Statistics II
Lecture notes

Section 3

Andrew Smith

School of Mathematical Sciences, University of Adelaide

Semester 1, 2017

Course outline

- **Section 01:** General information and introduction to Probability

- **Section 02:** Discrete random variables

- **Section 03:** Continuous random variables

- **Section 04:** Bivariate (multivariate) probability distributions

- **Section 05:** Discrete Time Markov Chains (DTMC)

Motivation example

Example 3.1

The time (in hours) a manager takes to interview a job applicant has an exponential distribution with $\lambda = 2$. The applicants are scheduled at quarter-hour intervals, beginning at 8:00 a.m., and the applicants arrive exactly on time. When the applicant with an 8:15 a.m. appointment arrives at the manager's office, what is the probability that he or she will have to wait before seeing the manager? $\triangleleft$

Continuous random variable

In some situations, the quantity we are measuring is not discrete, e.g. time until an event, yield of a crop, height of students.

How can we calculate probabilities for these cases?

For discrete random variables, we had a probability *mass* function, $f(x)$. It had an intuitive interpretation; $f(x) = \mathrm{P}(X = x)$. For continuous random variables $\mathrm{P}(X = x) = 0$. So we need a new function: a probability *density* function.

Probability density function

Definition 3.1:

Probability density function (PDF) A continuous random variable Y has a **probability density function** $f_Y : \mathcal{S}_Y \to [0, \infty)$, given by

$$\int_a^b f_Y(x)dx = P(a \leq X \leq b),$$

for some $a, b \in \mathcal{S}_Y$.

Cumulative distribution function

Recall the definition of the cumulative probability distribution function (CDF) in the last section, defined as

$$F(y) = P(Y \leq y), -\infty < y < \infty.$$

Cumulative distribution function

Properties of the cumulative distribution function

If $F(y)$ is a cumulative distribution function, then the following properties hold:

- $\lim_{y \to -\infty} F(y) = 0$.
- $\lim_{y \to \infty} F(y) = 1$.
- $F(y)$ is a non-decreasing function of y, i.e., if y_1 and y_2 are any values such that $y_1 < y_2$, then $F(y_1) \leq F(y_2)$.

Continuous random variables

It turns out that the nature of the CDF determines whether we call a random variable discrete or continuous.

Definition 3.2: CDF of a continuous random variable

A random variable Y with CDF $F(y)$ is said to be continuous if $F(y)$ is continuous for all y.

Probability density function

Definition 3.3: Density function

Let $F(y)$ be the cumulative distribution function for a continuous random variable Y. Then $f(y)$ given by

$$f(y) = F'(y) = \frac{dF(y)}{dy}$$

wherever the derivative exists, is called the *probability density function* (pdf) for the random variable Y.

Properties of the pdf

It Follows from definitions 3.2, 3.3 that the distribution function $F(y)$ of a random variable Y can be written as

$$F(y) \quad = \quad \int_{-\infty}^{y} f(t)dt,$$

where $f(\cdot)$ is the probability density function of the random variable Y.

If $f(y)$ is the pdf of a continuous random variable Y, then to be valid the following must hold

- $f(y) \geq 0$ for all $y, -\infty < y < \infty$,
- $F(\infty) = \int_{-\infty}^{\infty} f(y)dy = 1$.

Properties of the pdf

Example 3.2

Consider the random variable Y with pdf

$$f(y) = \begin{cases} 3y^2, & 0 \leq y \leq 1 \\ 0, & \text{elsewhere.} \end{cases}$$

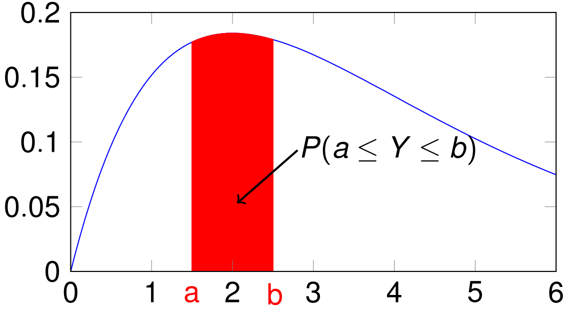
Is $f(y)$ a valid pdf?

$\triangleleft$

Calculating probabilities

Let Y be a random variable with pdf $f(y)$, then

$$P(a \le Y \le b) = F(b) - F(a) = \int_a^b f(y)dy.$$



Calculating probabilities

Example 3.3

Consider the random variable Y with pdf

$$f(y) = \begin{cases} 3y^2, & 0 \leq y \leq 1 \\ 0, & \text{elsewhere.} \end{cases}$$

What is $P(1/4 < Y < 1/2)$? $\triangleleft$

Expectation

Definition 3.4: Expectation of a continuous r.v.

Let Y be a continuous random variable with pdf $f(y)$, then the *expected value* of Y, $\mathbb{E}[Y]$, is defined as

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y\, f(y) dy,$$

provided this integral is absolutely convergent. That is,

$$\int_{-\infty}^{\infty} |y| f(y) < \infty.$$

Expectation

Example 3.4

Consider the random variable Y with pdf

$$f(y) = \begin{cases} 3y^2, & 0 \leq y \leq 1 \\ 0, & \text{elsewhere.} \end{cases}$$

What is $\mathbb{E}[Y]$? ◁

Expectation

Definition 3.5: Expectation of a f^{n} of a continuous r.v.

Let Y be a continuous random variable with pdf $f(y)$, and let $g(Y)$ be a real-valued function of Y, then the expected value of $g(Y)$ is given by

$$\mathbb{E}[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy,$$

provided the sum is absolutely convergent. That is,

$$\int_{-\infty}^{\infty} |g(y)|f(y) < \infty.$$

Expectation

Example 3.5

Consider the random variable Y with pdf

$$f(y) = \begin{cases} 3y^2, & 0 \leq y \leq 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Let $X = e^Y$, what is the expected value of X? ◁

Variance

Definition 3.6: Variance of a continuous random variable

If Y is a random variable with expected value $\mathbb{E}[Y] = \mu$, then the *variance* of Y is defined as

$$\begin{aligned} \text{var}(Y) &= \mathbb{E}\Big[(Y-\mu)^2\Big] \\ &= \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\,dy. \end{aligned}$$

The *standard deviation* of Y is the positive square root of var(Y).

Moment generating functions

Definition 3.7: mgf of a continuous random variable

The *moment generating function* (mgf) $m(t)$ of a continuous random variable Y is defined to be

$$m(t) = \mathbb{E}\left[e^{tY}\right]. \tag{3.1}$$

Moment generating functions

Example 3.6

Consider the random variable Y with pdf

$$f(y) = \begin{cases} 3y^2, & 0 \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

What is the moment generating function $m(t)$ of Y? ◁

Uniqueness

Theorem 3.1 (Uniqueness of moment generating functions)

Let $F_X(x)$ and $F_Y(y)$ be two cumulative distribution functions all of whose moments exist.

If the moment generating functions exist and

$$m_X(t) = m_Y(t),$$

for all t in some neighbourhood of 0, then for all u,

$$F_X(u) = F_Y(u).$$

Proof: Omitted.

Chebyshev's inequality

Theorem 3.2 (Chebyshev's inequality)

Let Y be a random variable with expected value μ and finite variance σ^2.
Then for any constant $k > 0$,

$$P(|Y - \mu)| < k\sigma) \quad \geq \quad 1 - \frac{1}{k^2}$$

$$\text{or} \quad P(|Y - \mu| \geq k\sigma) \quad \leq \quad \frac{1}{k^2}.$$

Proof of Chebyshev’s inequality

The Uniform distribution

Definition 3.8: The Uniform distribution

If $a < b$, a random variable Y is said to have a continuous *uniform* distribution on the interval (a, b) if and only if the density function of Y is

$$f(y) = \begin{cases} \frac{1}{b-a} & a \leq y \leq b \\ 0 & \text{elsewhere.} \end{cases}$$

This is denoted as $Y \sim U(a, b)$.

The Uniform distribution

Example 3.7 $Y \sim U(0, 1)$

Let Y be a random variable with a Uniform distribution on the interval $(0, 1)$. What is

1. $P(0 \leq Y \leq \alpha)$,   for $\alpha \in [0, 1]$?
2. $\mathbb{E}[Y]$?
3. $\text{var}(Y)$?

$\triangleleft$

Expectation and Variance

Mean and variance of the Uniform distribution

If $Y \sim U(a, b)$, then what is

- $\mathbb{E}[Y]$?
- $\text{var}(Y)$?
- $m_Y(t)$?

The Normal distribution

Definition 3.9: The Normal distribution

A random variable Y is said to have a *Normal probability distribution* if and only if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the pdf of Y is

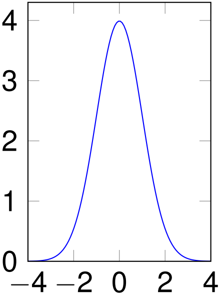
$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)}, -\infty < y < \infty.$$

This is denoted as $Y \sim N(\mu, \sigma^2)$.
If Y has a expected value of 0 and a variance of 1, then Y is said to have a *standard Normal distribution* $\sim N(0, 1)$.

Often the letter Z is used to denote the standard normal.

Calculating Normal probabilities

To calculate normal probabilities requires numerical integration, because it is not possible to integrate the standard Normal density function in terms of elementary functions.



Finding P($Z \leq z$)

Finding P($Z \leq z$)

Calculating Normal probabilities

Example 3.8

If $Y \sim N(4, 16)$, what is

1. $\mathrm{P}(0 \leq Y \leq 8)$?

2. $\mathrm{P}(Y \geq 1)$?

3. The value of a such that $\mathrm{P}(Y \leq a) = 0.25$.

◁

Calculating Normal probabilities

Calculating Normal probabilities

Calculating Normal probabilities

Mean, variance and mgf of a Normal distribution

Theorem 3.3

If Y is a random variable such that $Y \sim N(\mu, \sigma)$, then

- $\mathbb{E}[Y] = \mu$.
- $var(Y) = \sigma^2$.
- $m_Y(t) = e^{\mu t + (\sigma^2 t^2)/2}$.

Expectation of a Normal $(N(\mu, \sigma))$ r.v.

Variance of a Normal $(N(\mu, \sigma))$ r.v.

mgf of a Normal $\left(N(\mu, \sigma)\right)$ r.v.

The Cauchy distribution

Consider a random variable Y with pdf

$$f(y) = \frac{1}{\gamma\pi(1 + (\frac{y-\theta}{\gamma})^2)}, -\infty < y, \theta < \infty \text{ and } \gamma > 0.$$

A random variable with this pdf is said to have a *Cauchy distribution*. The Cauchy distribution is also a bell-shaped distribution symmetric around θ. When $\theta = 0$ and $\gamma = 1$ we get the standard Cauchy distribution.

Notes:

- The expected value does not exist.

- The ratio of two independent standard normal random variables is a Cauchy random variable with $\theta = 0$.

The Exponential distribution

Consider modelling the time until some event.

For example

- The survival time for cancer patients.
- The time to decay for radioactive atoms.
- The time to failure for a electronic component.

These have been successfully modelled by the *exponential distribution*, based on a parameter that describes the rate of events per unit time.

The Exponential distribution

Definition 3.10: The exponential probability distribution

Consider a random variable Y with probability density function

$$f(y) = \begin{cases} \lambda e^{-\lambda y}, & 0 \leq y < \infty, \lambda > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

A random variable with this pdf is said to have an *exponential distribution*, with rate parameter λ

Moments

Mean and variance of the exponential distribution

If Y has an exponential distribution with rate parameter λ, then

- $\mathbb{E}[Y] = \frac{1}{\lambda}$.
- $\text{var}(Y) = \left(\frac{1}{\lambda}\right)^2$.
- $m_Y(t) = \frac{\lambda}{\lambda - t}$.

Proof: Exercise

Memoryless property

Memoryless property of the exponential distribution

Let Y be an exponential random variable with rate λ, then for positive constants, a, b

$$P(Y > a + b | Y > a) = P(Y > b),$$

That is, the probability of surviving a further b units of time given you have already survived a units of time is the same as surviving b units of time.

Exponential distribution

Example 3.9

Let Y be the time for a hard drive to fail, known to be exponentially distributed with mean time to failure of 24 months. Calculate the probability that

- the hard drive will last less than 1 year.
- the hard drive will last more than 3 years.

Calculate the median time to failure. $\triangleleft$

Hazard function

Definition 3.11: Hazard function

The *hazard* function, $h_Y(y)$ is the conditional probability of failure in the interval $y + \delta$, (given that time has reached y) divided by the length of the interval δ as $\delta \to 0$.

$$h_Y(y) = \lim_{\delta \to 0} \frac{P(y \leq Y \leq y + \delta | Y \geq y)}{\delta},$$

which can be shown to be equal to $\dfrac{f(y)}{1 - F(y)}$.

Erlang distribution

Definition 3.12: The Erlang Distribution

The **Erlang distribution** of order n models the distribution of time until a sequence of n exponentially generated events with common rate parameter λ have occurred. Its density function is given by

$$f(y) = \begin{cases} 0 & y < 0 \\ \lambda e^{-\lambda y} \dfrac{(\lambda y)^{n-1}}{(n-1)!} & y \geq 0, \ \text{ with } \ n, \lambda > 0. \end{cases} \quad .$$

Its cumulative probability distribution function can be derived by integrating this, but is very messy to write down.

Erlang distribution

The expected value and variance of an Erlang distributed random variable Y are

$$\mathbb{E}[Y] \;=\; \frac{n}{\lambda} \quad \text{and} \quad \mathrm{var}(Y) \;=\; \frac{n}{\lambda^2}$$

$$\text{with} \;\; m_Y(t) \;=\; \frac{1}{\left(1 - \frac{t}{\lambda}\right)^n} \, .$$

Erlang distribution

Gamma distribution

Definition 3.13: Gamma distribution

A random variable Y is said to be Gamma distributed if it has the following density function

$$f(y) = \begin{cases} 0 & y < 0 \\ \lambda e^{-\lambda y} \frac{(\lambda y)^{\alpha-1}}{\Gamma(\alpha)} & y \geq 0, \text{ with } \alpha, \lambda > 0. \end{cases} .$$

where $\Gamma(\alpha)$ is the gamma function given by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy.$$

Properties of the gamma function

The gamma function $\Gamma(\alpha)$ has the following properties

- $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \alpha > 0$.
- $\Gamma(1) = 1$.
- If α is an integer n, then $\Gamma(n) = (n - 1)!$

Mean and Variance

Mean and variance of the gamma distribution

If Y is random variable with a gamma distribution with parameters α and λ, then similar to the Erlang distribution except in generality with $\alpha \in \mathbb{R}^+$

- $\mathbb{E}[Y] = \frac{\alpha}{\lambda}$.
- $\mathrm{var}(Y) = \frac{\alpha}{\lambda^2}$.
- $m_Y(t) = \frac{1}{\left(1-\frac{t}{\lambda}\right)^\alpha}$.

The Chi-square distribution

Definition 3.14: Chi-square distribution

Consider a gamma random variable Y with parameters $\alpha = \nu/2$ and $\lambda = 1/2$.
The pdf of Y is

$$f(y) = \begin{cases} \frac{y^{\left(\frac{\nu}{2}\right)-1} e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} & 0 \leq y < \infty, \nu > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

A random variable with this pdf is said to have a chi-square (χ^2) distribution with ν degrees of freedom.

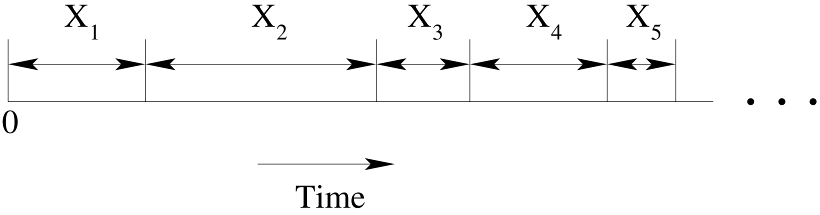
The Chi-square distribution

Point process

Consider events that happen at random moments in continuous time. For example, people arriving at a check out, lightning strikes, outbreaks of disease.

Sequences of events like this are called a *point process*.

We can count the number of events up to and including the time t denoted as $N(t)$ and we can also find the distribution of the time $S_n = X_1 + X_2 + \cdots + X_n$ taken until the n^{th} event.



The Poisson process

Assuming that the time between events $i-1$ and i given by X_i are iid *exponential* with rate λ, then using the fact that

$$\{N(t) \geq n\} = \{S_n \leq t\},$$

where S_n is the time to the nth event, then it can be shown that

Properties of the Poisson process

Independent increments:

Fix any set of successive times $0 = \tau_0 < \tau_1 < \ldots < \tau_n$, and define the number of events in time interval $\tau_i - \tau_{i-1}$ as

$$Y_i = N(\tau_i) - N(\tau_{i-1}),$$

then the random variables (or increments) $Y_1, Y_2, \ldots, Y_n$ are mutually independent.

Properties of the Poisson process

Stationary increments or shift invariance:

For any $s \geq 0$ and $h \geq 0$, the random variables (or increments) which are the number of events in time periods $s < t \leq s + h$ defined as

$$N(s+h) - N(s),$$

are stationary.

Call centre

Example 3.10 Helpline

Customers phone a help line according to a Poisson process with a rate of 3 per minute.
Calculate the probability that

- no calls arrive in the first 2 minutes.
- the first call arrives after 2 minutes.
- no calls arrive in the first 2 minutes and at most four calls arrive between $t = 2$ and $t = 3$.
- the fourth call arrives within 30 seconds of the third call.
- the time to the fifth call is greater than 2 minutes. ◁

Call centre

Example 3.10 Time between calls is $\sim \mathit{exp}(3)$

Call centre

Example 3.10 Time between calls is $\sim exp(3)$ (cont).

Transformations of random variables

Let Y be a random variable and let $U(Y)$ be a function of Y which we will denote as the random variable U.

What is the distribution of U?

Three methods we can employ

1. **Method of cumulative distribution functions.**
2. **Method of transformations.**
3. **Method of moment generating functions (see later).**

CDF method

Definition 3.15: Method of Cumulative Distribution functions

Let U be a function of the random variable Y, then the steps to find the density of U are

1. Find the region $U = u$ in the y space.
2. Find the region $U \leq u$.
3. Find the cumulative distribution function $F_U(u)$ of U, $P(U \leq u)$ by integrating $f(y)$ over the region $U \leq u$.
4. Find the pdf $f_U(u)$ of U by differentiating the cumulative distribution function $F_U(u)$ of U.

CDF method

Example 3.11

Let Y be a random variable with the pdf

$$f(y) = \begin{cases} 2y, & 0 \leq y \leq 1, \\ 0, & \text{elsewhere.} \end{cases}$$

If $U = 3Y - 1$, what is the pdf of U?

◁

CDF method

CDF method

Example 3.12

Let $Y \sim U(0,1)$ and let $U = Y^2$.
What is the pdf of U? $\triangleleft$

Increasing / Decreasing functions

Definition 3.16: Strictly increasing function

A function $h(y)$ is strictly increasing if for $y_1 < y_2$,

$$h(y_1) < h(y_2).$$

Definition 3.17: Strictly decreasing function

A function $h(y)$ is strictly decreasing if for $y_1 < y_2$,

$$h(y_1) > h(y_2).$$

Method of Transformations

Definition 3.18: Method of Transformations

Let Y have the pdf $f_Y(y)$.
If $h(y)$ is either a strictly increasing or decreasing function for all y where $f_Y(y) > 0$, then the random variable $U = h(Y)$ has pdf

$$f_U(u) = f_Y(h^{-1}(u))\left|\frac{dh^{-1}(u)}{du}\right|.$$

Using the Method of Transformations

Let $U = h(Y)$, where $h(y)$ is either strictly increasing or decreasing function of y for all y such that $f_Y(y) > 0$.

1. Find the inverse function $y = h^{-1}(u)$.

2. Evaluate $\frac{dh^{-1}(u)}{du}$.

3. Find $f_U(u)$ by

$$f_U(u) = f_Y(h^{-1}(u)) \left| \frac{dh^{-1}(u)}{du} \right| .$$

Using the transformation rule

Example 3.13

Let Y have pdf $f_Y(y) = \begin{cases} 2y & 0 \leq y \leq 1, \\ 0 & \text{elsewhere.} \end{cases}$

If $U = h(Y) = 4Y + 3$, what is the pdf of U? $\triangleleft$

Using the transformation rule

Example 3.14

Let $Y \sim N(0,1)$ and $U = Y^2$, what is the pdf of U?

$\triangleleft$

Using the transformation rule

Example 3.14 cont.

Let $Y \sim N(0,1)$ and $U = Y^2$, what is the pdf of U? $\triangleleft$

Order statistics

Many functions of random variables in practice depend on the relative magnitudes of the observed random variables. For example the fastest lap time observed in the Clipsal 500, or the highest score achieved on the last class exercise. Hence we often as not order the random variables from smallest to largest or vice versa, resulting in what we call *order statistics*.

For example if we collect a sequence of independent realisations of a continuous random variable $Y_1, Y_2, \ldots, Y_n$ with cumulative distribution function $F(y)$ and probability density function $f(y)$.
We denote the ordered random variables by $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ such that $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$, where the minimum is $Y_{(1)}$ and the maximum is $Y_{(n)}$.

Order statistics

What is the probability density function of the maximum?

Order statistics

What is the probability density function of the minimum?

Order statistics

What is the probability density function of the k^{th} order statistic?

Motivation example

Example 3.1

The time (in hours) a manager takes to interview a job applicant has an exponential distribution with $\lambda = 2$. The applicants are scheduled at quarter-hour intervals, beginning at 8:00 a.m., and the applicants arrive exactly on time. When the applicant with an 8:15 a.m. appointment arrives at the manager's office, what is the probability that he or she will have to wait before seeing the manager? ◁

Motivation example

MATHS 2103 / MATHS 7103 Probability and Statistics II Lecture notes

Section 4

Andrew Smith

School of Mathematical Sciences, University of Adelaide

Semester 1, 2017

Course outline

- **Section 01:** General information and introduction to Probability
- **Section 02:** Discrete random variables
- **Section 03:** Continuous random variables
- **Section 04:** Bivariate (multivariate) probability distributions
- **Section 05:** Discrete Time Markov Chains (DTMC)

Motivation example

Example 4.1

The number of eggs an insect lays follows a Poisson distribution with rate of 15 eggs per insect in a single brood.
The probability of each one of the eggs hatching is 0.001 independent of the other eggs.

What is the expected number of eggs per insect to hatch?

What is the variance of the number of hatching eggs per insect ?

◁

Discrete bivariate distributions

The joint probability of two or more events arises naturally in many experiments or in the consideration of nature and is therefore frequently of interest.

Discrete bivariate distributions

Definition 4.1: Bivariate probability mass function

Let Y_1 and Y_2 be discrete random variables, then the *joint or (bivariate) probability mass function* for Y_1 and Y_2 is given by

$$f(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) = \mathrm{P}((Y_1 = y_1) \cap (Y_2 = y_2)),$$
$$-\infty < y_1 < \infty, -\infty < y_2 < \infty.$$

Example 4.2

Toss a coin three times, let Y_1 be the number of heads on the first two coins and Y_2 be the total number of tails, what is the joint PMF of Y_1 and Y_2?

$\triangleleft$

[S4-5]

Discrete bivariate distributions

Discrete bivariate distributions

Properties of the bivariate PMF

If Y_1 and Y_2 are discrete random variables with joint probability mass function $f(y_1, y_2)$, then

- $f(y_1, y_2) \geq 0$ for all y_1, y_2.
- $\sum_{y_1} \sum_{y_2} f(y_1, y_2) = 1$, where the sum is over all the values (y_1, y_2) that are assigned nonzero probabilities.

Discrete bivariate distributions

Definition 4.2: Marginal probability mass function

Let Y_1 and Y_2 be jointly discrete random variables with probability mass function $f(y_1, y_2)$.
Then the *marginal probability mass functions* of Y_1 and Y_2, respectively, are given by

$$f_1(y_1) \quad = \quad \sum_{y_2} f(y_1, y_2)\,, \qquad \text{for each } y_1 \qquad \text{and}$$

$$f_2(y_2) \quad = \quad \sum_{y_1} f(y_1, y_2)\,, \qquad \text{for each } y_2.$$

Discrete bivariate distributions

Example 4.2 Marginal probability distribution $p_1(y_1)$

Discrete bivariate distributions

Example 4.2 Marginal probability distribution $p_2(y_2)$

Discrete bivariate distributions

Definition 4.3: Conditional probability distribution

If Y_1 and Y_2 are jointly discrete random variables with joint probability mass function $f(y_1, y_2)$ and marginal probability mass functions $f_1(y_1)$ and $f_2(y_2)$, respectively, then the *conditional discrete probability mass function* of Y_1 given Y_2 is

$$\begin{aligned} f(y_1|y_2) &= P(Y_1 = y_1|Y_2 = y_2) \\ &= \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)} = \frac{f(y_1, y_2)}{f_2(y_2)}, \end{aligned}$$

provided that $f_2(y_2) > 0$.

Discrete bivariate distributions

Example 4.2 Conditional distribution $f(y_1|y_2)$

◁

Discrete bivariate distributions

Example 4.2 Conditional distribution $f(y_2|y_1)$

◁

The binomial distribution (revisited)

Consider an experiment that has the following properties:

- n identical trials.
- There are only two outcomes $i = 1, 2$ to each trial.
- The probability p_i of outcome i is the same for each trial.
- The trials are independent.

Let Y_i be the number of trials $\in \{0, 1, \ldots, n\}$ that result in outcome $i = 1, 2$, so that $\sum_{i=1}^{2} Y_i = n$.
The bivariate probability mass function $f(y_1, y_2)$ is binomial

$$f(y_1, y_2) = \frac{n!}{y_1!y_2!} p_1^{y_1} p_2^{y_2} = \binom{n}{y_1} p_1^{y_1} (1 - p_1)^{n - y_1}.$$

where p_1 is probability of success, $p_2 = 1 - p_1$, $y_2 = n - y_1$.

The trinomial distribution

Consider now an experiment that has the following properties:

- n identical trials.
- There are only three outcomes $i = 1, 2, 3$ to each trial.
- The probability p_i of outcome i is the same for each trial.
- The trials are independent.

Let Y_i be the number of trials $\in \{0, 1, \ldots, n\}$ that result in outcome $i = 1, 2, 3$, such that

The trinomial distribution

Definition 4.4: Trinomial distribution

The joint (trivariate) probability mass function for random variables Y_1, Y_2, Y_3 given by

$$f(y_1, y_2, y_3) = \underbrace{\frac{n!}{y_1!y_2!y_3!}}_{\text{a triple partition}} p_1^{y_1} p_2^{y_2} p_3^{y_3},$$

where $p_i > 0$, such that

$$\sum_{i=1}^{3} p_i = 1, y_i \in \{0, 1, \ldots, n\} \text{ with } \sum_{i=1}^{3} y_i = n.$$

is said to have a trinomial distribution.

The trinomial distribution

Example 4.3

Toss two fair coins, let the possible outcomes of interest be two head, two tails, one head and one tail.

If this is repeated 10 times, what is the probability that

- the number of two head outcomes is 3
- the number of two tail outcomes is 3 and
- the rest of the trials consist of one head and one tail?

The trinomial distribution

The multinomial distribution

Consider an experiment that has the following properties:

- n identical trials.
- There are k outcomes $i = 1, 2, \ldots, k$ to each trial.
- The probability p_i of outcome i is the same for each trial.
- The trials are independent.

Let Y_i be the number of trials $\in \{0, 1, \ldots, n\}$ that result in outcome $i = 1, 2, \ldots, k$, so that

$$\sum_{i=1}^{k} Y_i = n, \quad p_i > 0 \text{ for } i = 1, 2, \ldots, k \text{ and } \sum_i p_i = 1.$$

The joint (multivariate) probability mass function $f(y_1, \ldots, y_k)$ has what is known as a multinomial distribution.

The multinomial distribution

Definition 4.5: Multinomial distribution

The joint (multivariate) probability mass function for random variables $Y_1, Y_2, \ldots, Y_k$ given by

$$f(y_1, y_2, \ldots, y_k) = \frac{n!}{y_1!y_2!\ldots y_k!} p_1^{y_1} p_2^{y_2} \ldots p_k^{y_k},$$

where $p_i > 0$, such that

$$\sum_{i=1}^{k} p_i = 1, y_i \in \{0, 1, \ldots, n\} \text{ with } \sum_{i=1}^{k} y_i = n.$$

is said to have a multinomial distribution.

The multinomial distribution

Properties of the multinomial distribution

If $Y_1, Y_2, \ldots, Y_k$ have a multinomial distribution with parameters n and $p_1, p_2, \ldots, p_k$, then

1. $\mathbb{E}[Y_i] = np_i$.
2. $\text{var}(Y_i) = np_i(1 - p_i)$, and
3. $\text{cov}(Y_s, Y_t) = -np_s p_t$, if $s \neq t$,

where the covariance $\text{cov}(Y_s, Y_t)$ is defined as follows.

Covariance of two random variables

Definition 4.6: Covariance of Y_1 and Y_2

If Y_1 and Y_2 are two (discrete or continuous) random variables with means μ_1 and μ_2 respectively, the covariance of Y_1 and Y_2 is

$$\text{cov}(Y_1, Y_2) = \mathbb{E}[(Y_1 - \mu_1)(Y_2 - \mu_2)]\,.$$

Continuous bivariate distributions

Properties of the bivariate pdf

If Y_1 and Y_2 are continuous random variables with joint probability density function $f(y_1, y_2)$, then

- $f(y_1, y_2) \geq 0$ for all y_1, y_2.
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 dy_2 = 1$.

To obtain the probabilities, we use integration in a similar way to the univariate case.

$$P\Big((a_1 \leq Y_1 \leq a_2), (b_1 \leq Y_2 \leq b_2)\Big) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f(y_1, y_2) dy_2 dy_1 .$$

Continuous bivariate distributions

Example 4.4

Let X_1 and X_2 have the joint pdf

$$f(x_1, x_2) = \begin{cases} 1 & 0 \leq x_1 \leq 1 \text{ and } 0 \leq x_2 \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

- Sketch the pdf.
- Find $F(0.2, 0.4)$.
- Find $P(0.1 \leq X_1 \leq 0.3, 0 \leq X_2 \leq 0.5)$.

◁

Continuous bivariate distributions

Example 4.4

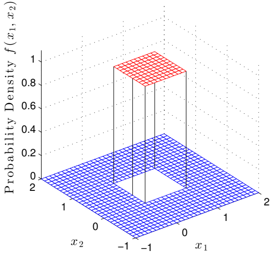


Figure 4.1: $f(x_1, x_2)$

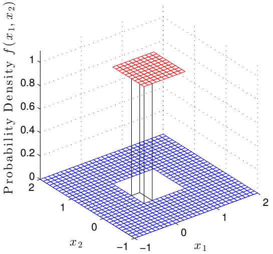


Figure 4.2: Volume

The pdf is sketched above in Figure 4.1 with the volume of the oblong cylinder in Figure 4.2 representing the probability $F(0.2, 0.4) = P(X_1 \leq 0.2, X_2 \leq 0.4)$ ◁

Continuous bivariate distributions

Continuous bivariate distributions

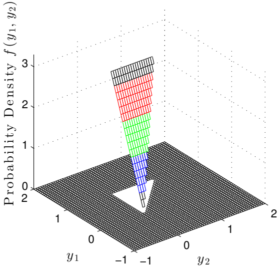
Continuous bivariate distributions

Example 4.5

$$\text{Let } f(y_1, y_2) = \begin{cases} 3y_1, & 0 \le y_2 \le y_1 \le 1, \\ & \\ 0, & \text{otherwise.} \end{cases}$$

Find $P(0 \le Y_1 \le 0.5, Y_2 > 0.25)$.

◁



Continuous bivariate distributions

Continuous bivariate distributions

Definition 4.7: Marginal probability distribution

Let Y_1 and Y_2 be jointly continuous random variables with probability density function $f(y_1, y_2)$.

Then the *marginal pdfs* of Y_1 and Y_2, respectively, are given by

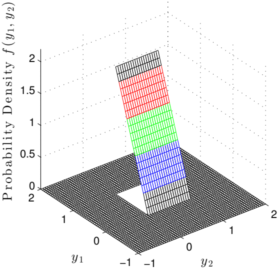
$$\begin{aligned} f_1(y_1) &= \int_{-\infty}^{\infty} f(y_1, y_2) dy_2 \\ \text{and} \quad f_2(y_2) &= \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 . \end{aligned}$$

Continuous bivariate distributions

Example 4.6

$$\text{Let } f(y_1, y_2) = \begin{cases} 2y_1, & 0 \le y_1 \le 1, 0 \le y_2 \le 1, \\ 0, & \text{otherwise.} \end{cases} .$$

Find the marginal pdf of Y_1 and Y_2. ◁



Continuous bivariate distributions

Definition 4.8: Conditional probability density function

Let Y_1 and Y_2 be jointly continuous random variables with joint pdf $f(y_1, y_2)$ and marginal densities $f_1(y_1)$ and $f_2(y_2)$, respectively.
The conditional density of Y_1 given $Y_2 = y_2$ is given by

$$f(y_1|y_2) \quad = \quad \frac{f(y_1, y_2)}{f_2(y_2)} \quad \text{for any } y_2 \text{ such that } f_2(y_2) > 0$$

and the conditional density of Y_2 given $Y_1 = y_1$ is given by

$$f(y_2|y_1) \quad = \quad \frac{f(y_1, y_2)}{f_1(y_1)} \quad \text{for any } y_1 \text{ such that } f_1(y_1) > 0.$$

Continuous bivariate distributions

Example 4.7

$$\text{Let } f(y_1, y_2) = \begin{cases} 1/2 & 0 \leq y_1 \leq y_2 \leq 2, \\ 0 & \text{otherwise.} \end{cases}.$$

Calculate $P(Y_1 \leq 1/2 | Y_2 = 1.5)$. $\triangleleft$

Continuous bivariate distributions

Example 4.7 (cont).

Independence of random variables

Definition 4.9: Independent random variables

Let Y_1 have cumulative distribution function $F_1(y_1)$,
let Y_2 have cumulative distribution function $F_2(y_2)$, and
let Y_1 and Y_2 have joint distribution function $F(y_1, y_2)$.
Then the random variables Y_1 and Y_2 are said to be *independent* if and only if

$$F(y_1, y_2) = F_1(y_1)F_2(y_2)$$

for every pair of real numbers (y_1, y_2).
If Y_1 and Y_2 are not independent, they are said to be *dependent*.

Independence of random variables

Theorem 4.1 (Discrete independent random variables)

If Y_1 and Y_2 are discrete random variables with joint probability mass function $p(y_1, y_2)$, where

Y_1 has marginal probability mass function $p_1(y_1)$ and
Y_2 has marginal probability mass function $p_2(y_2)$,

then Y_1 and Y_2 are independent if and only if

$$p(y_1, y_2) = p_1(y_1)p_2(y_2)$$

for all real numbers (y_1, y_2).

Independence of random variables

Theorem 4.2 (Continuous independent random variables)

If Y_1 and Y_2 are continuous random variables with joint probability density function $f(y_1, y_2)$, where

Y_1 has marginal probability density function $f_1(y_1)$ and Y_2 has marginal probability density function $f_2(y_2)$,

then Y_1 and Y_2 are independent if and only if

$$f(y_1, y_2) = f_1(y_1) f_2(y_2)$$

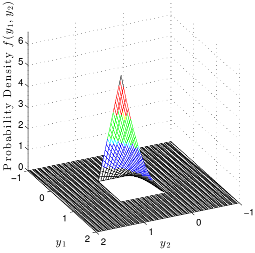
for all real numbers (y_1, y_2).

Independence of random variables

Example 4.8

$$\text{Let } f(y_1, y_2) = \begin{cases} 6y_1y_2^2 & 0 \le y_1 \le 1, 0 \le y_2 \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Show that Y_1 and Y_2 are independent. $\triangleleft$



Independence of random variables

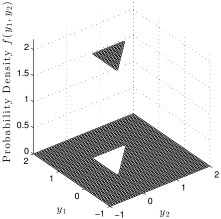
Example 4.8 cont.

Independence of random variables

Example 4.9

Let $f(y_1, y_2) = \begin{cases} 2 & 0 \le y_2 \le y_1 \le 1 \\ 0 & \text{elsewhere} \end{cases}$

Show that Y_1 and Y_2 are dependent.



Independence of random variables

Example 4.9 cont.

Joint (bivariate) random variables

Definition 4.10: $\mathbb{E}$[function of joint random variables]

Let $g(Y_1, Y_2)$ be a real-valued function of the random variables Y_1 and Y_2, then

$$\mathbb{E}[g(Y_1, Y_2)] \;=\; \sum_{y_1} \sum_{y_2} g(y_1, y_2) p(y_1, y_2),$$

if Y_1, Y_2 are discrete with joint pmf $p(y_1, y_2)$, or

$$\mathbb{E}[g(Y_1, Y_2)] \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y_1, y_2) f(y_1, y_2) dy_1 dy_2$$

if Y_1, Y_2 are continuous with joint pdf $f(y_1, y_2)$.

Joint (bivariate) random variables

Theorem 4.3

Let Y_1 and Y_2 be independent random variables and let $g(Y_1)$ be a function of only Y_1 and let $h(Y_2)$ be a function of only Y_2, then

$$\mathbb{E}[g(Y_1)h(Y_2)] \;=\; \mathbb{E}[g(Y_1)]\,\mathbb{E}[h(Y_2)]\,,$$

provided that the expectations exist.

Recall Definition 4.5: Covariance of Y_1 and Y_2

If Y_1 and Y_2 are two random variables with means μ_1 and μ_2 respectively, the covariance of Y_1 and Y_2 is

$$\operatorname{cov}(Y_1, Y_2) = \mathbb{E}[(Y_1 - \mu_1)(Y_2 - \mu_2)]\,.$$

Joint (bivariate) random variables

Theorem 4.4 (Properties of the covariance)

1. *cov*$(Y_1, Y_2) = \mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1]\,\mathbb{E}[Y_2]$*.*

2. *If* Y_1 *and* Y_2 *are independent then cov*$(Y_1, Y_2) = 0$*,*

Joint (bivariate) random variables

Definition 4.10: Correlation

If Y_1 and Y_2 are two (discrete or continuous) random variables with means μ_1 and μ_2 respectively, the correlation coefficient of Y_1 and Y_2 is

$$\rho \; = \; \text{corr}(Y_1, Y_2) \; = \; \frac{\text{cov}(Y_1, Y_2)}{\sqrt{\text{var}(Y_1)}\sqrt{\text{var}(Y_2)}}.$$

Joint (bivariate) random variables

Example 4.10 Recall Example 4.5, where we saw

$$f(y_1, y_2) = \begin{cases} 3y_1, & 0 \le y_2 \le y_1 \le 1, \\ \\ 0, & \text{otherwise.} \end{cases}$$

Show that Y_1 and Y_2 are dependent by finding the covariance of Y_1 and Y_2. ◁

Joint (bivariate) random variables

Example 4.10 cont.

◁

Joint (bivariate) random variables

Example 4.10 cont.

◁

Joint (bivariate) random variables

Example 4.10 cont.

Joint (bivariate) random variables

A linear combination of two random variables

Consider two random variables Y_1 and Y_2 and two real constants a and b, then

Mean: the mean of $aY_1 + bY_2$ is given by

$$\mathbb{E}[aY_1 + bY_2] \quad = \quad a\mathbb{E}[Y_1] + b\mathbb{E}[Y_2]\,,$$

and

Variance: the variance of $aY_1 + bY_2$ is given by

$$\mathrm{var}(aY_1 + bY_2) \quad = \quad a^2\mathrm{var}(Y_1) + b^2\mathrm{var}(Y_2) + 2ab\,\mathrm{cov}(Y_1, Y_2)\,.$$

Joint random variables

Definition 4.11: The MGF of bivariate random variables

Consider the random variables Y_1 and Y_2. The joint moment generating function of Y_1 and Y_2 is defined to be

$$m_{Y_1,Y_2}(t_1, t_2) = \mathbb{E}\left[e^{t_1 Y_1 + t_2 Y_2}\right].$$

Theorem 4.5 (Sum of independent random variables)

If $Y_1, \ldots, Y_n$ are n independent random variables with moment generating functions $m_{Y_1}(t), \ldots, m_{Y_n}(t)$, respectively, then the moment generating function of the sum $U = Y_1 + \cdots + Y_n = \sum_{i=1}^{n} Y_i$ is given by

$$m_U(t) = m_{Y_1}(t) m_{Y_2}(t) \ldots m_{Y_n}(t).$$

Joint random variables

Definition 4.12: Method of moment generating functions

Let U be a function of the random variables $Y_1, Y_2, \ldots, Y_n$.

1. Find the moment generating function, $m_U(t)$ for U.
2. Compare $m_U(t)$ with other well-known moment generating functions.

   If $m_U(t) = m_V(t)$ for all values of t,
   then by Theorem 3.1 (Uniqueness of MGFs),
   U has the same probability distribution as V.

Joint random variables

Example 4.11

If Y_1 and Y_2 are independent random variables, such that Y_1 is a Poisson random variable with parameter λ_1 and Y_2 is a Poisson random variable with parameter λ_2, then show that $X = Y_1 + Y_2$ is a Poisson random variable with parameter $(\lambda_1 + \lambda_2)$. ◁

Bivariate normal distribution

Definition 4.13: The bivariate normal distribution

Let the random variables Y_1 and Y_2 be such that

$$Y_1 \sim N(\mu_1, \sigma_1^2),\; Y_2 \sim N(\mu_2, \sigma_2^2) \text{ and } \rho = \text{corr}(Y_1, Y_2)\,,$$

then the joint probability density function of Y_1 and Y_2 is

$$f(y_1, y_2) \;=\; \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-Q/2}, \quad \text{with}$$

$$Q = \frac{1}{1-\rho^2}\left[\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2}\right]$$

where $-\infty < x, y < \infty, \;\; \sigma_1, \sigma_2 > 0$ and $-1 < \rho < 1$.

Bivariate normal distribution

Definition 4.14: Standard bivariate normal distribution

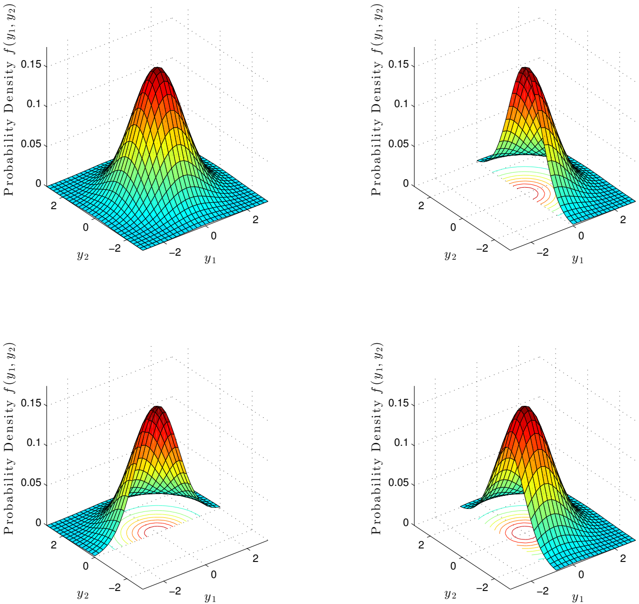
The random variables Y_1 and Y_2 are said to have a standard bivariate normal distribution if their joint probability density function is

$$f(y_1, y_2) = \frac{1}{2\pi\sqrt{1-\rho^2}}\; e^{\left(-\frac{y_1^2 + y_2^2 - 2\rho y_1 y_2}{2(1-\rho^2)}\right)},$$

such that $-\infty < x, y < \infty$ and $\rho = \text{corr}(Y_1, Y_2)$.

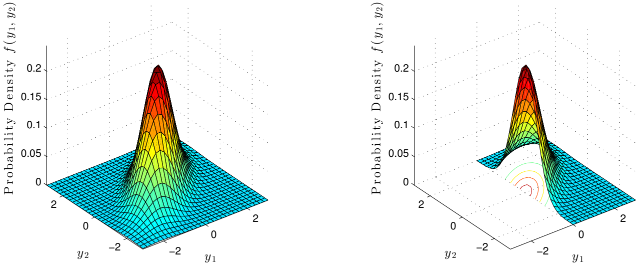
The std bivariate normal distribution

$\rho = 0.0$

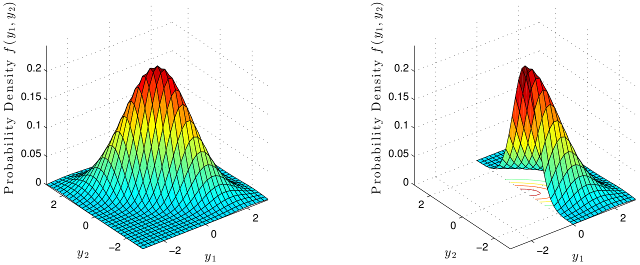


The std bivariate normal distribution

$\rho = 0.7$



$\rho = -0.7$



The std bivariate normal distribution

Example 4.12 The marginal distributions

If the random variables Y_1 and Y_2 have the standard bivariate normal distribution, then show that the marginal distribution of Y_1 is given by

$$f_1(y_1) = \frac{1}{\sqrt{2\pi}} e^{-y_1^2/2},$$

and the marginal distribution of Y_2 is given by

$$f_2(y_2) = \frac{1}{\sqrt{2\pi}} e^{-y_2^2/2}.$$

◁

The std bivariate normal distribution

Example 4.12 (cont.)

The std bivariate normal distribution

Example 4.12 (cont.)

The std bivariate normal distribution

Example 4.12 (cont.)

The std bivariate normal distribution

Example 4.13 Conditional distributions

If the random variables Y_1 and Y_2 have the standard bivariate normal distribution, then show that the conditional distribution of Y_1 given $Y_2 = y_2$ is

$$f_{Y_1|Y_2}(y_1|Y_2 = y_2) = \frac{1}{\sqrt{2\pi(1-\rho^2)}}\, e^{\left(-\frac{(y_1 - \rho y_2)^2}{2(1-\rho^2)}\right)}.$$

◁

The std bivariate normal distribution

Example 4.13 (cont).

◁

General bivariate normal distribution

Conditional distribution

In the case of the general bivariate normal distribution, where recall Y_1 and Y_2 are such that

$$Y_1 \sim N(\mu_1, \sigma_1^2),\ \ Y_2 \sim N(\mu_2, \sigma_2^2)\ \ \text{with}\ \ \rho = \text{corr}(Y_1, Y_2)\,,$$

it can be shown that the conditional distribution of Y_1 given $Y_2 = y_2$ for $-\infty < y_2 < \infty$ is normally distributed as

$$N\left(\left[\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2)\right], \left[\sigma_1^2(1 - \rho^2)\right]\right).$$

General bivariate normal distribution

The joint moment generating function

In the case of the general bivariate normal distribution, where recall Y_1 and Y_2 are such that

$$Y_1 \sim N(\mu_1, \sigma_1^2), \;\; Y_2 \sim N(\mu_2, \sigma_2^2) \;\text{ with }\; \rho = \operatorname{corr}(Y_1, Y_2)\,,$$

it can be shown that the joint moment generating function $m_{Y_1,Y_2}(t_1, t_2)$ of Y_1 and Y_2 is

$$\mathbb{E}\left[e^{t_1 Y_1 + t_2 Y_2}\right] \;\;=\;\; e^{\left(t_1\mu_1 + t_2\mu_2 + \frac{1}{2}\left(t_1^2\sigma_1^2 + 2t_1 t_2 \rho \sigma_1^2 \sigma_2^2 + t_2^2\sigma_2^2\right)\right)}.$$

General bivariate normal distribution

The marginal moment generating function

To obtain the marginal moment generating function of Y_i, set $t_j, j \neq i$ to zero, so for example the marginal moment generating function of Y_1 is

$$m_{Y_1}(t_1) = \mathbb{E}\left[e^{t_1 Y_1 + 0 Y_2}\right]$$

$$= e^{\left(t_1 \mu_1 + 0\mu_2 + \frac{1}{2}\left(t_1^2 \sigma_1^2 + 2 t_1 0 \rho \sigma_1^2 \sigma_2^2 + 0^2 \sigma_2^2\right)\right)}$$

$$= e^{\left(t_1 \mu_1 + \frac{1}{2}\left(t_1^2 \sigma_1^2\right)\right)}.$$

Linear combinations of rvs

Linear combinations of random variables

Let $Y_1, Y_2, \ldots, Y_n$ be a collection of random variables and let $X_1, X_2, \ldots, X_m$ be another collection of random variables such that

$$\mathbb{E}[Y_i] = \mu_i \text{ and } \mathbb{E}\big[X_j\big] = \xi_j$$

and define

$$U_1 = \sum_{i=1}^{n} a_i Y_i \text{ and } U_2 = \sum_{j=1}^{m} b_j X_j,$$

for real constants $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_m$.

Then the following hold...

Linear combinations of rvs

Mean, variance and covariance

1. $\mathbb{E}[U_1] = \sum_{i=1}^{n} a_i \mu_i.$

2. $\text{var}(U_1) = \sum_{i=1}^{n} a_i^2 \text{var}(Y_i) + 2 \sum_{j=2}^{n} \sum_{i=1}^{\min(j-1,n-1)} a_i a_j \text{cov}\big(Y_i, Y_j\big).$

3. $\text{cov}(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{cov}\big(Y_i, X_j\big).$

Linear combinations of rvs

Example 4.14

Let Y_1, Y_2, Y_3 be random variables such that

Variable	Mean	Variance
Y_1	1	1
Y_2	2	3
Y_3	-1	5

Covariances

$$\text{cov}(Y_1, Y_2) = -0.4$$

$$\text{cov}(Y_1, Y_3) = 0.5$$

$$\text{cov}(Y_2, Y_3) = 2.0$$

If $U = Y_1 - 2Y_2 + Y_3$ and $W = 3Y_1 + Y_2$, calculate the mean, variance and covariance of U and W. $\triangleleft$

Linear combinations of rvs

Example 4.14 (cont).

◁

Linear combinations of rvs

Example 4.14 (cont).

◁

Linear combinations of rvs

Example 4.14 (cont).

Linear combinations of iid rvs

Theorem 4.6 (Weak law of large numbers)

If $X_1, X_2, \ldots, X_i, \ldots$ is a sequence of independent identically distributed random variables such that

$$\mathbb{E}[X_i] \;=\; \mu,\; \textit{var}(X_i) \;=\; \sigma^2,\; \textit{with}\; \overline{X}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \textit{then}$$

$$P\Big(|\overline{X}_n - \mu| > \epsilon\Big) \to 0 \quad \textit{as } n \to \infty \textit{ for all } \epsilon > 0.$$

Linear combinations of iid rvs

Proof of the weak law of large numbers

Linear combinations of iid rvs

Proof of the weak law of large numbers

Linear combinations of iid rvs

Theorem 4.7 (The central limit theorem)

Let $Y_1, Y_2, \ldots, Y_i, \ldots$ be a sequence of independent identically distributed random variables such that

$$\mathbb{E}[Y_i] \;=\; \mu \;\; \textit{and} \;\; \textit{var}(Y_i) \;=\; \sigma^2.$$

$$\textit{Let} \;\; U_n \;=\; \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}, \quad \textit{where} \;\; \overline{Y} \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

then the distribution of U_n converges to the standard normal distribution function as $n \to \infty$. That is, for all u

$$\lim_{n\to\infty} P(U_n \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Linear combinations of iid rvs

Normal approximation to the Binomial distribution

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

can be approximated by

$$Y \sim N(np, np(1-p)).$$

As a ‘rule of thumb’, this approximation works well if both np and $n(1-p)$ are greater than or equal to 10.

Linear combinations of iid rvs

Continuity correction

As the approximation to the discrete Binomial distribution is by a continuous Normal distribution, then a continuity correction is required to improve the approximation (see figure on the next slide).

Let $X \sim \text{Bin}(n, p)$ and $Y \sim N(np, np(1-p))$, then

$$\text{P}(X \leq a) \approx \text{P}\left(Y \leq a + \frac{1}{2}\right),$$

$$\text{P}(X \leq a-1) = \text{P}(X < a) \approx \text{P}\left(Y \leq a - \frac{1}{2}\right) = \text{P}\left(Y \leq a - 1 + \frac{1}{2}\right).$$

Linear combinations of iid rvs

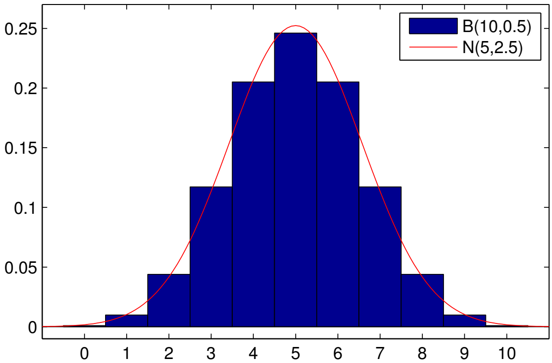


Figure 4.3: Continuity correction for a $N(5, 2.5)$ approximation to a $Bin(10, 0.5)$

Linear combinations of iid rvs

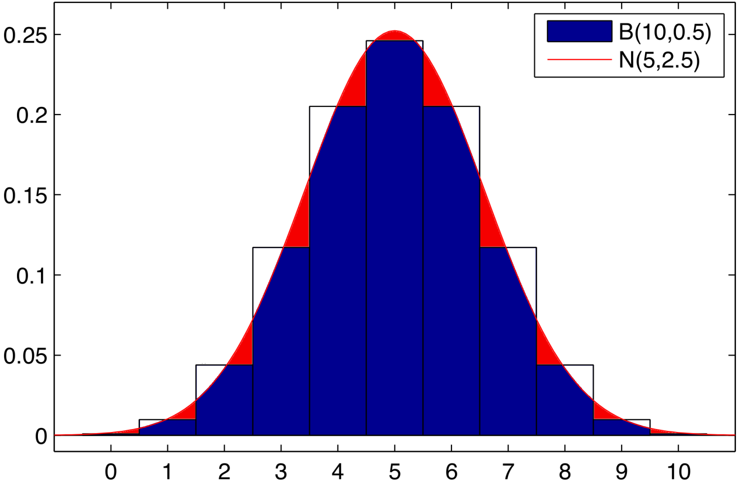


Figure 4.4: Continuity correction for a $N(5, 2.5)$ approximation to a $Bin(10, 0.5)$

Linear combinations of iid rvs

Definition 4.15: Conditional Expectation

Let Y_1 and Y_2 be two random variables, then the *conditional expectation* of $g(Y_1)$, given that $Y_2 = y_2$, is

$$\mathbb{E}[g(Y_1)|Y_2 = y_2] = \sum_{y_1} g(y_1)f(y_1|y_2)\,,$$

if Y_1 and Y_2 are jointly discrete and

$$\mathbb{E}[g(Y_1)|Y_2 = y_2] = \int_{-\infty}^{\infty} g(y_1)f(y_1|y_2)\,dy_1\,,$$

if Y_1 and Y_2 are jointly continuous.

Conditional Expectation

Example 4.15 Recall Example 4.7

$$\text{Let } f(y_1, y_2) = \begin{cases} 1/2, & 0 \leq y_1 \leq y_2 \leq 2, \\ \\ 0, & \text{otherwise.} \end{cases}$$

What is the expected value of Y_1 given that $Y_2 = 1.5$? $\triangleleft$

Conditional Expectation

Theorem 4.8 (Conditional Expectation Theorem)

Let Y_1 and Y_2 denote random variables, then

$$\mathbb{E}[Y_1] = \mathbb{E}[\mathbb{E}[Y_1 | Y_2]]$$

Proof (Discrete):

Conditional Expectation

Proof of Theorem 4.8 (Continuous):

Conditional Variance

Theorem 4.9 (Conditional Variance Theorem)

Let Y_1 and Y_2 denote random variables, then

$$var(Y_1) = \mathbb{E}[var(Y_1|Y_2)] + var(\mathbb{E}[Y_1|Y_2]) \,.$$

Proof:

Motivation example

Example 4.1 (recall the motivation example)

The number of eggs an insect lays follows a Poisson distribution with rate of 15 eggs per insect in a single brood.
The probability of each one of the eggs hatching is 0.001 independent of the other eggs.

What is the expected number of eggs per insect to hatch?

What is the variance of the number of hatching eggs per insect ?

◁

Motivation example

Example 4.1 (recall the motivation example)

MATHS 2103 / MATHS 7103 Probability and Statistics II Lecture notes

Section 5

Andrew Smith

School of Mathematical Sciences, University of Adelaide

Semester 1, 2017

Course outline

- **Section 01:** General information and introduction to Probability
- **Section 02:** Discrete random variables
- **Section 03:** Continuous random variables
- **Section 04:** Bivariate (multivariate) probability distributions
- **Section 05:** Discrete Time Markov Chains (DTMC)

Motivation example

Example 5.1 Expected time to extinction

Consider a population process that is observed and modelled in discrete time.

At time point $t \in \mathbb{Z}_0^+$, the population :

increases by one with probability p or

decreases by one with probability $(1 - p)$,

where $\boldsymbol{p} < \frac{1}{2}$.

What is the expected time until the population becomes extinct? ◁

Discrete Time Markov Chains

We often want to model processes of events where the probability associated with the next event only depends on what has just occurred. Markov chain models have this property that is often described as being “memoryless”. It essentially means that the next state that is visited by the Markov chain depends only on the present state, and not on any previous states.

Definition 5.1: Random processes (stochastic processes)

Discrete Time Markov Chains

Let's consider the case where the index n corresponds to discrete units of time, so that $T = \{0, 1, 2, \ldots\}$.

This type of random process is therefore said to occur in **discrete time**.

Furthermore, we restrict our attention to discrete random variables, which always have a countable state space.

Let $X_n, n \in \mathbb{N}$, be a sequence of random variables with a countable state space $S = \{x_0, x_1, x_2, \ldots, x_i, \ldots\}$.

Discrete Time Markov Chains

Definition 5.2: Discrete time Markov chain (DTMC)

A discrete time random process $X_n, n \in \mathbb{N}$, is a DTMC if it satisfies

$$\begin{aligned} \mathrm{P}(X_n = s \mid X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \\ = \mathrm{P}(X_n = s \mid X_{n-1} = x_{n-1}) \end{aligned} \tag{5.1}$$

for all $n \geq 1$ and all $s, x_0, x_1, \ldots, x_{n-1} \in S$.

Discrete Time Markov Chains

Since the state space *S* can be put into a one-to-one correspondence with some subset of the natural numbers $\mathbb{N} = \{0, 1, 2, \dots\}$, without loss of generality, we can assume that a subset of $\mathbb{N}$ is the state space.

This simplifies our notation considerably as we essentially can give up the physical interpretation of the states for a numerical one, and we can simply write more general expressions for the transition probabilities like

Discrete Time Markov Chains

Note that the “starting state” X_0 can be specified a priori or it can be chosen randomly from some distribution across S.

Definition 5.3: Time homogeneous Markov chain

A Markov chain X_n is called **time-homogeneous** if we have for all n and $i, j \in S$ that

$$\mathrm{P}(\, X_{n+1} = j \mid X_n = i) \;=\; \mathrm{P}(\, X_1 = j \mid X_0 = i)\,.$$

Discrete Time Markov Chains

Example 5.2 Repeated coin toss

A coin is tossed repeatedly that shows a head with probability p at each toss independently of all other tosses. If we count the number of heads seen in this experiment, then the state space will consist of states

$\triangleleft$

Discrete Time Markov Chains

Example 5.2 Repeated coin toss (cont)

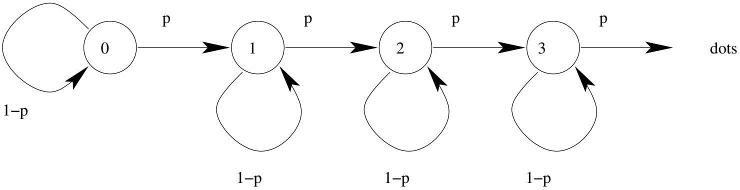


Figure 5.1: Counting the heads state space diagram.

The state space here is infinite as we have not imposed a restriction on the number of coin tosses, but we always start in state $X_0 = 0$ with probability 1 at the beginning of the experiment. $\triangleleft$

Discrete Time Markov Chains

Example 5.3 A vending machine

A vending machine can be in one of two states:

- Working
- Not working

Assume that if the machine is working on a particular day, then the probability that it will not be working on the next day is given by δ such that $0 < \delta < 1$.

When the machine is not working on a particular day, it will be in working condition on the next day with probability γ such that $0 < \gamma < 1$.

◁

Discrete Time Markov Chains

Example 5.3 A vending machine (cont)

If we say the machine is
in state 0 if not working and in state 1 if it is working, so that

and we can draw a state diagram here as in Figure 5.2.

Figure 5.2: Vending machine state space diagram.

Discrete Time Markov Chains

Example 5.3 A vending machine (cont)

In what state we start here could be set a-priori or maybe selected by some distribution such as $(p, 1-p)$.

If we assume that the machine starts in state 0, we can consider the evolution of the machine's condition for say values of $\delta = 0.9$ and $\gamma = 0.2$.

$$
\begin{aligned}
X_0 &= 0 \\
X_1 &= \begin{cases} 0 & \text{with probability} \quad 1-\gamma = 0.8 \\ 1 & \text{with probability} \quad \gamma = 0.2 \end{cases} \\
X_2 &= \begin{cases} 0 & \text{with probability} \quad 0.2(0.9) + 0.8(0.8) = 0.82 \\ 1 & \text{with probability} \quad 0.8(0.2) + 0.2(0.1) = 0.18 \end{cases}
\end{aligned}
$$

◁

Discrete Time Markov Chains

Example 5.3 A vending machine (cont)

Using simulation, since we know how to generate random numbers now and then generate the random variables X_n we can also investigate this evolution from each possible starting state.

Start	10 days	50 days	100 days	1000 days
$X_0 = 1$	0.5	0.86	0.82	0.812
$X_0 = 0$	0.9	0.82	0.84	0.815

Table 1: Proportion of time the machine is not working after · days

Every simulation yields a different sample path, but over a large horizon it appears as if the proportion of time not working is independent of X_0 and is in excess of 0.81 ◁

Discrete Time Markov Chains

Definition 5.4: Transition matrix

The **transition matrix** $\mathbb{P}$ of a discrete time time-homogeneous Markov chain is the $|S| \times |S|$ matrix of transition probabilities

$$p_{i,j} = \mathrm{P}(X_{n+1} = j \mid X_n = i) \qquad \rightarrow \qquad \mathbb{P} = [p_{i,j}].$$

Discrete Time Markov Chains

Example 5.4 A desperate gambler

Suppose a gambler has $1 and desperately wants to get $5. He is offered the chance to play repeated rounds of a game where he has the probability p of winning each round. He adopts the following “Bold Play” strategy.

- At each round he will bet all he has if winning will give him his goal of $5 or less.
- Otherwise he will bet the amount that would make his goal of $5 if he wins.

His initial bet is thus always $1 and if he wins, he has $2 which he bets with probability 1 in the next round. ◁

Discrete Time Markov Chains

Example 5.4 A desperate gambler (cont)

Discrete Time Markov Chains

Example 5.5 The gambler

A gambler begins with $1,000 and bets either

$100 with probability $\frac{3}{4}$, or

$200, with probability $\frac{1}{4}$,

on each hand of cards.
If the gambler bets

$100, she wins with probability 3/8, or if she bets
$200, she wins with probability 2/3.

(Don't ask me what game she is playing!).

If she has only $100 left, she always bets $100. ◁

Discrete Time Markov Chains

Example 5.5 The gambler (cont.)

We assume her opponent has an unlimited supply of money and we will describe the process whose state $S = \mathbb{N}$ is the amount of money in hundreds of dollars that the gambler has after each game as a DTMC.

Discrete Time Markov Chains

Example 5.5 The gambler (cont.)

Discrete Time Markov Chains

Example 5.5 The gambler (cont.)

If the gambler intends to play at most 60 hands in a night, the probability of interest may be $p^{(60)}_{10,j}$

in particular $p^{(60)}_{10,0}$ is the probability she loses the $1000

or $\displaystyle\sum_{j=0}^{\infty} j\, p^{(60)}_{10,j}$ is her average winnings.

◁

Definition 5.5: *m*-step transition matrix

The *m***-step transition matrix** $\mathbb{P}^{(m)} = [p_{i,j}^{(m)}]$ of a time-homogeneous Markov chain is the $|\mathcal{S}| \times |\mathcal{S}|$ matrix of transition probabilities

$$p_{i,j}^{(m)} = \mathrm{P}(\, X_{n+m} = j \mid X_n = i)\,.$$

Discrete Time Markov Chains

In general, there are multiple possible “paths” that result in this outcome.

The probability $p_{i,j}^{(m)}$ is essentially the sum of probabilities of all such paths.

Discrete Time Markov Chains

Theorem 5.1 (**The *m*-step transition matrix)**

$$\mathbb{P}^{(m)} = \mathbb{P}^{m} \, .$$

This theorem is useful because it tells us that the entry (i, j) of the matrix $\mathbb{P}^m$ is equal to the *m*-step transition probability $p_{i,j}^{(m)}$.

A proof of this Theorem follows, which should be well understood. We use induction on *m* because the thing we wish to prove has a general pattern in terms of powers that sits naturally with such a method of proof.

Discrete Time Markov Chains

Proof of Theorem 5.1:

We use induction on m: Since it is true that $\mathbb{P}^1 = \mathbb{P}^{(1)}$, we assume that $\mathbb{P}^{\,m-1} = \mathbb{P}^{(m-1)}$, and we condition $p_{i,j}^{(m)}$ on the state after $m-1$

steps.

Discrete Time Markov Chains

Example 5.6 (3×3) DTMC

$$\text{If } \mathbb{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \text{ What is } p_{1,2}^{(3)} \text{ ?}$$

Discrete Time Markov Chains

The “starting state” is given by the value of the random variable X_0, which can be set a priori or can be chosen randomly from some distribution across the state space S.

Let the vector $\boldsymbol{X}(0) = (P(X_0 = 0), P(X_0 = 1), \ldots)$

be the probability distribution (probability vector containing the probability mass function) describing the probability that the process is in each of the states at time 0.

Corollary 5.1

Assume the Markov chain starts with probability distribution $\boldsymbol{X}(0)$, then the distribution of states at time n is given by

$$\boldsymbol{X}(n) = \boldsymbol{X}(0)\mathbb{P}^n.$$

Discrete Time Markov Chains

Proof:

Discrete Time Markov Chains

Computer Demonstrations:

Discrete Time Markov Chains

Equilibrium Behaviour of Markov Chains

Assume that

$$p_{ij}^{(m)} \to \pi_j \text{ as } m \to \infty \tag{5.2}$$

That is, the probability of being in state *j* after many steps converges to some constant value and hence is independent of the initial state *i*. Note that this property does not always occur with all Markov Chains, as we will see later in this section.

In the proof of Theorem 5.1, we saw that

$$p_{ij}^{(m)} = \sum_k p_{ik}^{(m-1)} p_{kj},$$

in which we are now going to take the limit as $m \to \infty$.

Discrete Time Markov Chains

That is, we consider

$$\lim_{m\to\infty} p_{ij}^{(m)} = \lim_{m\to\infty} \sum_k p_{ik}^{(m-1)} p_{kj}.$$

The assumption made in equation 5.2 implies that the above limits give us

$$\pi_j = \sum_k \pi_k \, p_{kj}, \quad \text{for all } j. \tag{5.3}$$

Discrete Time Markov Chains

Example 5.7 A Queue

A queueing system is observed every minute and the number of customers in the system is determined. It is observed that the queue length increases by one, between observations, with probability p, decreases by one with probability q (provided it is not empty) and stays the same with probability $1 - p - q$. When the queue is empty, it increases by one with probability p and stays empty with probability $1 - p$. ◁

Discrete Time Markov Chains

In Example 5.7 we readily establish the following transition probabilities for a DTMC which models this queue on the inifinite state space $S = \{0, 1, 2, 3, \dots\}$.

Discrete Time Markov Chains

Consider now the equilibrium equations 5.3, which for this DTMC become

The solution to this system of equations if it exists, is given by a row vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots)$, which we recall is known as the equilibrium probability distribution for the DTMC that models this queueing system.

The equilibrium probability distribution given by $\boldsymbol{\pi}$ has three physical interpretations as given on the next slides.

Discrete Time Markov Chains

1. **Limiting:**

   By definition $\lim_{n\to\infty} p_{i,j}^{(n)} = \pi_j$ and so π_j is the limiting probability of the process being in state j.

2. **Stationary:**

   We showed that $\boldsymbol{\pi}\mathbb{P} = \boldsymbol{\pi}$, and so $\boldsymbol{\pi}$ is the equilibrium (or stationary) probability distribution of the process.

Discrete Time Markov Chains

3. **Ergodic:**

   It can be shown (and we won't) that there is also the ergodic interpretation.

Discrete Time Markov Chains

Methods for solving equilibrium equations:

There is no all-purpose automatic method, but instead the choice of method comes with experience. However,

1. **If the process has a finite number of states, *N*.**

   The equilibrium equations are a system of N linear equations in N unknowns. Note, there is always exactly one redundant equation (because the row sums are all one and so the last column can be deduced from this fact and all the other columns). However, there is also the normalising equation and so we now have N linearly independent equations in N unknowns and so the unique solution can be found by any of the usual matrix methods.

Discrete Time Markov Chains

Example 5.8 Finite queue

Discrete Time Markov Chains

2. **If the process has an infinite (countable) number of states:**

   and the transition probabilities p_{jk} do not depend on the actual value of j for $j \geq J$, but just $k - j$.

   That is, we have a homogeneous Markov chain for states above J and there are two natural methods that can be used in this case.

   Which method you choose will probably depend on what information you wish to extract from the equilibrium probability distribution after you have found it.

Discrete Time Markov Chains

Difference (or Recurrence) Equation Methods

If you just wish to determine equilibrium probability distribution, then this is probably the most efficient method.

In Example 5.7, the equilibrium equations were:

$$\begin{aligned}
\pi_n &= p\pi_{n-1} + (1-p-q)\pi_n + q\pi_{n+1}, \quad n \geq 1, \\
\pi_0 &= (1-p)\pi_0 + q\pi_1 \quad \text{such that} \\
\sum_i \pi_i &= 1, \quad \text{which we rewrite as}
\end{aligned}$$

$$(p+q)\pi_n = p\pi_{n-1} + q\pi_{n+1}, \quad n \geq 1 \tag{5.4}$$

$$p\pi_0 = q\pi_1. \tag{5.5}$$

Discrete Time Markov Chains

We shall solve equations 5.4 and 5.5 using a method that works on difference equations with constant coefficients.

It is similar to solving 2nd order linear DE's with constant coefficients - there we try a solution of the form $y = e^{\lambda t}$ to get a characteristic equation.

Discrete Time Markov Chains

Discrete Time Markov Chains

Probability Generating Function Method:

If you wish to extract summary statistics from the distribution, then it is most efficient to use a method based on probability generating functions, see Definition 2.13 of Section 2.

In Example 5.7, we saw that the equilibrium equations were

$$\begin{aligned} (p+q)\pi_n &= p\pi_{n-1} + q\pi_{n+1}, \quad n \geq 1 && (5.6) \\ p\pi_0 &= q\pi_1. && (5.7) \end{aligned}$$

Recall that for a probability distribution like $\{\pi_n, n \geq 0\}$, the pgf is given by $P(z) = \displaystyle\sum_{n=0}^{\infty} \pi_n z^n$, from which we have ready access to each π_n for $n = 0, 1, 2, \ldots$ and all of the moments.

Discrete Time Markov Chains

Therefore, we are trying to get an expression for $\sum_{n=0}^{\infty} \pi_n z^n$.

Discrete Time Markov Chains

Discrete Time Markov Chains

3. **If the process has an infinitely countable number of states and the transition probabilities p_{jk} do depend on the actual value of j and not just $k - j$.**

   This is of course the most general class of problems and many of these types of problems are impossible to solve analytically. However, there is a large sub-class of problems for which posses a very useful property, known as **partial balance**.

To define this sub-class of problems is very difficult, but it suffices to say that any problem you meet in this course will fit into either this sub-class or one of the other two classes we have previously discussed.

Discrete Time Markov Chains

Example 5.9 An optical fibre link

An optical fibre link can be assumed to have essentially infinite carrying capacity for telephone calls. The state of the link is the number of calls in progress, observed at each time point that a new call arrives or a call ends. The transition probabilities are

$$
\begin{aligned}
p_{i,i+1} &= \frac{p}{p+iq}, \quad i \geq 0 \\
p_{i,i-1} &= \frac{iq}{p+iq}, \quad i \geq 1 \\
p_{i,k} &= 0, \quad \text{otherwise.}
\end{aligned}
$$

◁

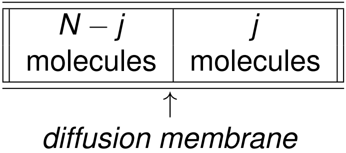
Discrete Time Markov Chains

Discrete Time Markov Chains

Example 5.10 The Ehrenfest Model

This model considers the diffusion of N gas molecules across a membrane in a diffusion chamber.

Let the state of the system j be the number of molecules in the right-hand chamber and assume that



$N-j$ molecules

j molecules

↑

diffusion membrane

$$p_{j,j+1} = \frac{N-j}{N}, \quad p_{j,j-1} = \frac{j}{N} \quad \text{and} \quad p_{j,k} = 0,$$

The equilibrium equations 5.3 for $1 \leq j \leq N-1$ are

$$\pi_j \;=\; \frac{N-(j-1)}{N}\pi_{j-1} + \frac{j+1}{N}\pi_{j+1}. \tag{5.12}$$

◁

Discrete Time Markov Chains

Example 5.10 The Ehrenfest Model (cont).

The two boundary equations for this system are

$$\pi_0 = \frac{1}{N}\pi_1 \quad \text{and} \quad \pi_N = \frac{1}{N}\pi_{N-1}.$$

We can solve this system similarly to example 5.9. ◁

Discrete Time Markov Chains

Discrete Time Markov Chains

Absorbing Markov Chains

There are many occasions where we wish to use Markov Chains to model situations in which the process stops making transitions if it gets into a particular state.

Prime examples of this are gambling games like that of example 5.5 in which the process stops if one of the players (Player *A*) loses all of their money.

Other examples occur in population models where the process stops if a species dies out.

States in which the process stops are called **absorbing states**.

Discrete Time Markov Chains

In such a Markov Chain, the states can always be re-ordered so that the transition matrix looks like

$$\begin{array}{cccc} & \text{non-absorbing} & \text{absorbing} & \\ \mathbb{P}_A = & \left(\begin{array}{c} R \\ 0 \end{array}\right. & \left.\begin{array}{c} S \\ I \end{array}\right) & \begin{array}{l}\text{non-absorbing}\\ \text{absorbing}\end{array} \end{array}$$

where 0 is the zero matrix and I is the identity matrix.

Example 5.11 Gambler's Ruin

Consider a gambling game played by two people, player A, who starts with $\$k$, and player B, who starts with $\$j$. The total fortune of the two players is $\$N = \$k + \$j$. ◁

Discrete Time Markov Chains

Example 5.11 Gambler’s Ruin (cont).

Assume that the players play games in succession until one loses all of their money, with player A having a probability *p* of winning each game. Denote the state of the process by the amount of money player A has. The transition matrix for the Markov chain describing this game is

$$\mathbb{P} = \begin{pmatrix} 1 & 0 & 0 & & & 0 \\ 1-p & 0 & p & 0 & & 0 \\ 0 & 1-p & 0 & p & & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & & 0 & 1-p & 0 & p \\ 0 & & & & 0 & 1 \end{pmatrix}.$$

Discrete Time Markov Chains

Discrete Time Markov Chains

We re-order the states to obtain useful expressions for $\mathbb{P}^n$ and $\lim_{n\to\infty} \mathbb{P}^n$ which is covered by the following two Theorems.

Theorem 5.2

For a transition matrix with the form $\mathbb{P}_A$

$$\mathbb{P}_A^n = \begin{pmatrix} R^n & \sum_{i=0}^{n-1} R^i S \\ 0 & I \end{pmatrix}.$$

Theorem 5.3

For a transition matrix with the form $\mathbb{P}_A$

$$\lim_{n\to\infty} \mathbb{P}_A^n = \begin{pmatrix} 0 & (I-R)^{-1} S \\ 0 & I \end{pmatrix}.$$

Discrete Time Markov Chains

Proof of theorem 5.2:

Discrete Time Markov Chains

Sketch of Proof of theorem 5.3:

Discrete Time Markov Chains

For i, a non-absorbing state, and j, an absorbing state,

the $(i, j)^{\text{th}}$ entry of $\lim_{n\to\infty} \mathbb{P}_A^n$ contains the probability that the process is eventually absorbed into j conditional on starting in i.

Thus, for example, in the gamblers ruin problem, the $(k, N)^{\text{th}}$ entry contains the probability that player A will eventually win all the money and the $(k, 0)^{\text{th}}$ entry contains the probability that player A eventually loses all the money.

Note that the $(k, j)^{\text{th}}$ entry is zero for all other j and so no other possibilities can occur.

Discrete Time Markov Chains

We can now evaluate the probability of being in each state after *n* games, and hence absorption in less than or equal to *n* games. We can also find the probability of eventual absorption into the different absorbing states. However, if the latter is all we want, then we can evaluate them in a more efficient way.

Theorem 5.4 (Absorption probability)

Let $X_j^{(N)}$ be the probability that the process is absorbed in state N given that it starts in state j. Let state 0 *be the only other absorbing state. Then $X_j^{(N)}$ satisfies the equation*

$$X_j^{(N)} = \sum_k p_{jk} X_k^{(N)}, \quad 1 \leq j \leq N-1$$

$$\textit{with } X_N^{(N)} = 1, X_0^{(N)} = 0.$$

Discrete Time Markov Chains

Proof:

Discrete Time Markov Chains

Example 5.12 Gamblers Ruin (cont).

Discrete Time Markov Chains

This gives us the probability that player A eventually wins, given that she starts with \$$j$ and alternatively we could have found the probability that she loses everything given that she starts with \$$j$. However, if we assume the game eventually ends, then we have both at once since $X_j^{(0)} = 1 - X_j^{(N)}$.

What if instead, we want to know how long it takes to win or lose everything? In other words “what is the expected time until absorption?”.

Such questions are particularly of interest in population models, where the mean time until extinction is of interest.

How do we calculate such a measure?

Discrete Time Markov Chains

Let's consider a general DTMC with state space $0, 1, 2, \dots$ and let 0 be the only absorbing state. Based on these suppositions, the following theorem gives us the approach we need to find the expected absorption time in such a DTMC.

Theorem 5.5 (Expected absorption time)

Let M_j be the mean time until absorption of the DTMC, conditional on starting in state j.
Then M_j satisfies the equations

$$M_j = 1 + \sum_k p_{j,k} M_k, \quad \textit{for } j \geq 1, \qquad \textit{with } M_0 = 0.$$

Discrete Time Markov Chains

Proof:

Let H_0 be random variable that is the
time to absorption in state 0. Then $M_j = \mathbb{E}\left[H_0 \middle| X_0 = j\right]$ is given by

Motivation example

Example 5.1 (recall the motivation example)

Consider a population process that is observed and modelled in discrete time.

At time point $t \in \mathbb{Z}_0^+$, the population increases by one with probability p or decreases by one with probability $(1 - p)$, where $\boldsymbol{p < \frac{1}{2}}$.

What is $\mathbb{E}$[time until the population becomes extinct]? $\triangleleft$

Discrete Time Markov Chains

Discrete Time Markov Chains

We can derive this result in another way, by considering the expected time that it takes for the process to move down from k to $k-1$. To do this, we pretend that $k-1$ is an absorbing state.

Discrete Time Markov Chains