COURSE NOTES

STATS 325

Stochastic Processes

The
University
of Auckland

Department of Statistics
University of Auckland

Contents

Chapter 1: Stochastic Processes

What are Stochastic Processes, and how do they fit in?

STATS 210 Foundations of Statistics and Probability		STATS 310 Statistics

STATS 210
**Foundations of
Statistics and Probability**

Tools for understanding randomness
(random variables, distributions)

STATS 310
Statistics

Randomness in Pattern

STATS 325
Probability

Randomness in Process
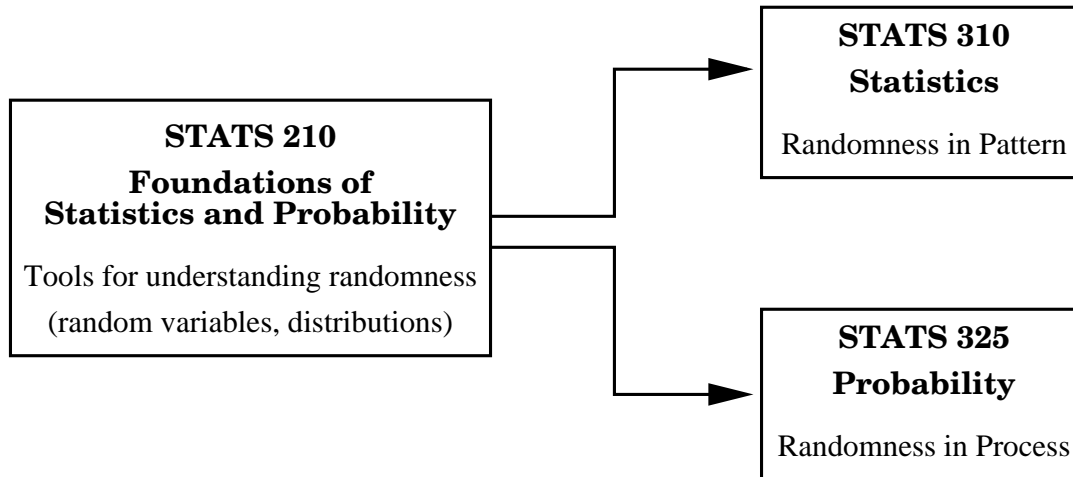
Stats 210: laid the foundations of both Statistics and Probability: the tools for understanding randomness.

Stats 310: develops the theory for understanding ***randomness in pattern:*** tools for estimating parameters (maximum likelihood), testing hypotheses, modelling patterns in data (regression models).

Stats 325: develops the theory for understanding ***randomness in process.*** A process is a sequence of events where each step follows from the last after a random choice.

What sort of problems will we cover in Stats 325?

Here are some examples of the sorts of problems that we study in this course.

Gambler's Ruin

You start with \$30 and toss a fair coin repeatedly. Every time you throw a Head, you win \$5. Every time you throw a Tail, you lose \$5. You will stop when you reach \$100 or when you lose everything. What is the probability that you lose everything?   ***Answer:*** 70%.

With permission, copyright Martin Ouellet

Winning at tennis

What is your probability of winning a game of tennis, starting from the even score Deuce (40-40), if your probability of winning each point is 0.3 and your opponent's is 0.7?

Answer: 15%.
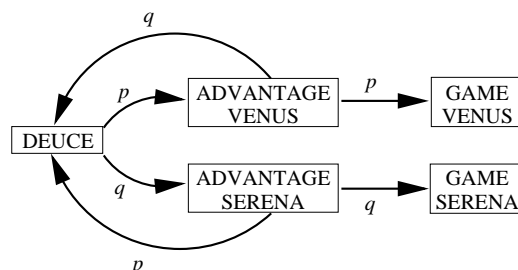
Winning a lottery

A million people have bought tickets for the weekly lottery draw. Each person has a probability of one-in-a-million of selecting the winning numbers. If more than one person selects the winning numbers, the winner will be chosen at random from all those with matching numbers.

You watch the lottery draw on TV and your numbers match the winning numbers!!! Only a one-in-a-million chance, and there were only a million players, so surely you will win the prize?

Not quite... What is the probability you will win? **Answer:** only 63%.

Drunkard's walk

A very drunk person staggers to left and right as he walks along. With each step he takes, he staggers one pace to the left with probability 0.5, and one pace to the right with probability 0.5. What is the expected number of paces he must take before he ends up one pace to the left of his starting point?

Arrived!
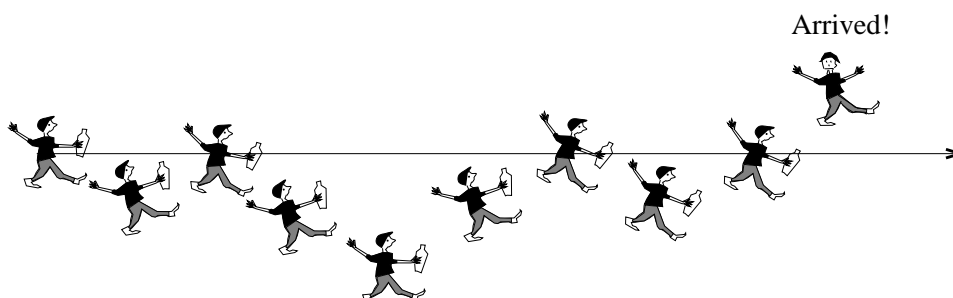
Answer: the expectation is infinite!

Pyramid selling schemes

Have you received a chain letter like this one? Just send $10 to the person whose name comes at the top of the list, and add your own name to the bottom of the list. Send the letter to as many people as you can. Within a few months, the letter promises, you will have received $77,000 in $10 notes! Will you?

I WAS AMAZED WHEN I SAW HOW MUCH MONEY CAME FLOODING THROUGH MY LETTER BOX...I TURNED $218 INTO $78190 WITHIN THE FIRST 80 DAYS OF OPERATING THIS BUSINESS PLAN

DO NOT BIN THIS IMMEDIATELY
THINK ABOUT IT FOR A FEW DAYS
FILE IN PENDING

My name is David Rhodes and in September 1997 I lost my job. At the time I was living at the edge of my means and in debt. Consequently, this started a chain reaction that ended with the repossession of my home and car. If that wasn't enough several debt collectors were constantly houndi imagine life looked bleak.

In January 1998 I received a letter telling me how to make ov time. I ignored it because I was sceptical. However by March in debt. I finally realised that I had absolutely nothing to lose apart from that, I couldn't stop myself from thinking what if

In the summer of 1999 my family and I went on a cruise and new Mercedes with cash and we are currently building our $ home and I don't owe a single cent.

To date I have made over $1,100,000. Even now as I write thi it hard to come to terms with the fact that like most people, I

THIS IS HOW THE SYSTEM WORKS
WITHIN 60 DAYS

You have sent off your $10 note then mailed 200 letters (minimum) your details are printed at No5 on each of them. Your tasks are now complete. Sit back and relax- you deserve it.

If only 3% of 200 people respond to your letter, 6 people will mail 200 letter each =1200 letters with your name at No4.

If only 3% of 1200 people respond to your letter, 36 people will mail 200 letters each = 7,200 letters with your name at No3

If only 3% of 7,200 people respond to your letter, 216 people will mail 200 letters each =$43,200 letters with your name at No2.

If only 3% of 43,200 people respond to your letter, 1296 people will mail 200 letters each. = 259,200 letters with your name at No1.

If only 3% of 259,200 people respond to their letters 7,776 people will send you $10 each. because your name is at No1 position therefore you will receive
$77,760.00 in $10 notes

Answer: it depends upon the response rate. However, with a fairly realistic assumption about response rate, we can calculate an expected return of $76 with a 64% chance of getting nothing!

Note: Pyramid selling schemes like this are prohibited under the Fair Trading Act, and it is illegal to participate in them.

Spread of SARS

The figure to the right shows the spread of the disease SARS (Severe Acute Respiratory Syndrome) through Singapore in 2003. With this pattern of infections, what is the probability that the disease eventually dies out of its own accord?

Answer: 0.997.
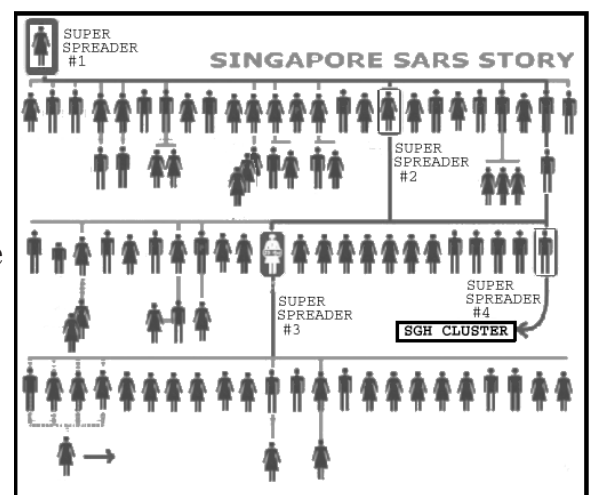
Markov's Marvellous Mystery Tours

Mr Markov's Marvellous Mystery Tours promises an All-Stochastic Tourist Experience for the town of Rotorua. Mr Markov has eight tourist attractions, to which he will take his clients completely at random with the probabilities shown below. He promises at least three exciting attractions per tour, ending at either the Lady Knox Geyser or the Tarawera Volcano. (Unfortunately he makes no mention of how the hapless tourist might get home from these places.)

What is the expected number of activities for a tour starting from the museum?
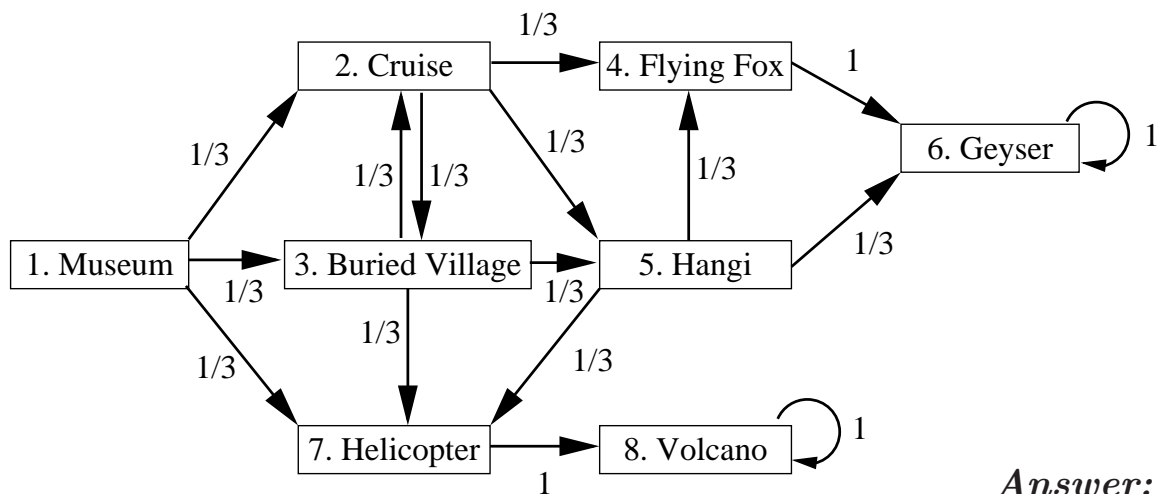
Answer: 4.2.

Structure of the course

- **Probability.** Revision of probability and random variables from Stats 210, with special focus on conditional probability. New material on the application of conditional probability to finding probabilities for stochastic processes.

- **Expectation.** Revision of expectation and variance. Introduction to conditional expectation, and its application in finding expected reaching times in stochastic processes.

- **Generating functions.** Introduction to probability generating functions, and their applications to stochastic processes.

- **Branching processes.** Branching processes are a simple model for reproduction. Examples are the pyramid selling scheme and the spread of SARS above.

- **_Markov chains._** Almost all the examples we look at throughout the course can be formulated as Markov chains. By developing a single unifying theory, we can easily tackle complex problems with many states and transitions like Markov's Marvellous Mystery Tours above.

The rest of this chapter covers:

- quick revision of sample spaces and random variables;
- formal definition of stochastic processes.

1.1 Revision: Sample spaces and random variables

Definition: A **random experiment** is a physical situation whose outcome cannot be predicted until it is observed.

Definition: A **sample space**, Ω, is a set of possible outcomes of a random experiment.

Example:

  Random experiment: Toss a coin once.
  Sample space: $\Omega = \{$_head, tail_$\}$

Definition: A **random variable**, X, is defined as a function from the sample space to the real numbers: $X : \Omega \to \mathbb{R}$.

  That is, _a random variable assigns a real number to every possible outcome of a random experiment._

Example:

  Random experiment: Toss a coin once.
  Sample space: $\Omega = \{$head, tail$\}$.
  An example of a random variable: $X : \Omega \to \mathbb{R}$ maps "head" $\to 1$, "tail" $\to 0$.

  Essential point: | A random variable is a way of producing random real numbers.

1.2 Stochastic Processes

Definition: A **stochastic process** is a *family of random variables,*
  $\{X(t) : t \in T\}$, *where t usually denotes time. That is, at every time t in the set T, a random number $X(t)$ is observed.*

Definition: $\{X(t) : t \in T\}$ is a **discrete-time process** if *the set T is finite or countable.*

In practice, this generally means $T = \{0, 1, 2, 3, \ldots\}$

Thus a discrete-time process is $\{X(0), X(1), X(2), X(3), \ldots\}$: *a new random number recorded at every time 0, 1, 2, 3, …*

Definition: $\{X(t) : t \in T\}$ is a **continuous-time process** if *T is <u>not</u> finite or countable.*

In practice, this generally means $T = [0, \infty)$, *or $T = [0, K]$ for some K.*

Thus a continuous-time process $\{X(t) : t \in T\}$ *has a random number $X(t)$ recorded at every instant in time.*

(Note that $X(t)$ need not *change* at every instant in time, but it is *allowed* to change at any time; i.e. not just at $t = 0, 1, 2, \ldots$, like a discrete-time process.)

Definition: The **state space**, S, is *the set of real values that $X(t)$ can take.*

Every $X(t)$ takes a value in $\mathbb{R}$, but S will often be a smaller set: $S \subseteq \mathbb{R}$. For example, if $X(t)$ is the outcome of a coin tossed at time t, then the state space is $S = \{0, 1\}$.

Definition: The state space S is **discrete** if it is *finite or countable.*
  Otherwise it is *continuous.*

> The state space S is the set of **states** that the stochastic process can be in.

For Reference: Discrete Random Variables

1. Binomial distribution

Notation: $X \sim \text{Binomial}(n, p)$.

Description: number of successes in n independent trials, each with probability p of success.

Probability function:

$$f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n.$$

Mean: $\mathbb{E}(X) = np$.

Variance: $\text{Var}(X) = np(1-p) = npq$, where $q = 1 - p$.

Sum: If $X \sim \text{Binomial}(n, p)$, $Y \sim \text{Binomial}(m, p)$, and X and Y are **independent**, then

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

2. Poisson distribution

Notation: $X \sim \text{Poisson}(\lambda)$.

Description: arises out of the Poisson process as the number of events in a fixed time or space, when events occur at a constant average rate. Also used in many other situations.

Probability function: $f_X(x) = \mathbb{P}(X = x) = \dfrac{\lambda^x}{x!} e^{-\lambda} \quad \text{for} \quad x = 0, 1, 2, \ldots$

Mean: $\mathbb{E}(X) = \lambda$.

Variance: $\text{Var}(X) = \lambda$.

Sum: If $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$, and X and Y are **independent**, then

$$X + Y \sim \text{Poisson}(\lambda + \mu).$$

3. Geometric distribution

Notation: $X \sim \text{Geometric}(p)$.

Description: number of failures before the **first** success in a sequence of independent trials, each with $\mathbb{P}(\text{success}) = p$.

Probability function: $f_X(x) = \mathbb{P}(X = x) = (1-p)^x p$ for $x = 0, 1, 2, \ldots$

Mean: $\mathbb{E}(X) = \dfrac{1-p}{p} = \dfrac{q}{p}$, where $q = 1 - p$.

Variance: $\text{Var}(X) = \dfrac{1-p}{p^2} = \dfrac{q}{p^2}$, where $q = 1 - p$.

Sum: if $X_1, \ldots, X_k$ are **independent**, and each $X_i \sim \text{Geometric}(p)$, then

$$X_1 + \ldots + X_k \sim \text{Negative Binomial}(k, p).$$

4. Negative Binomial distribution

Notation: $X \sim \text{NegBin}(k, p)$.

Description: number of failures before the **kth** success in a sequence of independent trials, each with $\mathbb{P}(\text{success}) = p$.

Probability function:

$$f_X(x) = \mathbb{P}(X = x) = \binom{k + x - 1}{x} p^k (1-p)^x \quad \text{for} \quad x = 0, 1, 2, \ldots$$

Mean: $\mathbb{E}(X) = \dfrac{k(1-p)}{p} = \dfrac{kq}{p}$, where $q = 1 - p$.

Variance: $\text{Var}(X) = \dfrac{k(1-p)}{p^2} = \dfrac{kq}{p^2}$, where $q = 1 - p$.

Sum: If $X \sim \text{NegBin}(k, p)$, $Y \sim \text{NegBin}(m, p)$, and X and Y are **independent**, then

$$X + Y \sim \text{NegBin}(k + m, p).$$

5. Hypergeometric distribution

Notation: $X \sim \text{Hypergeometric}(N, M, n)$.

Description: Sampling without replacement from a finite population. Given N objects, of which M are 'special'. Draw n objects without replacement. X is the number of the n objects that are 'special'.

Probability function:

$$f_X(x) = \mathbb{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad \text{for} \quad \left\{ \begin{array}{l} x = \max(0, \; n + M - N) \\ \text{to} \; x = \min(n, \; M). \end{array} \right.$$

Mean: $\mathbb{E}(X) = np$, where $p = \dfrac{M}{N}$.

Variance: $\text{Var}(X) = np(1-p)\left(\dfrac{N-n}{N-1}\right)$, where $p = \dfrac{M}{N}$.

6. Multinomial distribution

Notation: $\boldsymbol{X} = (X_1, \ldots, X_k) \sim \text{Multinomial}(n; \; p_1, p_2, \ldots, p_k)$.

Description: there are n independent trials, each with k possible outcomes. Let $p_i = \mathbb{P}(\text{outcome } i)$ for $i = 1, \ldots k$. Then $\boldsymbol{X} = (X_1, \ldots, X_k)$, where X_i is the number of trials with outcome i, for $i = 1, \ldots, k$.

Probability function:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

for $x_i \in \{0, \ldots, n\} \; \forall_i$ with $\displaystyle\sum_{i=1}^{k} x_i = n$, and where $p_i \geq 0 \; \forall_i$, $\displaystyle\sum_{i=1}^{k} p_i = 1$.

Marginal distributions: $X_i \sim \text{Binomial}(n, p_i)$ for $i = 1, \ldots, k$.

Mean: $\mathbb{E}(X_i) = np_i$ for $i = 1, \ldots, k$.

Variance: $\text{Var}(X_i) = np_i(1 - p_i)$, for $i = 1, \ldots, k$.

Covariance: $\text{cov}(X_i, X_j) = -np_i p_j$, for all $i \neq j$.

Continuous Random Variables

1. Uniform distribution

Notation: $X \sim \text{Uniform}(a, b)$.

Probability density function (pdf): $f_X(x) = \dfrac{1}{b-a}$ for $a < x < b$.

Cumulative distribution function:

$$F_X(x) = \mathbb{P}(X \le x) = \frac{x-a}{b-a} \quad \text{for } a < x < b.$$

$$F_X(x) = 0 \text{ for } x \le a, \text{ and } F_X(x) = 1 \text{ for } x \ge b.$$

Mean: $\mathbb{E}(X) = \dfrac{a+b}{2}$.

Variance: $\text{Var}(X) = \dfrac{(b-a)^2}{12}$.

2. Exponential distribution

Notation: $X \sim \text{Exponential}(\lambda)$.

Probability density function (pdf): $f_X(x) = \lambda e^{-\lambda x}$ for $0 < x < \infty$.

Cumulative distribution function:

$$F_X(x) = \mathbb{P}(X \le x) = 1 - e^{-\lambda x} \quad \text{for } 0 < x < \infty.$$

$$F_X(x) = 0 \text{ for } x \le 0.$$

Mean: $\mathbb{E}(X) = \dfrac{1}{\lambda}$.

Variance: $\text{Var}(X) = \dfrac{1}{\lambda^2}$.

Sum: if $X_1, \ldots, X_k$ are **independent**, and each $X_i \sim \text{Exponential}(\lambda)$, then

$$X_1 + \ldots + X_k \sim \text{Gamma}(k, \lambda).$$

3. Gamma distribution

Notation: $X \sim \text{Gamma}(k, \lambda)$.

Probability density function (pdf):

$$f_X(x) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} \quad \text{for } 0 < x < \infty,$$

where $\Gamma(k) = \int_0^\infty y^{k-1} e^{-y} \, dy$ (the Gamma function).

Cumulative distribution function: no closed form.

Mean: $\mathbb{E}(X) = \dfrac{k}{\lambda}$.

Variance: $\text{Var}(X) = \dfrac{k}{\lambda^2}$.

Sum: if $X_1, \ldots, X_n$ are **independent**, and $X_i \sim \text{Gamma}(k_i, \lambda)$, then

$$X_1 + \ldots + X_n \sim \text{Gamma}(k_1 + \ldots + k_n, \lambda).$$

4. Normal distribution

Notation: $X \sim \text{Normal}(\mu, \sigma^2)$.

Probability density function (pdf):

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\{-(x-\mu)^2/2\sigma^2\}} \quad \text{for } -\infty < x < \infty.$$

Cumulative distribution function: no closed form.

Mean: $\mathbb{E}(X) = \mu$.

Variance: $\text{Var}(X) = \sigma^2$.

Sum: if $X_1, \ldots, X_n$ are **independent**, and $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$, then

$$X_1 + \ldots + X_n \sim \text{Normal}(\mu_1 + \ldots + \mu_n, \ \sigma_1^2 + \ldots + \sigma_n^2).$$

Probability Density Functions

Uniform(a, b)

$f_X(x)$

$\dfrac{1}{b-a}$

$a \qquad b \qquad x$

Exponential(λ)

$\lambda = 2$

$\lambda = 1$

Gamma(k, λ)

$k = 2,\ \lambda = 1$

$k = 2,\ \lambda = 0.3$

Normal(μ, σ^2)

$\sigma = 2$

$\sigma = 4$

μ
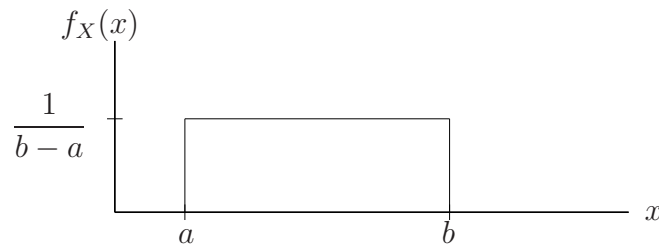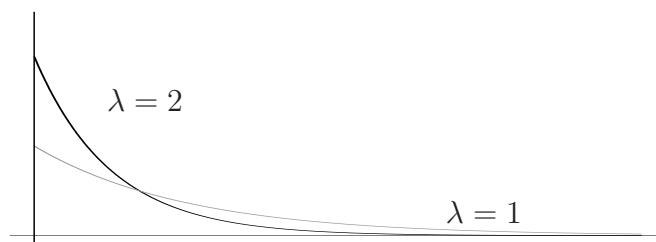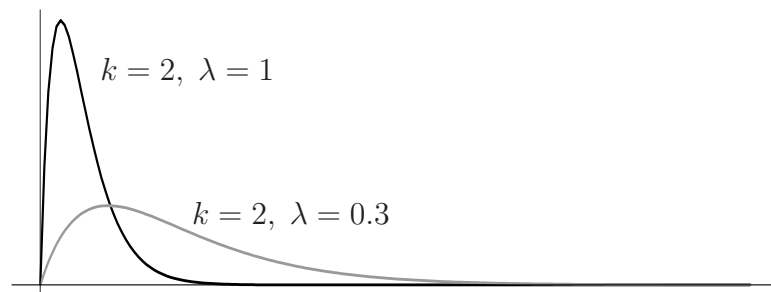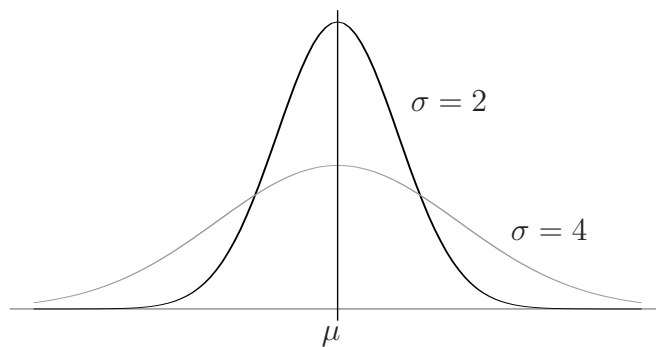
Chapter 2: Probability

The aim of this chapter is to revise the basic rules of probability. By the end of this chapter, you should be comfortable with:

- conditional probability, and what you can and can't do with conditional expressions;

- the Partition Theorem and Bayes' Theorem;

- First-Step Analysis for finding the probability that a process reaches some state, by conditioning on the outcome of the first step;

- calculating probabilities for continuous and discrete random variables.

2.1 Sample spaces and events

Definition: A **sample space**, Ω, is a *set of possible outcomes of a random experiment.*

Definition: An **event**, A, is a *subset of the sample space.*

This means that event A is simply *a collection of outcomes.*

Example:

Random experiment: Pick a person in this class at random.
Sample space: $\Omega = \{$*all people in class*$\}$
Event A: $A = \{$*all males in class*$\}$.

Definition: Event A **occurs** if *the outcome of the random experiment is a member of the set A.*

In the example above, event A occurs if *the person we pick is male.*

2.2 Probability Reference List

The following properties hold for all events A, B.

- $\mathbb{P}(\emptyset) = 0$.

- $0 \leq \mathbb{P}(A) \leq 1$.

- **Complement:** $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$.

- **Probability of a union:** $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
  For three events A, B, C:

  $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$.

  If A and B are **mutually exclusive**, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

- **Conditional probability:** $\mathbb{P}(A \,|\, B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

- **Multiplication rule:** $\mathbb{P}(A \cap B) = \mathbb{P}(A \,|\, B)\mathbb{P}(B) = \mathbb{P}(B \,|\, A)\mathbb{P}(A)$.

- **The Partition Theorem:** if $B_1, B_2, \ldots, B_m$ form a partition of Ω, then

$$\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{m} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i) \quad \text{for any event } A.$$

  As a special case, B and $\overline{B}$ partition Ω, so:

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap \overline{B}) \\ &= \mathbb{P}(A \,|\, B)\mathbb{P}(B) + \mathbb{P}(A \,|\, \overline{B})\mathbb{P}(\overline{B}) \quad \text{for any } A, B. \end{aligned}$$

- **Bayes' Theorem:** $\mathbb{P}(B \,|\, A) = \dfrac{\mathbb{P}(A \,|\, B)\mathbb{P}(B)}{\mathbb{P}(A)}$.

  More generally, if $B_1, B_2, \ldots, B_m$ form a partition of Ω, then

$$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{m} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)} \quad \text{for any } j.$$

- **Chains of events:** for any events $A_1, A_2, \ldots, A_n$,

  $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \,|\, A_1)\mathbb{P}(A_3 \,|\, A_2 \cap A_1) \ldots \mathbb{P}(A_n \,|\, A_{n-1} \cap \ldots \cap A_1)$.

2.3 Conditional Probability

Suppose we are working with sample space
$\Omega = \{$people in class$\}$. I want to find the
proportion of people in the class who ski. What do I do?

*Count up the number of people in the class who ski, and divide by the total
number of people in the class.*

$$\mathbb{P}(person\ skis) = \frac{number\ of\ skiers\ in\ class}{total\ number\ of\ people\ in\ class}.$$

Now suppose I want to find the proportion of *females* in the class who ski.
What do I do?

*Count up the number of females in the class who ski, and divide by the total
number of females in the class.*

$$\mathbb{P}(female\ skis) = \frac{number\ of\ female\ skiers\ in\ class}{total\ number\ of\ females\ in\ class}.$$

By changing from asking about everyone to asking about females only, we have:

- *restricted attention to the set of females only,*

or: *reduced the sample space* *from the set of everyone to the set of females,*

or: *conditioned* *on the event* $\{$*females*$\}$.

We could write the above as:

$$\mathbb{P}(skis\,|\,female) = \frac{number\ of\ female\ skiers\ in\ class}{total\ number\ of\ females\ in\ class}.$$

> Conditioning is like changing the sample space: we are now working in
> a new sample space of females in class.

In the above example, we could replace 'skiing' with *any* attribute B. We have:

$$\mathbb{P}(\text{skis}) = \frac{\#\text{ skiers in class}}{\#\text{ class}}\,; \qquad \mathbb{P}(\text{skis}\,|\,\text{female}) = \frac{\#\text{ female skiers in class}}{\#\text{ females in class}}\,;$$

so:

$$\mathbb{P}(B) = \frac{\#\text{ } B\text{'s in class}}{\text{total }\#\text{ people in class}}\,,$$

and:

$$\mathbb{P}(B\,|\,\text{female}) = \frac{\#\text{ female } B\text{'s in class}}{\text{total }\#\text{ females in class}}$$

$$= \frac{\#\text{ in class who are } B \text{ and female}}{\#\text{ in class who are female}}\,.$$

Likewise, we could replace 'female' with any attribute A:

$$\mathbb{P}(B\,|\,A) = \frac{\text{number in class who are } B \text{ and } A}{\text{number in class who are } A}\,.$$

This is how we get the definition of conditional probability:

$$\mathbb{P}(B\,|\,A) = \frac{\mathbb{P}(B \text{ and } A)}{\mathbb{P}(A).} = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}\,.$$

By conditioning on event A, we have *changed the sample space to the set of A's only.*

Definition: Let A and B be events on the same sample space: so $A \subseteq \Omega$ *and* $B \subseteq \Omega$. The **conditional probability of event B, given event A**, is

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

<u>**Multiplication Rule:**</u> (Immediate from above). For any events A and B,

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

Conditioning as 'changing the sample space'

The idea that *"conditioning" = "changing the sample space"* can be very helpful in understanding how to manipulate conditional probabilities.

Any 'unconditional' probability can be written as a conditional probability:

$$\mathbb{P}(B) = \mathbb{P}(B \,|\, \Omega).$$

Writing $\mathbb{P}(B) = \mathbb{P}(B \,|\, \Omega)$ just means that we are looking for the probability of event B, out of all possible outcomes in the set Ω.

In fact, the symbol $\mathbb{P}$ **belongs** to the set Ω: it has **no meaning without** Ω. To remind ourselves of this, we can write

$$\mathbb{P} = \mathbb{P}_\Omega.$$

Then $$\mathbb{P}(B) = \mathbb{P}(B \,|\, \Omega) = \mathbb{P}_\Omega(B).$$

Similarly, $\mathbb{P}(B \,|\, A)$ means that we are looking for the probability of event B, out of all possible outcomes in the set **A**.

So A is just another sample space. Thus *we can manipulate conditional probabilities $\mathbb{P}(\cdot \,|\, A)$ just like any other probabilities, <u>as long as</u> we always stay inside the same sample space A.*

The trick: Because we can think of A as just another sample space, let's write

$$\mathbb{P}(\cdot \,|\, A) = \mathbb{P}_A(\cdot)$$

Note: NOT
standard notation!

Then **we can use $\mathbb{P}_A$ just like $\mathbb{P}$, as long as we remember to keep the A subscript on EVERY $\mathbb{P}$ that we write.**

This helps us to make quite complex manipulations of conditional probabilities without thinking too hard or making mistakes. There is only one rule you need to learn to use this tool effectively:

$$\mathbb{P}_A(B \,|\, C) = \mathbb{P}(B \,|\, C \cap A) \textbf{ for any } A, B, C.$$

(Proof: Exercise).

The rules:

$$\mathbb{P}(\cdot \,|\, A) = \mathbb{P}_A(\cdot)$$

$$\mathbb{P}_A(B \,|\, C) = \mathbb{P}(B \,|\, C \cap A) \text{ for any } A, B, C.$$

Examples:

1. Probability of a union. In general,

   $$\mathbb{P}(B \cup C) = \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(B \cap C).$$

   So, $\mathbb{P}_A(B \cup C) = \mathbb{P}_A(B) + \mathbb{P}_A(C) - \mathbb{P}_A(B \cap C).$

   Thus, $\mathbb{P}(B \cup C \,|\, A) = \mathbb{P}(B \,|\, A) + \mathbb{P}(C \,|\, A) - \mathbb{P}(B \cap C \,|\, A).$

2. Which of the following is equal to $\mathbb{P}(B \cap C \,|\, A)$?

   (a) $\mathbb{P}(B \,|\, C \cap A).$ (c) $\mathbb{P}(B \,|\, C \cap A)\mathbb{P}(C \,|\, A).$

   (b) $\dfrac{\mathbb{P}(B \,|\, C)}{\mathbb{P}(A)}.$ (d) $\mathbb{P}(B \,|\, C)\mathbb{P}(C \,|\, A).$

   Solution:

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

   Thus the correct answer is (c).

3. Which of the following is true?

(a) $\mathbb{P}(\overline{B}\,|\,A) = 1 - \mathbb{P}(B\,|\,A)$.

(b) $\mathbb{P}(\overline{B}\,|\,A) = \mathbb{P}(B) - \mathbb{P}(B\,|\,A)$.

Solution:

$$\mathbb{P}(\overline{B}\,|\,A) = \mathbb{P}_A(\overline{B}) = 1 - \mathbb{P}_A(B) = 1 - \mathbb{P}(B\,|\,A).$$

Thus the correct answer is (a).

4. Which of the following is true?

(a) $\mathbb{P}(\overline{B} \cap A) = \mathbb{P}(A) - \mathbb{P}(B \cap A)$.

(b) $\mathbb{P}(\overline{B} \cap A) = \mathbb{P}(B) - \mathbb{P}(B \cap A)$.

Solution:

$$\begin{aligned}
\mathbb{P}(\overline{B} \cap A) = \mathbb{P}(\overline{B}\,|\,A)\mathbb{P}(A) &= \mathbb{P}_A(\overline{B})\mathbb{P}(A) \\
&= \left(1 - \mathbb{P}_A(B)\right)\mathbb{P}(A) \\
&= \mathbb{P}(A) - \mathbb{P}(B\,|\,A)\mathbb{P}(A) \\
&= \mathbb{P}(A) - \mathbb{P}(B \cap A).
\end{aligned}$$

Thus the correct answer is (a).

5. True or false: $\mathbb{P}(B\,|\,A) = 1 - \mathbb{P}(B\,|\,\overline{A})$?

Answer: *False.* $\mathbb{P}(B\,|\,A) = \mathbb{P}_A(B)$. *Once we have $\mathbb{P}_A$, we are stuck with it! There is no easy way of converting from $\mathbb{P}_A$ to $\mathbb{P}_{\overline{A}}$: or anything else. Probabilities in one sample space ($\mathbb{P}_A$) cannot tell us anything about probabilities in a different sample space ($\mathbb{P}_{\overline{A}}$).*

Exercise: if we wish to express $\mathbb{P}(B\,|\,A)$ in terms of only B and $\overline{A}$, show that
$$\mathbb{P}(B\,|\,A) = \frac{\mathbb{P}(B) - \mathbb{P}(B\,|\,\overline{A})\mathbb{P}(\overline{A})}{1 - \mathbb{P}(\overline{A})}.$$
Note that this does not simplify nicely!

2.4 The Partition Theorem (Law of Total Probability)

Definition: Events A and B are **mutually exclusive**, or **disjoint**, if $A \cap B = \emptyset$.

This means events A and B cannot happen together. If A happens, it excludes B from happening, and vice-versa.
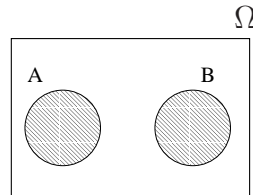
If A and B are mutually exclusive, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
For all other A and B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Definition: Any number of events $B_1, B_2, \ldots, B_k$ are **mutually exclusive** if every pair of the events is mutually exclusive: ie. $B_i \cap B_j = \emptyset$ *for all* i, j *with* $i \neq j$.
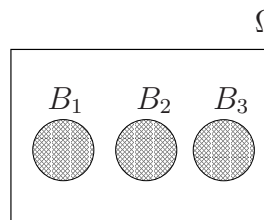
Definition: A **partition** of Ω is a *collection of mutually exclusive events whose union is Ω.*

That is, sets $B_1, B_2, \ldots, B_k$ form a partition of Ω if

$$B_i \cap B_j \;=\; \emptyset \text{ for all } i, j \text{ with } i \neq j\,,$$

$$\textbf{\underline{and}} \quad \bigcup_{i=1}^{k} B_i \;=\; B_1 \cup B_2 \cup \ldots \cup B_k \;=\; \Omega.$$

> $B_1, \ldots, B_k$ form a partition of Ω if they *have no overlap*
> *and collectively cover all possible outcomes.*

Examples:
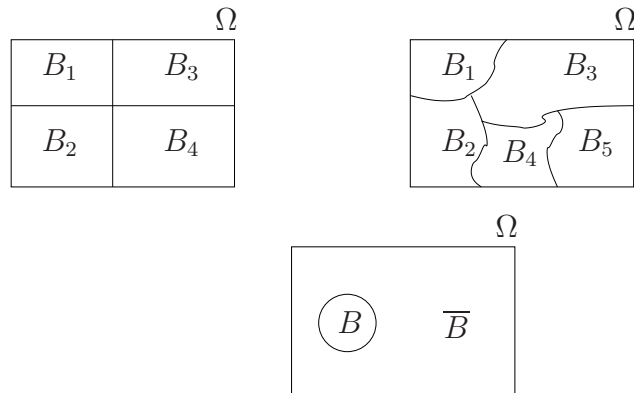
Partitioning an event A

Any set A can be partitioned: it doesn't have to be Ω.

In particular, if $B_1, \ldots, B_k$ form a partition of Ω, then $(A \cap B_1), \ldots, (A \cap B_k)$ form a partition of A.
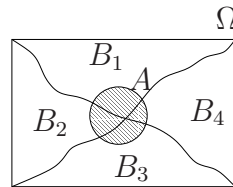
Theorem 2.4: The Partition Theorem (Law of Total Probability)

Let $B_1, \ldots, B_m$ form a partition of Ω. Then for any event A,

$$\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{m} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)$$

Both formulations of the Partition Theorem are very widely used, but especially the conditional formulation $\sum_{i=1}^{m} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)$.

Intuition behind the Partition Theorem:

The Partition Theorem is easy to understand because it simply states that "the whole is the sum of its parts."

$A \cap B_1$ $A \cap B_2$

$A \cap B_3$ $A \cap B_4$

A
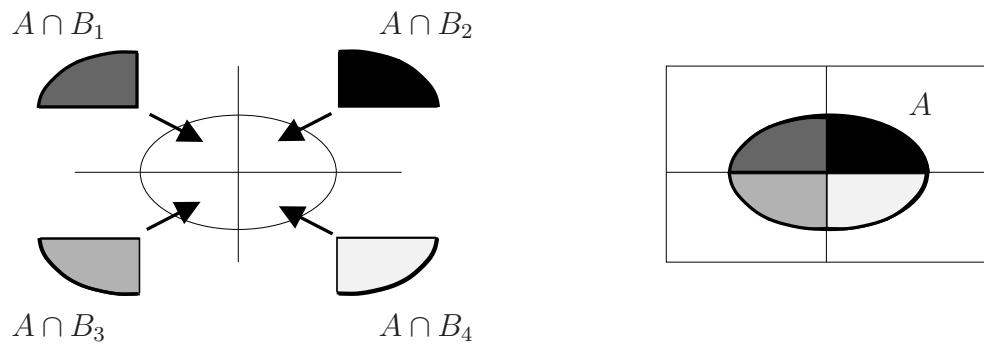
$$\mathbb{P}(A) = \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) + \mathbb{P}(A \cap B_3) + \mathbb{P}(A \cap B_4).$$

2.5 Bayes' Theorem: inverting conditional probabilities

Bayes' Theorem allows us to "invert" a conditional statement, ie. *to express* $\mathbb{P}(B \mid A)$ *in terms of* $\mathbb{P}(A \mid B)$.

Theorem 2.5: Bayes' Theorem

For any events A and B:
$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

Proof:

$$
\begin{aligned}
\mathbb{P}(B \cap A) &= \mathbb{P}(A \cap B) \\
\mathbb{P}(B \mid A)\mathbb{P}(A) &= \mathbb{P}(A \mid B)\mathbb{P}(B) \qquad \text{(multiplication rule)} \\
\therefore \qquad \mathbb{P}(B \mid A) &= \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}. \qquad \square
\end{aligned}
$$

Extension of Bayes' Theorem

Suppose that $B_1, B_2, \ldots, B_m$ form a partition of Ω. By the Partition Theorem,

$$\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Thus, for **any single partition member** B_j, put $B = B_j$ in Bayes' Theorem to obtain:

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$
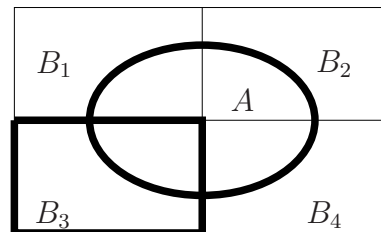
Special case: $m = 2$

Given any event B, the events B and $\overline{B}$ form a partition of Ω. Thus:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid \overline{B})\mathbb{P}(\overline{B})}.$$

Example: In screening for a certain disease, the probability that a healthy person wrongly gets a positive result is 0.05. The probability that a diseased person wrongly gets a negative result is 0.002. The overall rate of the disease in the population being screened is 1%. If my test gives a positive result, what is the probability I actually have the disease?

1. Define events:

$$D = \{have\ disease\} \qquad \overline{D} = \{do\ not\ have\ the\ disease\}$$

$$P = \{positive\ test\} \qquad N = \overline{P} = \{negative\ test\}$$

2. Information given:

False positive rate is 0.05 $\quad \Rightarrow \quad \mathbb{P}(P \mid \overline{D}) = 0.05$

False negative rate is 0.002 $\quad \Rightarrow \quad \mathbb{P}(N \mid D) = 0.002$

Disease rate is 1% $\quad \Rightarrow \quad \mathbb{P}(D) = 0.01.$

3. Looking for $\mathbb{P}(D \mid P)$*:*

$$\textit{We have} \quad \mathbb{P}(D \mid P) \;=\; \frac{\mathbb{P}(P \mid D)\mathbb{P}(D)}{\mathbb{P}(P)}.$$

$$\begin{aligned}
\textit{Now} \quad \mathbb{P}(P \mid D) &= 1 - \mathbb{P}(\overline{P} \mid D) \\
&= 1 - \mathbb{P}(N \mid D) \\
&= 1 - 0.002 \\
&= 0.998.
\end{aligned}$$

$$\begin{aligned}
\textit{Also} \quad \mathbb{P}(P) &= \mathbb{P}(P \mid D)\mathbb{P}(D) + \mathbb{P}(P \mid \overline{D})\mathbb{P}(\overline{D}) \\
&= 0.998 \times 0.01 + 0.05 \times (1 - 0.01) \\
&= 0.05948.
\end{aligned}$$

Thus

$$\mathbb{P}(D \mid P) = \frac{0.998 \times 0.01}{0.05948} = 0.168.$$

Given a positive test, my chance of having the disease is only 16.8%.

2.6 First-Step Analysis for calculating probabilities in a process

In a stochastic process, what happens at the next step depends upon the current state of the process. We are often interested to know the probability of *eventually* reaching some particular state, given our current position.

Throughout this course, we will tackle this sort of problem using a technique called *First-Step Analysis.*

The idea is to consider all possible first steps away from the current state. We derive a system of equations that specify the probability of the eventual outcome given each of the possible first steps. We then try to solve these equations for the probability of interest.

First-Step Analysis depends upon **conditional probability and the Partition Theorem.** Let $S_1, \ldots, S_k$ be the k possible first steps we can take away from our current state. We wish to find the probability that event E happens eventually. First-Step Analysis calculates $\mathbb{P}(E)$ as follows:

$$\mathbb{P}(E) = \mathbb{P}(E|S_1)\mathbb{P}(S_1) + \ldots + \mathbb{P}(E|S_k)\mathbb{P}(S_k).$$

Here, $\mathbb{P}(S_1), \ldots, \mathbb{P}(S_k)$ give the probabilities of taking the different first steps $1, 2, \ldots, k$.

Example 1: Tennis game at Deuce.

Venus and Serena are playing tennis, and have reached the score Deuce (40-40). (*Deuce* comes from the French word *Deux* for 'two', meaning that each player needs to win two consecutive points to win the game.)

For each point, let:

$$p = \mathbb{P}(\text{Venus wins point}), \qquad q = 1 - p = \mathbb{P}(\text{Serena wins point}).$$

Assume that all points are independent.

Let v be the probability that Venus wins the game eventually, starting from Deuce. Find v.
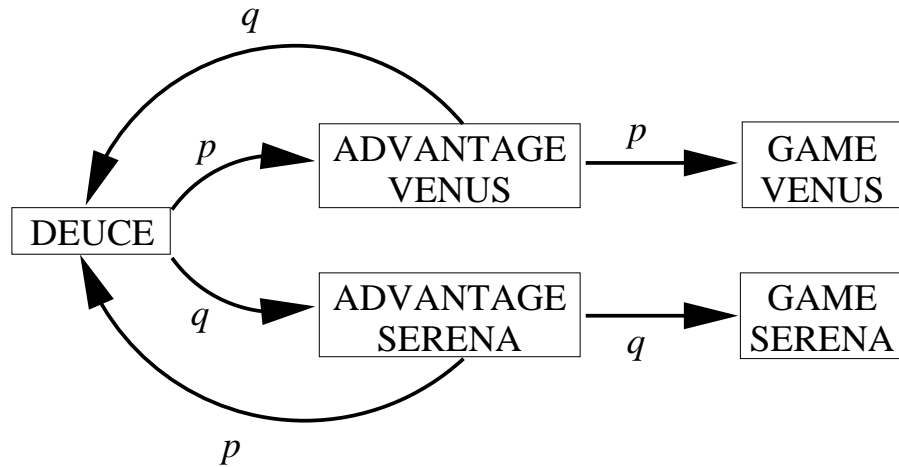
Use First-step analysis. The possible first steps starting from Deuce are:

1. *Venus wins the next point (probability p): state AV;*

2. *Serena wins the next point (probability q): state AS.*

Thus:

$$v = \mathbb{P}(\textit{Venus wins} \,|\, \textit{Deuce}) \;=\; \mathbb{P}_D(VW \,|\, AV)\mathbb{P}_D(AV) + \mathbb{P}_D(VW \,|\, AS)\mathbb{P}_D(AS)$$

$$= \; \mathbb{P}(VW \,|\, AV)p + \mathbb{P}(VW \,|\, AS)q. \qquad (\star)$$

Now we need to find $\mathbb{P}(VW \,|\, AV)$, and $\mathbb{P}(VW \,|\, AS)$, again using First-step analysis:

$$\mathbb{P}(VW \,|\, AV) \;=\; \mathbb{P}(VW \,|\, \textit{Game Venus})p + \mathbb{P}(VW \,|\, \textit{Deuce})q$$

$$= \; 1 \times p + v \times q$$

$$= \; p + qv. \qquad (a)$$

Similarly,

$$\mathbb{P}(VW \,|\, AS) \;=\; \mathbb{P}(VW \,|\, \textit{Game Serena})q + \mathbb{P}(VW \,|\, \textit{Deuce})p$$

$$= \; 0 \times q + v \times p$$

$$= \; pv. \qquad (b)$$

Substituting (a) and (b) into $(\star)$,

$$\begin{aligned}
v &= (p+qv)p + (pv)q \\
v &= p^2 + 2pqv \\
v(1-2pq) &= p^2 \\
v &= \frac{p^2}{1-2pq}.
\end{aligned}$$

Note: Because $p + q = 1$, we have:

$$1 = (p+q)^2 = p^2 + q^2 + 2pq.$$

So the final probability that Venus wins the game is:

$$v = \frac{p^2}{1-2pq} = \frac{p^2}{p^2 + q^2}.$$

Note how this result makes intuitive sense. For the game to finish from Deuce, either Venus has to win two points in a row (probability p^2), or Serena does (probability q^2). The ratio $p^2/(p^2+q^2)$ describes Venus's 'share' of the winning probability.

Alternative approach: two-step analysis

The diagram above is a simple one, and we can see that the *only* way of winning from Deuce is to take two steps away from Deuce. In situations like these, it is worth considering whether a two-step approach is possible.

For a two-step approach, we look at the ***possible <u>pairs</u> of steps we can take out of the state Deuce.***

Pair of steps	Probability	Outcome
$AV \to GV$	p^2	*Venus wins*
$AS \to GS$	q^2	Serena wins
$AV \to D$	pq	*back to start (Deuce)*
$AS \to D$	qp	*back to start (Deuce)*

Our approach to finding $v = \mathbb{P}(\text{Venus wins})$ can be summarized as:

$$\mathbb{P}(\textit{Venus wins}) = v = \sum_{\textit{options}} \mathbb{P}(\textit{Venus wins} \,|\, \textit{option})\mathbb{P}(\textit{option})$$

$$v = 1 \times p^2 + 0 \times q^2 + v \times pq + v \times qp$$

$$v = p^2 + 2pqv$$

$$v(1 - 2pq) = p^2$$

$$v = \frac{p^2}{1 - 2pq} \qquad \text{as before.}$$

With permission, copyright Martin Ouellet

Example 2: **Gambler's Ruin.**

This is a famous problem in probability. A gambler starts with $\$x$. She tosses a fair coin repeatedly.

If she gets a Head, she wins $\$1$. If she gets a Tail, she loses $\$1$.

The coin tossing is repeated until the gambler has either $\$0$ or $\$N$. What is the probability of the Gambler's Ruin (i.e. that the gambler ends up with $\$0$)?

Define events:

$$S_x = \{\textit{starts with } \$x\} \qquad R = \{\textit{ends with ruin}\}$$

Looking for:

$$\mathbb{P}(\textit{ruin} \,|\, \textit{starts with } x) = \mathbb{P}(R \,|\, S_x)$$

Define: $\qquad\qquad p_x = \mathbb{P}(R \,|\, S_x).$

What we know:

$$p_0 = \mathbb{P}(\textit{ruin} \,|\, \textit{starts with } \$0) = 1,$$
$$p_N = \mathbb{P}(\textit{ruin} \,|\, \textit{starts with } \$N) = 0.$$

Possible steps from state S_x	Probability	Outcome
Throw Head (event H)	0.5	Win \$1, start again from \$(x+1): State S_{x+1}
Throw Tail (event T)	0.5	Lose \$1, start again from \$(x-1): State S_{x-1}

First-step analysis:

$$
\begin{aligned}
p_x &= \mathbb{P}(R \,|\, S_x) \\
&= \mathbb{P}_{S_x}(R) \qquad \text{in notation of Section 2.3} \\
&= \mathbb{P}_{S_x}(R \,|\, H)\mathbb{P}_{S_x}(H) \;+\; \mathbb{P}_{S_x}(R \,|\, T)\mathbb{P}_{S_x}(T)
\end{aligned}
$$

So:

$$
\begin{aligned}
p_x &= \mathbb{P}_{S_x}(R \,|\, H) \times \tfrac{1}{2} \;+\; \mathbb{P}_{S_x}(R \,|\, T) \times \tfrac{1}{2} \\
&= \mathbb{P}(R \,|\, S_{x+1}) \times \tfrac{1}{2} \;+\; \mathbb{P}(R \,|\, S_{x-1}) \times \tfrac{1}{2} \\
p_x &= \tfrac{1}{2}p_{x+1} \;+\; \tfrac{1}{2}p_{x-1}. \qquad\qquad (\star)
\end{aligned}
$$

True for $1 \le x \le N - 1$ with boundary conditions $p_0 = 1$, $p_N = 0$.

Solution of difference equation $(\star)$:

We have:

$$
\begin{aligned}
p_x &= \tfrac{1}{2}p_{x+1} \;+\; \tfrac{1}{2}p_{x-1} \\
\tfrac{1}{2}p_x + \tfrac{1}{2}p_x &= \tfrac{1}{2}p_{x+1} \;+\; \tfrac{1}{2}p_{x-1} \\
\text{Rearranging:} \qquad p_{x-1} - p_x &= p_x - p_{x+1}.
\end{aligned}
$$

Boundaries: $p_0 = 1$, $p_N = 0$.

There are several ways of solving this equation.

1. By inspection

There are N steps to go down from $p_0 = 1$ to $p_N = 0$.
Each step is the same size, because
$(p_{x-1} - p_x) = (p_x - p_{x+1})$ *for all x.*
So each step has size $1/N$,
$\Rightarrow \quad p_0 = 1, \ p_1 = 1 - 1/N,$
$\qquad p_2 = 1 - 2/N,$ *etc.*

So $\boxed{p_x = 1 - \dfrac{x}{N}.}$
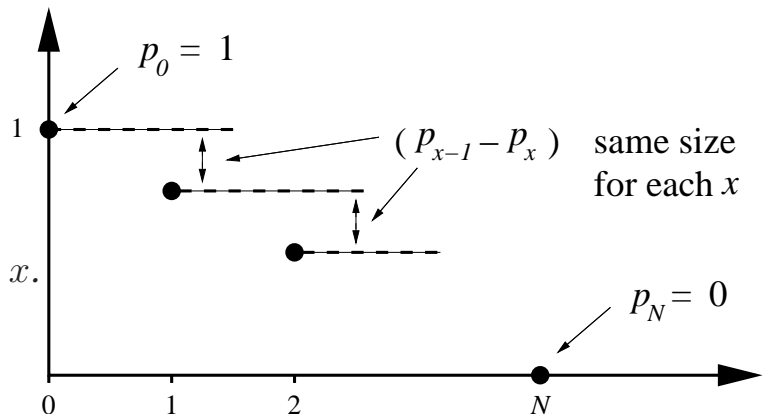
2. Theory of linear 2nd order difference equations

Theory tells us that the general solution of $(\star)$ is $p_x = A + Bx$ for some constants A, B.

Boundary conditions:

$$
\begin{aligned}
p_0 &= A + B \times 0 = 1 \quad \Rightarrow \quad A = 1 \\
p_N &= A + B \times N = 1 + BN = 0 \quad \Rightarrow \quad B = \frac{-1}{N}.
\end{aligned}
$$

So $\boxed{p_x = A + B\,x = 1 - \dfrac{x}{N}}$ *as before.*

3. Repeated substitution: rearrange $(\star)$ to give:

$$
\begin{aligned}
p_{x+1} &= 2p_x - p_{x-1} \\
\Rightarrow \quad (x=1) \qquad p_2 &= 2p_1 - 1 \qquad \text{(recall } p_0 = 1) \\
(x=2) \qquad p_3 &= 2p_2 - p_1 = 2(2p_1 - 1) - p_1 = 3p_1 - 2 \\
(x=3) \qquad p_4 &= 2p_3 - p_2 = 2(3p_1 - 2) - (2p_1 - 1) = 4p_1 - 3 \qquad etc
\end{aligned}
$$

$$\text{\textit{giving}} \qquad p_x \;=\; x p_1 - (x-1) \qquad \text{\textit{in general,}} \qquad (\star\star)$$

$$\text{\textit{likewise}} \qquad p_N \;=\; N p_1 - (N-1) \qquad \text{\textit{at endpoint.}}$$

Boundary condition: $\quad p_N = 0 \quad \Rightarrow \quad N p_1 - (N-1) = 0 \quad \Rightarrow \quad p_1 = 1 - 1/N.$

Substitute in $(\star\star)$*:*

$$
\begin{aligned}
p_x \;&=\; x p_1 - (x-1) \\
&=\; x\left(1 - \tfrac{1}{N}\right) - (x-1) \\
&=\; x - \tfrac{x}{N} - x + 1 \\
p_x \;&=\; 1 - \tfrac{x}{N} \qquad \text{\textit{as before.}} \qquad \square
\end{aligned}
$$

2.7 Independence

Definition: Events A and B are **statistically independent** if and only if

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

This implies that A and B are statistically independent if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A).$

Note: If events are *physically* independent, they will also be statistically indept.

For interest: more than two events

Definition: For more than two events, $A_1, A_2, \ldots, A_n$, we say that $A_1, A_2, \ldots, A_n$ are **mutually independent** if

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i) \qquad \text{\textit{for ALL finite subsets }} J \subseteq \{1, 2, \ldots, n\}.$$

Example: events A_1, A_2, A_3, A_4 are mutually independent if

    i) $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all i, j with $i \neq j$; AND

    ii) $\mathbb{P}(A_i \cap A_j \cap A_k) = \mathbb{P}(A_i)\mathbb{P}(A_j)\mathbb{P}(A_k)$ for all i, j, k that are all different; AND

    iii) $\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)\mathbb{P}(A_4)$.

Note: For mutual independence, it is ***not*** enough to check that $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $i \neq j$. Pairwise independence does not imply mutual independence.

Definition: Events, $A_1, A_2, \ldots, A_n$ are **pairwise independent** if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \quad \textit{for all } i \neq j.$$

Pairwise independence does NOT imply mutual independence. See example in Stats 210 notes.

2.8   The Continuity Theorem

The Continuity Theorem states that probability is a *continuous set function*:

Theorem 2.8:   The Continuity Theorem

a) Let $A_1, A_2, \ldots$ be an *increasing sequence of events*: i.e.

$$A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n \subseteq A_{n+1} \subseteq \ldots .$$

Then

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

Note: because $A_1 \subseteq A_2 \subseteq \ldots$, we have: $\quad \lim_{n \to \infty} A_n = \bigcup_{n=1}^{\infty} A_n.$

b) Let $B_1, B_2, \ldots$ be a *decreasing sequence of events*: i.e.

$$B_1 \supseteq B_2 \supseteq \ldots \supseteq B_n \supseteq B_{n+1} \supseteq \ldots .$$

Then

$$\mathbb{P}\left(\lim_{n\to\infty} B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n).$$

Note: because $B_1 \supseteq B_2 \supseteq \ldots$, we have: $\quad \lim_{n\to\infty} B_n = \bigcap_{n=1}^{\infty} B_n.$

Proof (a) only: for (b), take complements and use (a).

Define $C_1 = A_1$, and $C_i = A_i \backslash A_{i-1}$ for $i = 2, 3, \ldots$. Then $C_1, C_2, \ldots$ are mutually exclusive, and $\bigcup_{i=1}^{n} C_i = \bigcup_{i=1}^{n} A_i$, and likewise, $\bigcup_{i=1}^{\infty} C_i = \bigcup_{i=1}^{\infty} A_i$.

Thus

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} C_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(C_i) \quad (C_i \text{ mutually exclusive})$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(C_i)$$

$$= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} C_i\right)$$

$$= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \lim_{n\to\infty} \mathbb{P}(A_n). \quad \square$$

2.9 Random Variables

Definition: A **random variable**, X, is defined as *a function from the sample space to the real numbers:* $X : \Omega \to \mathbb{R}$.

A random variable therefore *assigns a real number to every possible outcome of a random experiment.*

A random variable is essentially *a rule or mechanism for generating random real numbers.*

The Distribution Function

Definition: The **cumulative distribution function** of a random variable X is given by

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

$F_X(x)$ is often referred to as simply the **distribution function**.

Properties of the distribution function

1) $F_X(-\infty) = \mathbb{P}(X \leq -\infty) = 0$.
   $F_X(+\infty) = \mathbb{P}(X \leq \infty) = 1$.

2) $F_X(x)$ is a non-decreasing function of x:
   if $x_1 < x_2$*, then* $F_X(x_1) \leq F_X(x_2)$.

3) If $b > a$, then $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.

4) F_X is right-continuous: i.e. $\lim_{h \downarrow 0} F_X(x + h) = F_X(x)$.

2.10 Continuous Random Variables

Definition: The random variable X is **continuous** if *the distribution function $F_X(x)$ is a continuous function.*

In practice, this means that a continuous random variable *takes values in a continuous subset of $\mathbb{R}$: e.g. $X : \Omega \to [0, 1]$ or $X : \Omega \to [0, \infty)$.*
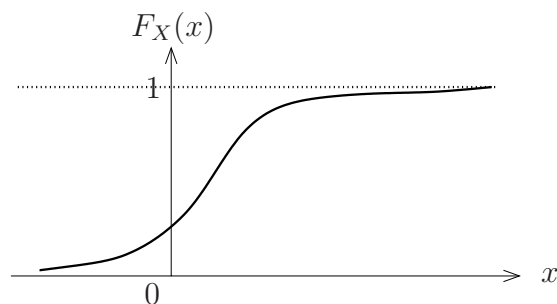
Probability Density Function for continuous random variables

Definition: Let X be a continuous random variable with continuous distribution function $F_X(x)$. The **probability density function (p.d.f.)** of X is defined as

$$f_X(x) = F'_X(x) = \frac{d}{dx}(F_X(x))$$

> The pdf, $f_X(x)$, gives the *shape* of the distribution of X.
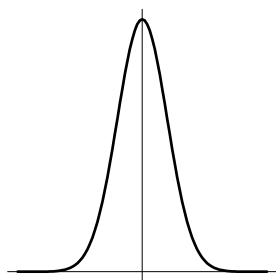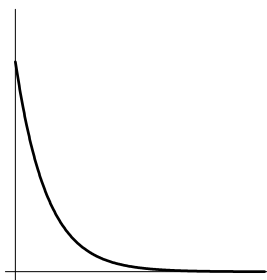
Normal distribution   Exponential distribution   Gamma distribution
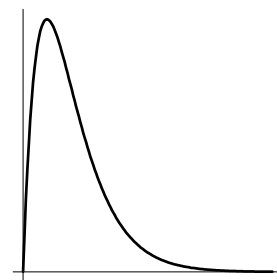
By the Fundamental Theorem of Calculus, the distribution function $F_X(x)$ can be written in terms of the probability density function, $f_X(x)$, as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$$

Endpoints of intervals

For continuous random variables, every point x has $\mathbb{P}(X = x) = 0$. This means that the endpoints of intervals are not important for continuous random variables.

Thus, $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b)$.

This is **only** true for **continuous** random variables.

Calculating probabilities for continuous random variables

To calculate $\mathbb{P}(a \leq X \leq b)$, use **either**

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$$

or

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx$$

Example: Let X be a continuous random variable with p.d.f.

$$f_X(x) = \begin{cases} 2x^{-2} & \text{for } 1 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the cumulative distribution function, $F_X(x)$.

(b) Find $\mathbb{P}(X \leq 1.5)$.

a) $\quad F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = \int_{1}^{x} 2u^{-2}\, du = \left[\dfrac{2u^{-1}}{-1}\right]_{1}^{x} = 2 - \dfrac{2}{x}$ $\qquad$ *for* $1 < x < 2$.

Thus $\qquad F_X(x) = \begin{cases} 0 & \textit{for } x \leq 1, \\ 2 - \frac{2}{x} & \textit{for } 1 < x < 2, \\ 1 & \textit{for } x \geq 2. \end{cases}$

b) $\mathbb{P}(X \leq 1.5) = F_X(1.5) = 2 - \dfrac{2}{1.5} = \dfrac{2}{3}.$

2.11 Discrete Random Variables

Definition: The random variable X is **discrete** if X *takes values in a finite or countable subset of* $\mathbb{R}$*: thus,* $X : \Omega \to \{x_1, x_2, \ldots\}$.

When X is a discrete random variable, the distribution function $F_X(x)$ is a *step function.*
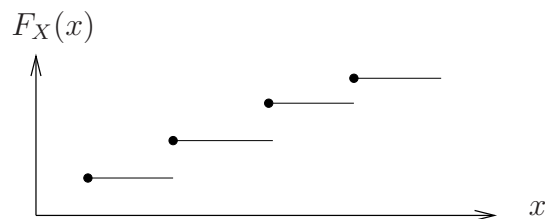
Probability function

Definition: Let X be a discrete random variable with distribution function $F_X(x)$. The **probability function** of X is defined as

$$f_X(x) = \mathbb{P}(X = x).$$

Endpoints of intervals

For discrete random variables, *individual points can have* $\mathbb{P}(X = x) > 0$.

This means that *the endpoints of intervals ARE important for discrete random variables.*

For example, if X takes values $0, 1, 2, \ldots$, and a, b are integers with $b > a$, then

$$\mathbb{P}(a \le X \le b) = \mathbb{P}(a-1 < X \le b) = \mathbb{P}(a \le X < b+1) = \mathbb{P}(a-1 < X < b+1).$$

Calculating probabilities for discrete random variables

To calculate $\mathbb{P}(X \in A)$ for any countable set A, use

$$\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x).$$

Partition Theorem for probabilities of discrete random variables

Recall the Partition Theorem: for any event A, and for events $B_1, B_2, \ldots$ that form a **partition** of Ω,
$$\mathbb{P}(A) = \sum_y \mathbb{P}(A \mid B_y)\mathbb{P}(B_y).$$

We can use the Partition Theorem to find probabilities for random variables. Let X and Y be **discrete** random variables.

- *Define event A as $A = \{X = x\}$.*

- *Define event B_y as $B_y = \{Y = y\}$ for $y = 0, 1, 2, \ldots$ (or whatever values Y takes).*

- *Then, by the Partition Theorem,*

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

2.12 Independent Random Variables

Random variables X and Y are independent if they have no effect on each other. This means that the probability that they both take specified values simultaneously is the product of the individual probabilities.

Definition: Let X and Y be random variables. The **joint distribution function** of X and Y is given by

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x \text{ and } Y \leq y) = \mathbb{P}(X \leq x, Y \leq y).$$

Definition: Let X and Y be any random variables (continuous or discrete). X and Y are **independent** if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \text{ for ALL } x,y \in \mathbb{R}.$$

If X and Y are **discrete**, they are independent if and only if their joint probability function is the product of their individual probability functions:

$$\text{Discrete } X, Y \text{ are indept} \iff \mathbb{P}(X = x \text{ AND } Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$
$$\text{for ALL } x, y$$
$$\iff f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \text{for ALL } x, y.$$

Chapter 3: Expectation and Variance

In the previous chapter we looked at probability, with three major themes:

1. Conditional probability: $\mathbb{P}(A \,|\, B)$.

2. First-step analysis for calculating eventual probabilities in a stochastic process.

3. Calculating probabilities for continuous and discrete random variables.

In this chapter, we look at the same themes for **expectation** and **variance**. The expectation of a random variable is the *long-term average of the random variable.*

Imagine observing many thousands of independent random values from the random variable of interest. Take the average of these random values. The expectation is the value of this average as the sample size tends to infinity.

We will repeat the three themes of the previous chapter, but in a different order.

1. Calculating expectations for continuous and discrete random variables.

2. Conditional expectation: the expectation of a random variable X, *conditional* on the value taken by another random variable Y. If the value of Y affects the value of X (i.e. X and Y are *dependent*), the conditional expectation of X given the value of Y will be different from the overall expectation of X.

3. First-step analysis for calculating the expected amount of time needed to reach a particular state in a process (e.g. the expected number of shots before we win a game of tennis).

We will also study similar themes for variance.

3.1 Expectation

The **mean**, **expected value**, or **expectation** of a random variable X is written as $\mathbb{E}(X)$ or μ_X. If we observe N random values of X, then the mean of the N values will be approximately equal to $\mathbb{E}(X)$ for large N. The expectation is defined differently for continuous and discrete random variables.

Definition: Let X be a **continuous** random variable with p.d.f. $f_X(x)$. The expected value of X is

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

Definition: Let X be a **discrete** random variable with probability function $f_X(x)$. The expected value of X is

$$\mathbb{E}(X) = \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x).$$

Expectation of $g(X)$

Let $g(X)$ be a function of X. We can imagine a long-term average of $g(X)$ just as we can imagine a long-term average of X. This average is written as $\mathbb{E}(g(X))$. Imagine observing X many times (N times) to give results $x_1, x_2, \ldots, x_N$. Apply the function g to each of these observations, to give $g(x_1), \ldots, g(x_N)$. The mean of $g(x_1), g(x_2), \ldots, g(x_N)$ approaches $\mathbb{E}(g(X))$ as the number of observations N tends to infinity.

Definition: Let X be a **continuous** random variable, and let g be a function. The expected value of $g(X)$ is

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

Definition: Let X be a **discrete** random variable, and let g be a function. The expected value of $g(X)$ is

$$\mathbb{E}\Big(g(X)\Big) = \sum_x g(x) f_X(x) = \sum_x g(x) \mathbb{P}(X = x).$$

Expectation of XY: the definition of $\mathbb{E}(XY)$

Suppose we have two random variables, X and Y. These might be independent, in which case the value of X has no effect on the value of Y. Alternatively, X and Y might be *dependent*: when we observe a random value for X, it might influence the random values of Y that we are most likely to observe. For example, X might be the height of a randomly selected person, and Y might be the weight. On the whole, larger values of X will be associated with larger values of Y.

To understand what $\mathbb{E}(XY)$ means, think of observing a large number of *pairs* $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. If X and Y are dependent, the value x_i might affect the value y_i, and vice versa, so we have to keep the observations together in their pairings. As the number of pairs N tends to infinity, the average $\frac{1}{N} \sum_{i=1}^{N} x_i \times y_i$ approaches the expectation $\mathbb{E}(XY)$.

For example, if X is height and Y is weight, $\mathbb{E}(XY)$ is the average of (height $\times$ weight). We are interested in $\mathbb{E}(XY)$ because it is used for calculating the *covariance* and *correlation*, which are measures of how closely related X and Y are (see Section 3.2).

Properties of Expectation

i) Let g and h be functions, and let a and b be constants. For any random variable X (discrete or continuous),
$$\mathbb{E}\Big\{ ag(X) + bh(X) \Big\} = a\mathbb{E}\Big\{ g(X) \Big\} + b\mathbb{E}\Big\{ h(X) \Big\}.$$
In particular,
$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

ii) Let X and Y be ANY random variables (discrete, continuous, independent, or non-independent). Then
$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

More generally, for ANY random variables $X_1, \ldots, X_n$,
$$\mathbb{E}(X_1 + \ldots + X_n) = \mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n).$$

iii) Let X and Y be **independent** random variables, and g, h be functions. Then

$$\mathbb{E}(XY) \;=\; \mathbb{E}(X)\mathbb{E}(Y)$$
$$\mathbb{E}\Big(g(X)h(Y)\Big) \;=\; \mathbb{E}\Big(g(X)\Big)\mathbb{E}\Big(h(Y)\Big).$$

Notes: 1. $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ is ONLY generally true if X and Y are *INDEPENDENT.*

2. If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. However, the converse is not generally true: it is possible for $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ even though X and Y are dependent.

Probability as an Expectation

Let A be any event. We can write $\mathbb{P}(A)$ as an expectation, as follows. Define the ***indicator function:***

$$I_A = \begin{cases} 1 & \text{\textit{if event A occurs,}} \\ 0 & \text{\textit{otherwise.}} \end{cases}$$

Then I_A is a ***random variable***, and

$$\begin{aligned} \mathbb{E}(I_A) \;&=\; \sum_{r=0}^{1} r\mathbb{P}(I_A = r) \\ &=\; 0 \times \mathbb{P}(I_A = 0) + 1 \times \mathbb{P}(I_A = 1) \\ &=\; \mathbb{P}(I_A = 1) \\ &=\; \mathbb{P}(A). \end{aligned}$$

Thus

$$\boxed{\mathbb{P}(A) = \mathbb{E}(I_A) \text{ \textit{for any event} } A.}$$

3.2 Variance, covariance, and correlation

The variance of a random variable X is a measure of how **spread out** it is. Are the values of X clustered tightly around their mean, or can we commonly observe values of X a long way from the mean value? The *variance* measures how far the values of X are from their mean, on average.

Definition: Let X be any random variable. The **variance** of X is

$$Var(X) = \mathbb{E}\Big((X - \mu_X)^2\Big) = \mathbb{E}(X^2) - \Big(\mathbb{E}(X)\Big)^2.$$

The variance is the **mean squared deviation** of a random variable from its own mean.

If X has *high variance*, we can observe values of X a long way from the mean.

If X has *low variance*, the values of X tend to be clustered tightly around the mean value.

Example: Let X be a continuous random variable with p.d.f.

$$f_X(x) = \begin{cases} 2x^{-2} & \text{for } 1 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{E}(X)$ and $Var(X)$.

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_{1}^{2} x \times 2x^{-2}\, dx \ = \ \int_{1}^{2} 2x^{-1}\, dx$$

$$= \ \Big[\, 2\log(x) \,\Big]_{1}^{2}$$

$$= \ 2\log(2) - 2\log(1)$$

$$= \ 2\log(2).$$

For Var(X), we use

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

Now

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \, f_X(x) \, dx = \int_1^2 x^2 \times 2x^{-2} \, dx \ = \ \int_1^2 2 \, dx$$

$$= \ \Big[\, 2x \,\Big]_1^2$$

$$= \ 2 \times 2 - 2 \times 1$$

$$= \ 2.$$

Thus

$$\begin{aligned}
\text{Var}(X) \ &= \ \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 \\
&= \ 2 - \{2\log(2)\}^2 \\
&= \ 0.0782.
\end{aligned}$$

Covariance

Covariance is a measure of the association or dependence between two random variables X and Y. Covariance can be either positive or negative. (*Variance* is always positive.)

Definition: Let X and Y be any random variables. The **covariance** between X and Y is given by

$$cov(X, Y) = \mathbb{E}\Big\{(X - \mu_X)(Y - \mu_Y)\Big\} = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

where $\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$.

1. $\text{cov}(X, Y)$ will be *positive* if large values of X tend to occur with large values of Y, and small values of X tend to occur with small values of Y. For example, if X is height and Y is weight of a randomly selected person, we would expect $\text{cov}(X, Y)$ to be positive.

2. $\text{cov}(X, Y)$ will be *negative* if large values of X tend to occur with small values of Y, and small values of X tend to occur with large values of Y. For example, if X is age of a randomly selected person, and Y is heart rate, we would expect X and Y to be negatively correlated (older people have slower heart rates).

3. If X and Y are independent, then there is no pattern between large values of X and large values of Y, so $\text{cov}(X, Y) = 0$. However, $\text{cov}(X, Y) = 0$ does NOT imply that X and Y are independent, unless X and Y are Normally distributed.

Properties of Variance

i) Let g be a function, and let a and b be constants. For any random variable X (discrete or continuous),

$$\textit{Var}\Big\{ag(X) + b\Big\} = a^2\, \textit{Var}\Big\{g(X)\Big\}.$$

In particular, $\textit{Var}(aX + b) = a^2\, \textit{Var}(X)$.

ii) Let X and Y be **independent** random variables. Then

$$\textit{Var}(X + Y) = \textit{Var}(X) + \textit{Var}(Y).$$

iii) If X and Y are **NOT independent**, then

$$\textit{Var}(X + Y) = \textit{Var}(X) + \textit{Var}(Y) + 2\textit{cov}(X, Y).$$

Correlation (non-examinable)

The correlation coefficient of X and Y is a measure of the linear association between X and Y. It is given by the covariance, scaled by the overall variability in X and Y. As a result, the correlation coefficient is always between -1 and $+1$, so it is easily compared for different quantities.

Definition: The **correlation** between X and Y, also called the **correlation coefficient**, is given by

$$\boxed{\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}\,.}$$

The correlation measures <u>linear</u> association between X and Y. It takes values only between -1 and $+1$, and has the same sign as the covariance.

The correlation is ± 1 if and only if there is a perfect linear relationship between X and Y, i.e. $\mathrm{corr}(X, Y) = 1 \iff Y = aX + b$ for some constants a and b.

The correlation is 0 if X and Y are independent, but a correlation of 0 does not *imply* that X and Y are independent.

3.3 Conditional Expectation and Conditional Variance

Throughout this section, we will assume for simplicity that X and Y are discrete random variables. However, exactly the same results hold for continuous random variables too.

Suppose that X and Y are discrete random variables, possibly dependent on each other. Suppose that we fix Y at the value y. This gives us a set of conditional probabilities $\mathbb{P}(X = x \,|\, Y = y)$ for all possible values x of X. This is called the *conditional distribution of X, given that $Y = y$.*

Definition: Let X and Y be discrete random variables. The **conditional probability function** of X, given that $Y = y$, is:

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

We write the conditional probability function as:

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

Note: The conditional probabilities $f_{X\,|\,Y}(x \,|\, y)$ sum to one, just like any other probability function:

$$\sum_x \mathbb{P}(X = x \,|\, Y = y) = \sum_x \mathbb{P}_{\{Y=y\}}(X = x) = 1,$$

using the subscript notation $\mathbb{P}_{\{Y=y\}}$ of Section 2.3.

We can also find the expectation and variance of X with respect to this conditional distribution. That is, if we know that the value of Y is fixed at y, then we can find the mean value of X *given that* Y takes the value y, and also the variance of X given that $Y = y$.

Definition: Let X and Y be discrete random variables. The **conditional expectation of X, given that $Y = y$**, is

$$\mu_{X \mid Y=y} = \mathbb{E}(X \mid Y = y) = \sum_x x f_{X \mid Y}(x \mid y).$$

$\mathbb{E}(X \mid Y = y)$ is the *mean value of X, when Y is fixed at* y.

Conditional expectation as a random variable

The unconditional expectation of X, $\mathbb{E}(X)$, is just *a number:*
e.g. $\mathbb{E}X = 2$ *or* $\mathbb{E}X = 5.8$.

The conditional expectation, $\mathbb{E}(X \mid Y = y)$, is *a number depending on y.*

If Y has an influence on the value of X, then Y will have an influence on the *average* value of X. So, for example, we would expect $\mathbb{E}(X \mid Y = 2)$ to be different from $\mathbb{E}(X \mid Y = 3)$.

We can therefore view $\mathbb{E}(X \mid Y = y)$ as a *function of y, say $\mathbb{E}(X \mid Y{=}y) = h(y)$.*

To evaluate this function, $h(y) = \mathbb{E}(X \mid Y = y)$, we:

i) *fix Y at the chosen value y;*

ii) *find the expectation of X when Y is fixed at this value.*

However, we could also evaluate the function at a *random value* of Y:

i) *observe a random value of Y;*

ii) *fix Y at that observed random value;*

iii) *evaluate $\mathbb{E}(X \mid Y = \text{observed random value}).$*

We obtain a <u>*random variable*</u>: $\mathbb{E}(X \mid Y) = h(Y).$

The randomness comes from the randomness in Y, not in X.

> *Conditional expectation, $\mathbb{E}(X \mid Y)$, is a random variable with randomness inherited from Y, not X.*

Example: Suppose $Y = \begin{cases} 1 & \text{with probability } 1/8\,, \\ 2 & \text{with probability } 7/8\,, \end{cases}$

and $\quad X \mid Y = \begin{cases} 2Y & \text{with probability } 3/4\,, \\ 3Y & \text{with probability } 1/4\,. \end{cases}$

Conditional expectation of X given $Y = y$ is a number depending on y:

If $Y = 1$, then: $X \mid (Y = 1) = \begin{cases} 2 & \text{with probability } 3/4 \\ 3 & \text{with probability } 1/4 \end{cases}$

$$so \quad \mathbb{E}(X \mid Y = 1) = 2 \times \tfrac{3}{4} + 3 \times \tfrac{1}{4} = \tfrac{9}{4}.$$

If $Y = 2$, then: $X \mid (Y = 2) = \begin{cases} 4 & \text{with probability } 3/4 \\ 6 & \text{with probability } 1/4 \end{cases}$

$$so \quad \mathbb{E}(X \mid Y = 2) = 4 \times \tfrac{3}{4} + 6 \times \tfrac{1}{4} = \tfrac{18}{4}.$$

Thus $\mathbb{E}(X \mid Y = y) = \begin{cases} 9/4 & \text{if } y = 1 \\ 18/4 & \text{if } y = 2. \end{cases}$

So $\mathbb{E}(X \mid Y = y)$ is a number depending on y, or a function of y.

Conditional expectation of X given random Y is a random variable:

From above, $\quad \mathbb{E}(X \,|\, Y) = \begin{cases} 9/4 & \textit{if } Y = 1 \textit{ (\textbf{probability} } 1/8), \\ 18/4 & \textit{if } Y = 2 \textit{ (\textbf{probability} } 7/8). \end{cases}$

So $\quad \mathbb{E}(X \,|\, Y) = \begin{cases} 9/4 & \textit{with probability } 1/8, \\ 18/4 & \textit{with probability } 7/8. \end{cases}$

Thus $\mathbb{E}(X \,|\, Y)$ is a <u>random variable</u>.

The randomness in $\mathbb{E}(X \,|\, Y)$ is inherited from Y, not from X.

Conditional expectation is a very useful tool for finding the **unconditional** expectation of X (see below). Just like the Partition Theorem, it is useful because it is often easier to specify conditional probabilities than to specify overall probabilities.

Conditional variance

The conditional variance is similar to the conditional expectation.

- $\text{Var}(X \,|\, Y = y)$ is the variance of X, when Y is fixed at the value $Y = y$.

- $\text{Var}(X \,|\, Y)$ is a random variable, giving the variance of X when Y is fixed at a value to be selected randomly.

Definition: Let X and Y be random variables. The **conditional variance of X, given Y,** is given by

$$\textit{Var}(X \,|\, Y) = \mathbb{E}(X^2 \,|\, Y) - \left\{\mathbb{E}(X \,|\, Y)\right\}^2 = \mathbb{E}\left\{(X - \mu_{X\,|\,Y})^2 \,|\, Y\right\}$$

Like expectation, $\textit{Var}(X \,|\, Y = y)$ is a number depending on y (a function of y), while $\textit{Var}(X \,|\, Y)$ is a <u>random variable</u> with randomness inherited from Y.

Laws of Total Expectation and Variance

If all the expectations below are finite, then for ANY random variables X and Y, we have:

i) $$\boxed{\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X \,|\, Y)\Big)}$$ *Law of Total Expectation.*

Note that we can pick any r.v. Y, to make the expectation as easy as we can.

ii) $\mathbb{E}(g(X)) = \mathbb{E}_Y\Big(\mathbb{E}(g(X) \,|\, Y)\Big)$ *for any function g.*

iii) $$\boxed{Var(X) = \mathbb{E}_Y\Big(Var(X \,|\, Y)\Big) + Var_Y\Big(\mathbb{E}(X \,|\, Y)\Big)}$$

Law of Total Variance.

Note: $\mathbb{E}_Y$ and Var_Y denote <u>expectation over Y</u> and <u>variance over Y</u>,

i.e. the expectation or variance is computed over the distribution of the random variable Y.

The Law of Total Expectation says that *the total average is the average of case-by-case averages.*

- The total average is $\mathbb{E}(X)$*;*

- The case-by-case averages are $\mathbb{E}(X \,|\, Y)$ *for the different values of Y ;*

- The average of case-by-case averages is *the average over Y of the Y-case averages:* $\mathbb{E}_Y\Big(\mathbb{E}(X \,|\, Y)\Big)$.

Example: In the example above, we had: $\mathbb{E}(X \,|\, Y) = \begin{cases} 9/4 & \text{with probability } 1/8, \\ 18/4 & \text{with probability } 7/8. \end{cases}$

The total average is:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X \,|\, Y)\Big\} = \frac{9}{4} \times \frac{1}{8} + \frac{18}{4} \times \frac{7}{8} = 4.22.$$

Proof of (i), (ii), (iii):

(i) is a special case of (ii), so we just need to prove (ii). Begin at RHS:

$$
\begin{aligned}
\text{RHS} \;=\; \mathbb{E}_Y\Big[\mathbb{E}(g(X) \,|\, Y)\Big] \;&=\; \mathbb{E}_Y\Big[\sum_x g(x)\mathbb{P}(X = x \,|\, Y)\Big] \\
&=\; \sum_y \Big[\sum_x g(x)\mathbb{P}(X = x \,|\, Y = y)\Big]\mathbb{P}(Y = y) \\
&=\; \sum_y \sum_x g(x)\mathbb{P}(X = x \,|\, Y = y)\mathbb{P}(Y = y) \\
&=\; \sum_x g(x) \sum_y \mathbb{P}(X = x \,|\, Y = y)\mathbb{P}(Y = y) \\
&=\; \sum_x g(x)\mathbb{P}(X = x) \quad \text{(partition rule)} \\
&=\; \mathbb{E}(g(X)) \quad=\quad \text{LHS.}
\end{aligned}
$$

(iii) Wish to prove $\text{Var}(X) = \mathbb{E}_Y[\text{Var}(X \,|\, Y)] + \text{Var}_Y[\mathbb{E}(X \,|\, Y)]$. Begin at RHS:

$\mathbb{E}_Y[\text{Var}(X \,|\, Y)] + \text{Var}_Y[\mathbb{E}(X \,|\, Y)]$

$$= \mathbb{E}_Y\Big\{\mathbb{E}(X^2 \,|\, Y) - (\mathbb{E}(X \,|\, Y))^2\Big\} + \Big\{\mathbb{E}_Y\Big\{[\mathbb{E}(X \,|\, Y)]^2\Big\} - \Big[\underbrace{\mathbb{E}_Y(\mathbb{E}(X \,|\, Y))}_{\mathbb{E}(X) \text{ by part (i)}}\Big]^2\Big\}$$

$$= \underbrace{\mathbb{E}_Y\{\mathbb{E}(X^2 \,|\, Y)\}}_{\mathbb{E}(X^2) \text{ by part (i)}} - \mathbb{E}_Y\{[\mathbb{E}(X \,|\, Y)]^2\} + \mathbb{E}_Y\{[\mathbb{E}(X \,|\, Y)]^2\} - (\mathbb{E}X)^2$$

$$= \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

$$= \text{Var}(X) = \text{LHS.} \qquad \square$$

3.4 Examples of Conditional Expectation and Variance

1. Swimming with dolphins

Fraser runs a dolphin-watch business.
Every day, he is unable to run the trip
due to bad weather with probability p,
independently of all other days. Fraser works every day except the bad-weather
days, which he takes as holiday.

Let Y be the number of consecutive days Fraser has to work between bad-weather days. Let X be the total number of customers who go on Fraser's trip in this period of Y days. Conditional on Y, the distribution of X is

$$(X \,|\, Y) \sim \text{Poisson}(\mu Y).$$

(a) Name the distribution of Y, and state $\mathbb{E}(Y)$ and $\text{Var}(Y)$.

(b) Find the expectation and the variance of the number of customers Fraser sees between bad-weather days, $\mathbb{E}(X)$ and $\text{Var}(X)$.

(a) *Let 'success' be 'bad-weather day' and 'failure' be 'work-day'.*
Then $\mathbb{P}(success) = \mathbb{P}(bad\text{-}weather) = p$.
Y is the number of failures before the first success.
So
$$Y \sim Geometric(p).$$

Thus

$$\mathbb{E}(Y) \;=\; \frac{1-p}{p},$$

$$Var(Y) \;=\; \frac{1-p}{p^2}.$$

(b) *We know $(X \,|\, Y) \sim Poisson(\mu Y)$: so*

$$\mathbb{E}(X \,|\, Y) = Var(X \,|\, Y) = \mu Y.$$

By the Law of Total Expectation:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X\,|\,Y)\Big\}$$
$$= \mathbb{E}_Y(\mu Y)$$
$$= \mu\mathbb{E}_Y(Y)$$

$$\therefore \quad \mathbb{E}(X) = \frac{\mu(1-p)}{p}\,.$$

By the Law of Total Variance:

$$\mathbf{Var}(X) = \mathbb{E}_Y\Big(\mathbf{Var}(X\,|\,Y)\Big) + \mathbf{Var}_Y\Big(\mathbb{E}(X\,|\,Y)\Big)$$

$$= \mathbb{E}_Y\Big(\mu Y\Big) + \mathbf{Var}_Y\Big(\mu Y\Big)$$

$$= \mu\mathbb{E}_Y(Y) + \mu^2\,\mathbf{Var}_Y(Y)$$

$$= \mu\left(\frac{1-p}{p}\right) + \mu^2\left(\frac{1-p}{p^2}\right)$$

$$= \frac{\mu(1-p)(p+\mu)}{p^2}\,.$$

Checking your answer in R:

If you know how to use a statistical package like R, you can check your answer to the question above as follows.

```
> # Pick a value for p, e.g. p = 0.2.
> # Pick a value for mu, e.g. mu = 25
>
> # Generate 10,000 random values of Y ~ Geometric(p = 0.2):
> y <- rgeom(10000, prob=0.2)
>
> # Generate 10,000 random values of X conditional on Y:
> # use (X | Y) ~ Poisson(mu * Y) ~ Poisson(25 * Y)
> x <- rpois(10000, lambda = 25*y)
```

```
> # Find the sample mean of X (should be close to E(X)):
> mean(x)
[1] 100.6606
>
> # Find the sample variance of X (should be close to var(X)):
> var(x)
[1] 12624.47
>
> # Check the formula for E(X):
> 25 * (1 - 0.2) / 0.2
[1] 100
>
> # Check the formula for var(X):
> 25 * (1 - 0.2) * (0.2 + 25) / 0.2^2
[1] 12600
```

The formulas we obtained by working give $\mathbb{E}(X) = 100$ and $\text{Var}(X) = 12600$. The sample mean was $\overline{x} = 100.6606$ (close to 100), and the sample variance was 12624.47 (close to 12600). Thus our working seems to have been correct.

---

2. Randomly stopped sum

This model arises very commonly in stochastic processes. A random number N of events occur, and each event i has associated with it some cost, penalty, or reward X_i. The question is to find the mean and variance of the total cost / reward:
$T_N = X_1 + X_2 + \ldots + X_N.$
The difficulty is that the number N of terms in the sum is itself random.

T_N is called a *randomly stopped sum: it is a sum of X_i's, randomly stopped at the random number of N terms.*

Example: Think of a cash machine, which has to be loaded with enough money to cover the day's business. The number of customers per day is a random number N. Customer i withdraws a random amount X_i. The total amount withdrawn during the day is a randomly stopped sum: $T_N = X_1 + \ldots + X_N$.

Cash machine example

The citizens of Remuera withdraw money from a cash machine according to the following probability function (X):

Amount, x (\$)	50	100	200
$\mathbb{P}(X = x)$	0.3	0.5	0.2

The number of customers per day has the distribution $N \sim \text{Poisson}(\lambda)$.

Let $T_N = X_1 + X_2 + \ldots + X_N$ be the total amount of money withdrawn in a day, where each X_i has the probability function above, and $X_1, X_2, \ldots$ are independent of each other and of N.

T_N is a randomly stopped sum, stopped by the random number of N customers.

(a) Show that $\mathbb{E}(X) = 105$, and $\text{Var}(X) = 2725$.

(b) Find $\mathbb{E}(T_N)$ and $\text{Var}(T_N)$: the mean and variance of the amount of money withdrawn each day.

Solution

(a) Exercise.

(b) *Let $T_N = \sum_{i=1}^{N} X_i$. If we knew how many terms were in the sum, we could easily find $\mathbb{E}(T_N)$ and $\text{Var}(T_N)$ as the mean and variance of a sum of independent r.v.s.*

So 'pretend' we know how many terms are in the sum: i.e. <u>condition</u> on N.

$$
\begin{aligned}
\mathbb{E}(T_N \,|\, N) &= \mathbb{E}(X_1 + X_2 + \ldots + X_N \,|\, N) \\
&= \mathbb{E}(X_1 + X_2 + \ldots + X_N) \\
&\qquad \text{(because all } X_i\text{s are independent of } N) \\
&= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_N) \\
&\qquad \text{where } N \text{ is now considered constant;} \\
&\qquad \text{(we do NOT need independence of } X_i\text{'s for this)} \\
&= N \times \mathbb{E}(X) \quad \text{(because all } X_i\text{'s have same mean, } \mathbb{E}(X)) \\
&= 105N.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathit{Var}(T_N \,|\, N) &= \mathit{Var}(X_1 + X_2 + \ldots + X_N \,|\, N) \\
&= \mathit{Var}(X_1 + X_2 + \ldots + X_N) \\
&\qquad\qquad \textit{where } N \textit{ is now considered constant;} \\
&\qquad\qquad \textit{(because all } X_i\textit{'s are independent of } N) \\
&= \mathit{Var}(X_1) + \mathit{Var}(X_2) + \ldots + \mathit{Var}(X_N) \\
&\qquad \textit{(we DO need independence of } X_i\textit{'s for this)} \\
&= N \times \mathit{Var}(X) \quad \textit{(because all } X_i\textit{'s have same variance, } \mathit{Var}(X)) \\
&= 2725N.
\end{aligned}
$$

So

$$
\begin{aligned}
\mathbb{E}(T_N) &= \mathbb{E}_N\left\{\mathbb{E}(T_N \,|\, N)\right\} \\
&= \mathbb{E}_N(105N) \\
&= 105\mathbb{E}_N(N) \\
&= 105\lambda,
\end{aligned}
$$

because $N \sim Poisson(\lambda)$ so $\mathbb{E}(N) = \lambda$.

Similarly,

$$
\begin{aligned}
\mathit{Var}(T_N) &= \mathbb{E}_N\left\{\mathit{Var}(T_N \,|\, N)\right\} + \mathit{Var}_N\left\{\mathbb{E}(T_N \,|\, N)\right\} \\
&= \mathbb{E}_N\left\{2725N\right\} + \mathit{Var}_N\left\{105N\right\} \\
&= 2725\mathbb{E}_N(N) + 105^2\mathit{Var}_N(N) \\
&= 2725\lambda + 11025\lambda \\
&= 13750\lambda,
\end{aligned}
$$

because $N \sim Poisson(\lambda)$ so $\mathbb{E}(N) = \mathit{Var}(N) = \lambda$.

Check in R (advanced)

```
> # Create a function tn.func to calculate a single value of T_N
> # for a given value N=n:
> tn.func <- function(n){
        sum(sample(c(50, 100, 200), n, replace=T,
        prob=c(0.3, 0.5, 0.2)))
  }


> # Generate 10,000 random values of N, using lambda=50:
> N <- rpois(10000, lambda=50)
> # Generate 10,000 random values of T_N, conditional on N:
> TN <- sapply(N, tn.func)
> # Find the sample mean of T_N values, which should be close to
> # 105 * 50 = 5250:
> mean(TN)
[1] 5253.255
> # Find the sample variance of T_N values, which should be close
> # to 13750 * 50 = 687500:
> var(TN)
[1] 682469.4
```

All seems well. Note that the sample variance is often some distance from the true variance, even when the sample size is 10,000.

General result for randomly stopped sums:

Suppose $X_1, X_2, \ldots$ each have the same mean μ and variance σ^2, and $X_1, X_2, \ldots$, and N are mutually independent. Let $T_N = X_1 + \ldots + X_N$ be the randomly stopped sum. By following similar working to that above:

$$
\mathbb{E}(T_N) = \mathbb{E}\left\{\sum_{i=1}^{N} X_i\right\} = \mu\,\mathbb{E}(N)
$$

$$
\mathrm{Var}(T_N) = \mathrm{Var}\left\{\sum_{i=1}^{N} X_i\right\} = \sigma^2\,\mathbb{E}(N) + \mu^2\,\mathrm{Var}(N).
$$

3.5 First-Step Analysis for calculating expected reaching times

Remember from Section 2.6 that we use First-Step Analysis for finding the probability of eventually reaching a particular state in a stochastic process. First-step analysis for probabilities uses *conditional probability and the Partition Theorem (Law of Total Probability).*

In the same way, we can use first-step analysis for finding the *expected reaching time for a state.*

This is the expected number of steps that will be needed to reach a particular state from a specified start-point, or the expected length of time it will take to get there if we have a continuous time process.

Just as first-step analysis for probabilities uses conditional probability and the law of total probability (Partition Theorem), first-step analysis for expectations uses *conditional expectation and the law of total expectation.*

First-step analysis for probabilities:

The first-step analysis procedure for probabilities can be summarized as follows:

$$\mathbb{P}(\textit{eventual goal}) = \sum_{\substack{\textit{first-step} \\ \textit{options}}} \mathbb{P}(\textit{eventual goal} \,|\, \textit{option})\mathbb{P}(\textit{option}) \,.$$

This is because the first-step options form a *partition of the sample space.*

First-step analysis for expected reaching times:

The expression for expected reaching times is very similar:

$$\mathbb{E}(\textit{reaching time}) = \sum_{\substack{\textit{first-step} \\ \textit{options}}} \mathbb{E}(\textit{reaching time} \,|\, \textit{option})\mathbb{P}(\textit{option}) \,.$$

This follows immediately from the law of total expectation:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X \mid Y)\Big\} = \sum_y \mathbb{E}(X \mid Y = y)\mathbb{P}(Y = y).$$

Let X be the reaching time, and let Y be the label for possible options:
i.e. $Y = 1, 2, 3, \ldots$ for options 1, 2, 3, …

We then obtain:

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

i.e. $\quad \mathbb{E}(\textit{reaching time}) = \sum_{\substack{\textit{first-step} \\ \textit{options}}} \mathbb{E}(\textit{reaching time} \mid \textit{option})\mathbb{P}(\textit{option}).$

Example: Mouse in a maze

A mouse is trapped in a room with three exits at the centre of a maze.

- Exit 1 leads outside the maze after 3 minutes.

- Exit 2 leads back to the room after 5 minutes.

- Exit 3 leads back to the room after 7 minutes.

Every time the mouse makes a choice, it is equally likely to choose any of the three exits. What is the expected time taken for the mouse to leave the maze?

Let X = time taken for mouse to leave maze.

Let Y = exit the mouse chooses first (1, 2, or 3).

Exit 2

5 mins   1/3

1/3

Room   → Exit 1

3 mins

7 mins   1/3

Exit 3
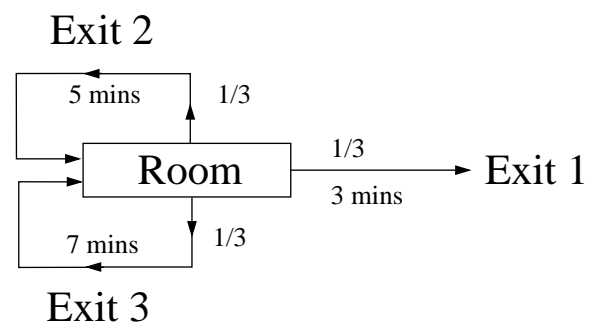
Then:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X\,|\,Y)\Big)$$

$$= \sum_{y=1}^{3} \mathbb{E}(X\,|\,Y=y)\mathbb{P}(Y=y)$$

$$= \mathbb{E}(X\,|\,Y=1)\times\tfrac{1}{3} + \mathbb{E}(X\,|\,Y=2)\times\tfrac{1}{3} + \mathbb{E}(X\,|\,Y=3)\times\tfrac{1}{3}.$$

But:

$$\mathbb{E}(X\,|\,Y=1) = 3 \text{ minutes}$$
$$\mathbb{E}(X\,|\,Y=2) = 5 + \mathbb{E}(X) \text{ (after 5 minutes, mouse is back at the beginning)}$$
$$\mathbb{E}(X\,|\,Y=3) = 7 + \mathbb{E}(X) \text{ (after 7 minutes, back to start)}$$

So

$$\mathbb{E}(X) = 3\times\tfrac{1}{3} + \Big(5+\mathbb{E}X\Big)\times\tfrac{1}{3} + \Big(7+\mathbb{E}X\Big)\times\tfrac{1}{3}$$

$$= 15\times\tfrac{1}{3} + 2(\mathbb{E}X)\times\tfrac{1}{3}$$

$$\tfrac{1}{3}\,\mathbb{E}(X) = 15\times\tfrac{1}{3}$$

$$\Rightarrow \quad \mathbb{E}(X) = 15 \text{ minutes.}$$

Two approaches to first-step analysis for expectations

In some problems, all possible first-step options incur the same initial time penalty. For example, in the mouse example, suppose that the mouse *always takes 0.5 minutes to choose an exit every time it enters the room.*
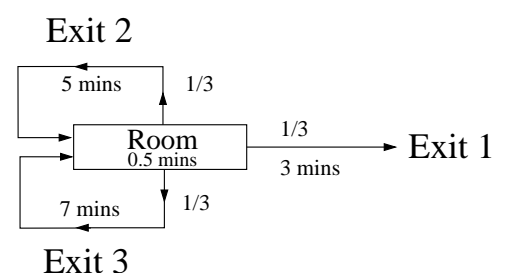
In this case, we have two choices:

1. *Add the extra 0.5 minutes onto each option separately:*

$$\mathbb{E}(X) = \left(0.5 + 3\right) \times \tfrac{1}{3} + \left(0.5 + 5 + \mathbb{E}X\right) \times \tfrac{1}{3} + \left(0.5 + 7 + \mathbb{E}X\right) \times \tfrac{1}{3}.$$

2. *Add the extra 0.5 minutes on at the beginning:*

$$\mathbb{E}(X) = 0.5 + \left\{ 3 \times \tfrac{1}{3} + (5 + \mathbb{E}X) \times \tfrac{1}{3} + (7 + \mathbb{E}X) \times \tfrac{1}{3} \right\}.$$

In each case, we will get the same answer (check). This is because the option probabilities sum to 1, *so in Method 1 we are adding in* $0.5 \times (\tfrac{1}{3} + \tfrac{1}{3} + \tfrac{1}{3}) = 0.5 \times 1 = 0.5$, *just as we are in Method 2.*

3.6 Probability as a conditional expectation

Recall from Section 3.1 that for any event A, we can write $\mathbb{P}(A)$ as an expectation as follows.

Define the indicator random variable: $I_A = \begin{cases} 1 & \text{if event } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$

Then $\mathbb{E}(I_A) = \mathbb{P}(I_A = 1) = \mathbb{P}(A)$.

We can refine this expression further, using the idea of conditional expectation. Let Y be any random variable. Then

$$\mathbb{P}(A) = \mathbb{E}(I_A) = \mathbb{E}_Y\Big(\mathbb{E}(I_A \,|\, Y)\Big).$$

But

$$
\begin{aligned}
\mathbb{E}(I_A \,|\, Y) &= \sum_{r=0}^{1} r\mathbb{P}(I_A = r \,|\, Y) \\
&= 0 \times \mathbb{P}(I_A = 0 \,|\, Y) + 1 \times \mathbb{P}(I_A = 1 \,|\, Y) \\
&= \mathbb{P}(I_A = 1 \,|\, Y) \\
&= \mathbb{P}(A \,|\, Y).
\end{aligned}
$$

Thus

$$\mathbb{P}(A) = \mathbb{E}_Y\Big(\mathbb{E}(I_A \,|\, Y)\Big) = \mathbb{E}_Y\Big(\mathbb{P}(A \,|\, Y)\Big).$$

This means that for **any** random variable X (discrete or continuous), and for any set of values S (a discrete set or a continuous set), we can write:

- for any **discrete** random variable Y,

$$\mathbb{P}(X \in S) = \sum_y \mathbb{P}(X \in S \,|\, Y = y)\mathbb{P}(Y = y).$$

- for any **continuous** random variable Y,

$$\mathbb{P}(X \in S) = \int_y \mathbb{P}(X \in S \,|\, Y = y)f_Y(y)\,dy.$$

Example of probability as a conditional expectation: winning a lottery

Suppose that a million people have bought tickets for the weekly lottery draw. Each person has a probability of one-in-a-million of selecting the winning numbers. If more than one person selects the winning numbers, the winner will be chosen at random from all those with matching numbers.

You watch the lottery draw on TV and your numbers match the winning numbers!!! Only a one-in-a-million chance, and there were only a million players, but what is your probability of winning the prize....?

Define Y to be the number of OTHER matching tickets out of the OTHER 1 million tickets sold. (If you are lucky, $Y = 0$ so you are the only person who can win the prize!)

If there are 1 million tickets and each ticket has a one in a million chance of having the winning numbers, then

$$Y \sim \textit{Poisson}(1) \textit{ approximately.}$$

The relationship $Y \sim \text{Poisson}(1)$ arises because of the Poisson approximation to the Binomial distribution.

(a) What is the probability function of Y, $f_Y(y)$?

$$f_Y(y) = \mathbb{P}(Y = y) = \frac{1^y}{y!} e^{-1} = \frac{1}{e \times y!} \qquad \textit{for } y = 0, 1, 2, \ldots.$$

(b) What is the probability that yours is the only matching ticket?

$$\mathbb{P}(\textit{only one matching ticket}) = \mathbb{P}(Y = 0) = \frac{1}{e} = 0.368.$$

(c) The prize is chosen at random from all those who have matching tickets. What is the probability that you win if there are $Y = y$ OTHER matching tickets?

Let W be the event that I win.

$$\mathbb{P}(W \mid Y = y) = \frac{1}{y+1}.$$

(d) Overall, what is the probability that you win, given that you have a matching ticket?

$$\mathbb{P}(W) = \mathbb{E}_Y \left\{ \mathbb{P}(W \mid Y = y) \right\}$$

$$= \sum_{y=0}^{\infty} \mathbb{P}(W \mid Y = y)\mathbb{P}(Y = y)$$

$$= \sum_{y=0}^{\infty} \left(\frac{1}{y+1} \right) \left(\frac{1}{e \times y!} \right)$$

$$= \frac{1}{e} \sum_{y=0}^{\infty} \frac{1}{(y+1)y!}$$

$$= \frac{1}{e} \sum_{y=0}^{\infty} \frac{1}{(y+1)!}$$

$$= \frac{1}{e} \left\{ \sum_{y=0}^{\infty} \frac{1}{y!} - \frac{1}{0!} \right\}$$

$$= \frac{1}{e} \{e - 1\}$$

$$= 1 - \frac{1}{e}$$

$$= 0.632.$$

Disappointing...?

Chapter 4: Generating Functions

This chapter looks at Probability Generating Functions (PGFs) for **_discrete_** random variables. Probability generating functions are useful tools for dealing with sums and limits of random variables.

By the end of this chapter, you should be able to:

- find the sum of Geometric, Binomial, and Exponential series;
- know the definition of the PGF, and use it to calculate the mean, variance, and probabilities;
- calculate the PGF for Geometric, Binomial, and Poisson distributions;
- calculate the PGF for a randomly stopped sum;
- calculate the PGF for first reaching times in the random walk.

4.1 Common sums

1. Geometric Series

$$1 + r + r^2 + r^3 + \ldots \; = \; \sum_{x=0}^{\infty} r^x \; = \; \frac{1}{1-r}, \qquad \textit{when } |r| < 1.$$

This formula proves that $\sum_{x=0}^{\infty} \mathbb{P}(X = x) = 1$ when $X \sim \text{Geometric}(p)$:

$$\mathbb{P}(X = x) = p(1-p)^x \quad \Rightarrow \quad \sum_{x=0}^{\infty} \mathbb{P}(X = x) \; = \; \sum_{x=0}^{\infty} p(1-p)^x$$

$$= \; p \sum_{x=0}^{\infty} (1-p)^x$$

$$= \; \frac{p}{1-(1-p)} \quad \textit{(because } |1-p| < 1\textit{)}$$

$$= \; 1.$$

2. Binomial Theorem

For any $p, q \in \mathbb{R}$, and integer n,

$$(p+q)^n = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}.$$

Note that $\binom{n}{x} = \dfrac{n!}{(n-x)!\, x!}$ *(^nC_r button on calculator.)*

The Binomial Theorem proves that $\sum_{x=0}^{n} \mathbb{P}(X = x) = 1$ when $X \sim \text{Binomial}(n, p)$:

$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ *for* $x = 0, 1, \ldots, n$, *so*

$$
\begin{aligned}
\sum_{x=0}^{n} \mathbb{P}(X = x) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \\
&= \Big(p + (1-p)\Big)^n \\
&= 1^n \\
&= 1.
\end{aligned}
$$

3. Exponential Power Series

$$\text{For any } \lambda \in \mathbb{R}, \qquad \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^\lambda.$$

This proves that $\sum_{x=0}^{\infty} \mathbb{P}(X = x) = 1$ when $X \sim \text{Poisson}(\lambda)$:

$\mathbb{P}(X = x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$ *for* $x = 0, 1, 2, \ldots$, *so*

$$
\begin{aligned}
\sum_{x=0}^{\infty} \mathbb{P}(X = x) = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\
&= e^{-\lambda} e^{\lambda} \\
&= 1.
\end{aligned}
$$

Note: Another useful identity is: $\qquad e^\lambda = \lim_{n \to \infty} \left(1 + \dfrac{\lambda}{n}\right)^n \qquad$ *for* $\lambda \in \mathbb{R}$.

4.2 Probability Generating Functions

The **probability generating function (PGF)** is a useful tool for dealing with **_discrete_** random variables taking values $0, 1, 2, \ldots$. Its particular strength is that it gives us an easy way of characterizing the distribution of $X + Y$ when X and Y are independent. In general it is difficult to find the distribution of a sum using the traditional probability function. The PGF transforms a sum into a product and enables it to be handled much more easily.

Sums of random variables are particularly important in the study of stochastic processes, because many stochastic processes are formed from the sum of a sequence of repeating steps: for example, the Gambler's Ruin from Section 2.6.

The name _probability generating function_ also gives us another clue to the role of the PGF. The PGF can be used to generate all the probabilities of the distribution. This is generally tedious and is not often an efficient way of calculating probabilities. However, the fact that it _can_ be done demonstrates that **the PGF tells us everything there is to know about the distribution.**

Definition: Let X be a discrete random variable taking values in the non-negative integers $\{0, 1, 2, \ldots\}$. The **probability generating function (PGF)** of X is $G_X(s) = \mathbb{E}(s^X)$, _for all $s \in \mathbb{R}$ for which the sum converges._

Calculating the probability generating function

$$G_X(s) = \mathbb{E}\left(s^X\right) = \sum_{x=0}^{\infty} s^x \mathbb{P}(X = x).$$

Properties of the PGF:

1. $G_X(0) = \mathbb{P}(X = 0)$:

$$
\begin{aligned}
G_X(0) &= 0^0 \times \mathbb{P}(X = 0) + 0^1 \times \mathbb{P}(X = 1) + 0^2 \times \mathbb{P}(X = 2) + \ldots \\
\therefore \quad G_X(0) &= \mathbb{P}(X = 0).
\end{aligned}
$$

2. $G_X(1) = 1$: $\qquad G_X(1) = \displaystyle\sum_{x=0}^{\infty} 1^x \mathbb{P}(X = x) = \sum_{x=0}^{\infty} \mathbb{P}(X = x) = 1.$

Example 1: <u>**Binomial Distribution**</u>

Let $X \sim \text{Binomial}(n, p)$, so $\mathbb{P}(X = x) = \dbinom{n}{x} p^x q^{n-x}$ for $x = 0, 1, \ldots, n$.

$$
\begin{aligned}
G_X(s) &= \sum_{x=0}^{n} s^x \binom{n}{x} p^x q^{n-x} \\
&= \sum_{x=0}^{n} \binom{n}{x} (ps)^x q^{n-x} \\
&= (ps + q)^n \quad \textit{by the Binomial Theorem: true for all } s.
\end{aligned}
$$

Thus $G_X(s) = (ps + q)^n$ for all $s \in \mathbb{R}$.

<u>**Check $G_X(0)$:**</u>
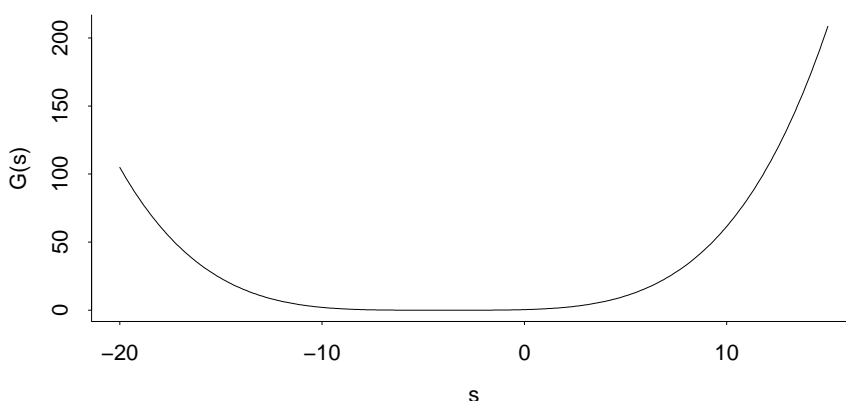
$$
\begin{aligned}
G_X(0) &= (p \times 0 + q)^n \\
&= q^n \\
&= \mathbb{P}(X = 0).
\end{aligned}
$$

X ~ Bin(n=4, p=0.2)

<u>**Check $G_X(1)$:**</u>

$$
\begin{aligned}
G_X(1) &= (p \times 1 + q)^n \\
&= (1)^n \\
&= 1.
\end{aligned}
$$

Example 2: Poisson Distribution

Let $X \sim \text{Poisson}(\lambda)$, so $\mathbb{P}(X = x) = \dfrac{\lambda^x}{x!}e^{-\lambda}$ for $x = 0, 1, 2, \ldots$.

$$G_X(s) = \sum_{x=0}^{\infty} s^x \frac{\lambda^x}{x!} e^{-\lambda} \;\; = \;\; e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda s)^x}{x!}$$

$$= \;\; e^{-\lambda} e^{(\lambda s)} \quad \textit{for all } s \in \mathbb{R}.$$

Thus $\quad G_X(s) = e^{\lambda(s-1)} \quad$ *for all* $s \in \mathbb{R}$.

X ~ Poisson(4)
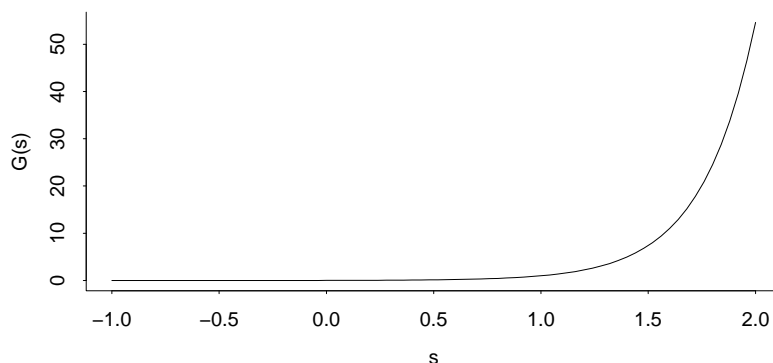
Example 3: Geometric Distribution

Let $X \sim \text{Geometric}(p)$, so $\mathbb{P}(X = x) = p(1-p)^x = pq^x$ for $x = 0, 1, 2, \ldots$, where $q = 1 - p$.

$$G_X(s) \;\; = \;\; \sum_{x=0}^{\infty} s^x p q^x$$

$$= \;\; p \sum_{x=0}^{\infty} (qs)^x$$

X ~ Geom(0.8)

$$= \;\; \frac{p}{1 - qs} \quad \textit{for all } s \textit{ such that } |qs| < 1.$$

Thus $\quad G_X(s) = \dfrac{p}{1 - qs} \quad$ *for* $\;\; |s| < \dfrac{1}{q}$.

4.3 Using the probability generating function to calculate probabilities

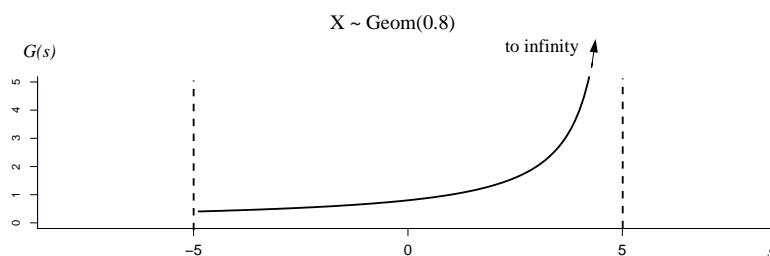The probability generating function gets its name because the power series can be expanded and differentiated to reveal the individual probabilities. Thus, *given only the PGF $G_X(s) = \mathbb{E}(s^X)$, we can recover all probabilities $\mathbb{P}(X = x)$.*

For shorthand, write $p_x = \mathbb{P}(X = x)$. Then

$$G_X(s) = \mathbb{E}(s^X) \;=\; \sum_{x=0}^{\infty} p_x s^x = p_0 + p_1 s + p_2 s^2 + p_3 s^3 + p_4 s^4 + \ldots$$

Thus $p_0 = \mathbb{P}(X = 0) = G_X(0)$.

First derivative: $\quad G'_X(s) \;=\; p_1 + 2p_2 s + 3p_3 s^2 + 4p_4 s^3 + \ldots$

Thus $p_1 = \mathbb{P}(X = 1) = G'_X(0)$.

Second derivative: $\quad G''_X(s) \;=\; 2p_2 + (3 \times 2)p_3 s + (4 \times 3)p_4 s^2 + \ldots$

Thus $p_2 = \mathbb{P}(X = 2) = \dfrac{1}{2} G''_X(0)$.

Third derivative: $\quad G'''_X(s) \;=\; (3 \times 2 \times 1)p_3 + (4 \times 3 \times 2)p_4 s + \ldots$

Thus $p_3 = \mathbb{P}(X = 3) = \dfrac{1}{3!} G'''_X(0)$.

In general:

$$\boxed{\; p_n = \mathbb{P}(X = n) = \left(\frac{1}{n!}\right) G_X^{(n)}(0) = \left(\frac{1}{n!}\right) \frac{d^n}{ds^n}\left(G_X(s)\right)\Big|_{s=0}. \;}$$

Example: Let X be a discrete random variable with PGF $G_X(s) = \dfrac{s}{5}(2 + 3s^2)$. Find the distribution of X.

$$G_X(s) = \frac{2}{5}s + \frac{3}{5}s^3 : \qquad G_X(0) = \mathbb{P}(X = 0) = 0.$$

$$G_X'(s) = \frac{2}{5} + \frac{9}{5}s^2 : \qquad G_X'(0) = \mathbb{P}(X = 1) = \frac{2}{5}.$$

$$G_X''(s) = \frac{18}{5}s : \qquad \frac{1}{2}G_X''(0) = \mathbb{P}(X = 2) = 0.$$

$$G_X'''(s) = \frac{18}{5} : \qquad \frac{1}{3!}G_X'''(0) = \mathbb{P}(X = 3) = \frac{3}{5}.$$

$$G_X^{(r)}(s) = 0 \; \forall r \geq 4 : \qquad \frac{1}{r!}G_X^{(r)}(s) = \mathbb{P}(X = r) = 0 \; \forall r \geq 4.$$

Thus

$$X = \begin{cases} 1 \text{ \emph{with probability} } 2/5, \\ 3 \text{ \emph{with probability} } 3/5. \end{cases}$$

Uniqueness of the PGF

The formula $p_n = \mathbb{P}(X = n) = \left(\dfrac{1}{n!}\right) G_X^{(n)}(0)$ shows that the whole sequence of probabilities $p_0, p_1, p_2, \ldots$ is determined by the values of the PGF and its derivatives at $s = 0$. It follows that the PGF specifies a ***unique*** set of probabilities.

Fact: If two power series agree on any interval containing 0, however small, then all terms of the two series are equal.

Formally: let $A(s)$ and $B(s)$ be PGFs with $A(s) = \sum_{n=0}^{\infty} a_n s^n$, $B(s) = \sum_{n=0}^{\infty} b_n s^n$. If there exists some $R' > 0$ such that $A(s) = B(s)$ for all $-R' < s < R'$, then $a_n = b_n$ for all n.

Practical use: If we can show that two random variables have the same PGF in some interval containing 0, then we have shown that *the two random variables have the same distribution.*

Another way of expressing this is to say that *the PGF of X tells us everything there is to know about the distribution of X.*

4.4 Expectation and moments from the PGF

As well as calculating probabilities, we can also use the PGF to calculate the moments of the distribution of X. The moments of a distribution are *the mean, variance, etc.*

Theorem 4.4: Let X be a discrete random variable with PGF $G_X(s)$. Then:

1. $\mathbb{E}(X) = G'_X(1)$.

2. $\mathbb{E}\Big\{X(X-1)(X-2)\ldots(X-k+1)\Big\} = G_X^{(k)}(1) = \dfrac{d^k\, G_X(s)}{ds^k}\bigg|_{s=1}.$
   (This is the <u>kth factorial moment</u> of X.)

Proof: (Sketch: see Section 4.8 for more details)

1.
$$G_X(s) \;=\; \sum_{x=0}^{\infty} s^x\, p_x,$$

so
$$G'_X(s) \;=\; \sum_{x=0}^{\infty} x s^{x-1} p_x$$

$$\Rightarrow \qquad G'_X(1) \;=\; \sum_{x=0}^{\infty} x p_x = \mathbb{E}(X)$$
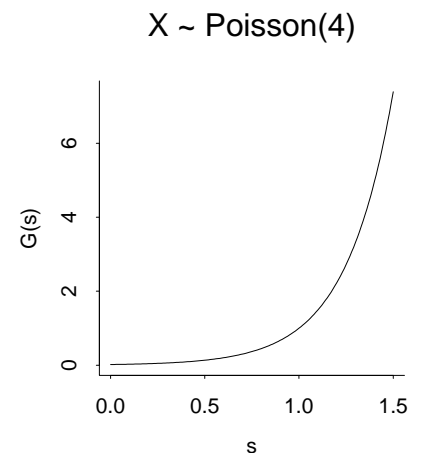
X ~ Poisson(4)

2.
$$G_X^{(k)}(s) = \frac{d^k\, G_X(s)}{ds^k} \;=\; \sum_{x=k}^{\infty} x(x-1)(x-2)\ldots(x-k+1)s^{x-k}p_x$$

so
$$G_X^{(k)}(1) \;=\; \sum_{x=k}^{\infty} x(x-1)(x-2)\ldots(x-k+1)p_x$$

$$\;=\; \mathbb{E}\Big\{X(X-1)(X-2)\ldots(X-k+1)\Big\}. \qquad \square$$

Example: Let $X \sim \text{Poisson}(\lambda)$. The PGF of X is $G_X(s) = e^{\lambda(s-1)}$. Find $\mathbb{E}(X)$ and $\text{Var}(X)$.

X ~ Poisson(4)

Solution:

$$G'_X(s) \;=\; \lambda e^{\lambda(s-1)}$$

$$\Rightarrow \;\; \mathbb{E}(X) \;=\; G'_X(1) = \lambda.$$

For the variance, consider

$$\mathbb{E}\Big\{X(X-1)\Big\} \;=\; G''_X(1) = \lambda^2 e^{\lambda(s-1)}|_{s=1} = \lambda^2.$$

So

$$
\begin{aligned}
\text{Var}(X) \;&=\; \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\
&=\; \mathbb{E}\Big\{X(X-1)\Big\} + \mathbb{E}X - (\mathbb{E}X)^2 \\
&=\; \lambda^2 + \lambda - \lambda^2 \\
&=\; \lambda.
\end{aligned}
$$
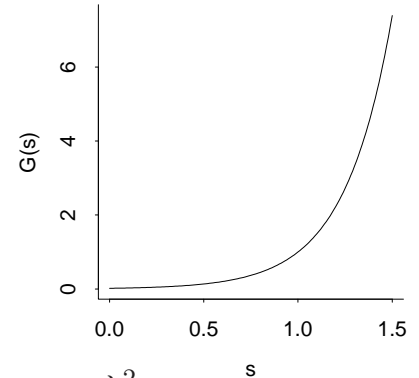
4.5 Probability generating function for a sum of independent r.v.s

One of the PGF's greatest strengths is that it turns a sum into a product:

$$\mathbb{E}\left(s^{(X_1+X_2)}\right) = \mathbb{E}\left(s^{X_1}s^{X_2}\right).$$

This makes the PGF useful for finding the probabilities and moments of *a sum of independent random variables.*

__Theorem 4.5:__ Suppose that $X_1, \ldots, X_n$ are ***independent*** random variables, and let $Y = X_1 + \ldots + X_n$. Then

$$G_Y(s) = \prod_{i=1}^{n} G_{X_i}(s).$$

Proof:

$$G_Y(s) = \mathbb{E}(s^{(X_1+\ldots+X_n)})$$

$$= \mathbb{E}(s^{X_1}s^{X_2}\ldots s^{X_n})$$

$$= \mathbb{E}(s^{X_1})\mathbb{E}(s^{X_2})\ldots\mathbb{E}(s^{X_n})$$

(because $X_1, \ldots, X_n$ are independent)

$$= \prod_{i=1}^{n} G_{X_i}(s). \quad \text{as required.} \quad \square$$

Example: Suppose that X and Y are independent with $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$. Find the distribution of $X + Y$.

Solution:

$$G_{X+Y}(s) = G_X(s) \cdot G_Y(s)$$

$$= e^{\lambda(s-1)}e^{\mu(s-1)}$$

$$= e^{(\lambda+\mu)(s-1)}.$$

But this is the PGF of the Poisson($\lambda + \mu$) distribution. So, by the uniqueness of PGFs, $X + Y \sim$ Poisson($\lambda + \mu$).

4.6 Randomly stopped sum

Remember the randomly stopped sum model from Section 3.4. A random number N of events occur, and each event i has associated with it a cost or reward X_i. The question is to find the distribution of the total cost or reward: $T_N = X_1 + X_2 + \ldots + X_N$. T_N is called a *randomly stopped sum* because it has a random number of terms.

Example: Cash machine model. N customers arrive during the day. Customer i withdraws amount X_i. The total amount withdrawn during the day is $T_N = X_1 + \ldots + X_N$.

In Chapter 3, we used the Laws of Total Expectation and Variance to show that $\mathbb{E}(T_N) = \mu \, \mathbb{E}(N)$ and $\mathrm{Var}(T_N) = \sigma^2 \, \mathbb{E}(N) + \mu^2 \, \mathrm{Var}(N)$, where $\mu = \mathbb{E}(X_i)$ and $\sigma^2 = \mathrm{Var}(X_i)$.

In this chapter we will now use probability generating functions to investigate the *whole distribution of T_N*.

Theorem 4.6: Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with common PGF G_X. Let N be a random variable, independent of the X_i's, with PGF G_N, and let $T_N = X_1 + \ldots + X_N = \sum_{i=1}^{N} X_i$. Then the PGF of T_N is:

$$G_{T_N}(s) = G_N\Big(G_X(s)\Big).$$

Proof:

$$
\begin{aligned}
G_{T_N}(s) \;&=\; \mathbb{E}(s^{T_N}) = \mathbb{E}\left(s^{X_1+\ldots+X_N}\right) \\[2mm]
&=\; \mathbb{E}_N\left\{ \mathbb{E}\left(s^{X_1+\ldots+X_N} \,\Big|\, N\right) \right\} \quad \text{\textit{(conditional expectation)}} \\[2mm]
&=\; \mathbb{E}_N\left\{ \mathbb{E}\left(s^{X_1}\ldots s^{X_N} \,|\, N\right) \right\} \\[2mm]
&=\; \mathbb{E}_N\left\{ \mathbb{E}\left(s^{X_1}\ldots s^{X_N}\right) \right\} \quad \text{\textit{(X_i's are indept of N)}} \\[2mm]
&=\; \mathbb{E}_N\left\{ \mathbb{E}\left(s^{X_1}\right)\ldots \mathbb{E}\left(s^{X_N}\right) \right\} \quad \text{\textit{(X_i's are indept of each other)}} \\[2mm]
&=\; \mathbb{E}_N\left\{ \left(G_X(s)\right)^N \right\} \\[2mm]
&=\; G_N\Big(G_X(s)\Big) \quad \text{\textit{(by definition of G_N).}} \qquad \square
\end{aligned}
$$

Example: Let $X_1, X_2, \ldots$ and N be as above. Find the mean of T_N.

$$
\begin{aligned}
\mathbb{E}(T_N) = G'_{T_N}(1) &= \left. \frac{d}{ds} G_N(G_X(s)) \right|_{s=1} \\[2mm]
&= \left. G'_N(G_X(s)) \cdot G'_X(s) \right|_{s=1} \\[2mm]
&= G'_N(1) \cdot G'_X(1) \qquad \textit{Note: } G_X(1) = 1 \textit{ for any r.v. } X \\[2mm]
&= \mathbb{E}(N) \cdot \mathbb{E}(X_1), \qquad \textit{— same answer as in Chapter 3.}
\end{aligned}
$$

Example: <u>Heron goes fishing</u>

My aunt was asked by her neighbours to feed the prize
goldfish in their garden pond while they were on holiday.
Although my aunt dutifully went and fed them every day,
she never saw a single fish for the whole three weeks. It
turned out that all the fish had been eaten by a heron
when she wasn't looking!

Let N be the number of times the heron visits the pond
during the neighbours' absence. Suppose that $N \sim \text{Geometric}(1 - \theta)$,
so $\mathbb{P}(N = n) = (1 - \theta)\theta^n$, for $n = 0, 1, 2, \ldots$. When the heron visits the pond
it has probability p of catching a prize goldfish, independently of what happens
on any other visit. (This assumes that there are infinitely many goldfish to be
caught!) Find the distribution of

$$T = \text{total number of goldfish caught.}$$

Solution:

$$
\text{Let} \quad X_i = \begin{cases} 1 & \textit{if heron catches a fish on visit } i, \\ 0 & \textit{otherwise.} \end{cases}
$$

Then $T = X_1 + X_2 + \ldots + X_N$ *(randomly stopped sum), so*

$$G_T(s) = G_N(G_X(s)).$$

Now

$$G_X(s) = \mathbb{E}(s^X) = s^0 \times \mathbb{P}(X = 0) + s^1 \times \mathbb{P}(X = 1) = 1 - p + ps.$$

Also,

$$
\begin{aligned}
G_N(r) = \sum_{n=0}^{\infty} r^n \mathbb{P}(N = n) &= \sum_{n=0}^{\infty} r^n (1 - \theta)\theta^n \\
&= (1 - \theta) \sum_{n=0}^{\infty} (\theta r)^n \\
&= \frac{1 - \theta}{1 - \theta r}. \qquad (r < 1/\theta).
\end{aligned}
$$

So

$$G_T(s) = \frac{1 - \theta}{1 - \theta\, G_X(s)} \qquad \text{(\textit{putting } } r = G_X(s)),$$

giving:

$$
\begin{aligned}
G_T(s) &= \frac{1 - \theta}{1 - \theta(1 - p + ps)} \\[2mm]
&= \frac{1 - \theta}{1 - \theta + \theta p - \theta ps}
\end{aligned}
$$

[could this be Geometric? $G_T(s) = \dfrac{1 - \pi}{1 - \pi s}$ *for some* π*?]*

$$
\begin{aligned}
&= \frac{1 - \theta}{(1 - \theta + \theta p) - \theta ps} \\[3mm]
&= \frac{\left(\dfrac{1 - \theta}{1 - \theta + \theta p}\right)}{\left(\dfrac{(1 - \theta + \theta p) - \theta ps}{1 - \theta + \theta p}\right)}
\end{aligned}
$$

$$= \frac{\left(\dfrac{1 - \theta + \theta p - \theta p}{1 - \theta + \theta p}\right)}{1 - \left(\dfrac{\theta p}{1 - \theta + \theta p}\right) s}$$

$$= \frac{1 - \left(\dfrac{\theta p}{1 - \theta + \theta p}\right)}{1 - \left(\dfrac{\theta p}{1 - \theta + \theta p}\right) s}.$$

This is the PGF of the Geometric $\left(1 - \dfrac{\theta p}{1 - \theta + \theta p}\right)$ *distribution, so by uniqueness of PGFs, we have:*

$$T \sim \text{Geometric} \left(\frac{1 - \theta}{1 - \theta + \theta p}\right).$$

Why did we need to use the PGF?

We could have solved the heron problem without using the PGF, but it is much more difficult. PGFs are very useful for dealing with sums of random variables, which are difficult to tackle using the standard probability function.

Here are the first few steps of solving the heron problem without the PGF. Recall the problem:

- Let $N \sim \text{Geometric}(1 - \theta)$, so $\mathbb{P}(N = n) = (1 - \theta)\theta^n$;

- Let $X_1, X_2, \ldots$ be independent of each other and of N, with $X_i \sim \text{Binomial}(1, p)$ (remember $X_i = 1$ with probability p, and 0 otherwise);

- Let $T = X_1 + \ldots + X_N$ be the randomly stopped sum;

- Find the distribution of T.

Without using the PGF, we would tackle this by looking for an expression for $\mathbb{P}(T = t)$ for any t. Once we have obtained that expression, we might be able to see that T has a distribution we recognise (e.g. Geometric), or otherwise we would just state that T is defined by the probability function we have obtained.

To find $\mathbb{P}(T = t)$, we have to *partition over different values of N:*

$$\mathbb{P}(T = t) = \sum_{n=0}^{\infty} \mathbb{P}(T = t \mid N = n)\mathbb{P}(N = n). \qquad (\star)$$

Here, we are *lucky* that we can write down the distribution of $T \mid N = n$:

- if $N = n$ is fixed, then $T = X_1 + \ldots + X_n$ is a sum of n independent Binomial$(1, p)$ random variables, so

$$(T \mid N = n) \sim Binomial(n, p).$$

For most distributions of X, *it would be difficult or impossible to write down the distribution of $X_1 + \ldots + X_n$:*

we would have to use an expression like

$$\mathbb{P}(X_1 + \ldots + X_N = t \mid N = n) = \sum_{x_1=0}^{t} \sum_{x_2=0}^{t-x_1} \ldots \sum_{x_{n-1}=0}^{t-(x_1+\ldots+x_{n-2})} \Big\{ \mathbb{P}(X_1 = x_1) \times$$

$$\mathbb{P}(X_2 = x_2) \times \ldots \times \mathbb{P}(X_{n-1} = x_{n-1}) \times \mathbb{P}[X_n = t - (x_1 + \ldots + x_{n-1})] \Big\}.$$

Back to the heron problem: we are lucky in this case that we know the distribution of $(T \mid N = n)$ is Binomial$(N = n, p)$, so

$$\mathbb{P}(T = t \mid N = n) = \binom{n}{t} p^t (1 - p)^{n-t} \quad \text{for } t = 0, 1, \ldots, n.$$

Continuing from $(\star)$:

$$\mathbb{P}(T = t) = \sum_{n=0}^{\infty} \mathbb{P}(T = t \mid N = n)\mathbb{P}(N = n)$$

$$= \sum_{n=t}^{\infty} \binom{n}{t} p^t (1-p)^{n-t} (1-\theta)\theta^n$$

$$= (1-\theta) \left(\frac{p}{1-p}\right)^t \sum_{n=t}^{\infty} \binom{n}{t} \left[\theta(1-p)\right]^n \qquad (\star\star)$$

$$= \ldots?$$

As it happens, we can evaluate the sum in $(\star\star)$ using the fact that Negative Binomial probabilities sum to 1. You can try this if you like, but it is quite tricky. [Hint: use the Negative Binomial $(t+1, 1-\theta(1-p))$ distribution.]

Overall, we obtain the same answer that $T \sim \text{Geometric}\left(\dfrac{1-\theta}{1-\theta+\theta p}\right)$, but hopefully you can see why the PGF is so useful.

Without the PGF, we have two major difficulties:

1. *Writing down $\mathbb{P}(T = t \mid N = n)$;*

2. *Evaluating the sum over n in $(\star\star)$.*

For a general problem, both of these steps might be too difficult to do without a computer. The PGF has none of these difficulties, and even if $G_T(s)$ does not simplify readily, it still tells us everything there is to know about the distribution of T.

4.7 Summary: Properties of the PGF

Definition: $G_X(s) = \mathbb{E}(s^X)$

Used for: Discrete r.v.s with values 0, 1, 2, . . .

Moments: $\mathbb{E}(X) = G'_X(1)$ $\mathbb{E}\left\{X(X-1)\ldots(X-k+1)\right\} = G_X^{(k)}(1)$

Probabilities: $\mathbb{P}(X = n) = \dfrac{1}{n!}G_X^{(n)}(0)$

Sums: $G_{X+Y}(s) = G_X(s)G_Y(s)$ for independent X, Y

4.8   Convergence of PGFs

We have been using PGFs throughout this chapter without paying much attention to their mathematical properties. For example, are we sure that the power series $G_X(s) = \sum_{x=0}^{\infty} s^x \mathbb{P}(X = x)$ converges? Can we differentiate and integrate the infinite power series term by term as we did in Section 4.4? When we said in Section 4.4 that $\mathbb{E}(X) = G'_X(1)$, can we be sure that $G_X(1)$ and its derivative $G'_X(1)$ even exist?

This technical section introduces the important notion of the **radius of convergence** of the PGF. Although it isn't obvious, it is always safe to assume convergence of $G_X(s)$ at least for $|s| < 1$. Also, there are results that assure us that $\mathbb{E}(X) = G'_X(1)$ will work for all non-defective random variables X.

Definition: The **radius of convergence** of a probability generating function is *a number $R > 0$, such that the sum $G_X(s) = \sum_{x=0}^{\infty} s^x \mathbb{P}(X = x)$ converges if $|s| < R$ and diverges ($\to \infty$) if $|s| > R$.*

(No general statement is made about what happens when $|s| = R$.)

Fact: For any PGF, the radius of convergence exists.

It is always ≥ 1: every PGF converges for at least $s \in (-1, 1)$.

The radius of convergence could be anything from $R = 1$ to $R = \infty$.

Note: This gives us the surprising result that the set of s for which the PGF $G_X(s)$ converges is symmetric about 0: the PGF converges for all $s \in (-R, R)$, and for no $s < -R$ or $s > R$.

This is surprising because the PGF itself is not usually symmetric about 0: i.e. $G_X(-s) \neq G_X(s)$ in general.

Example 1: **Geometric distribution**

Let $X \sim$ Geometric($p = 0.8$). What is the radius of convergence of $G_X(s)$?

As in Section 4.2,

$$G_X(s) = \sum_{x=0}^{\infty} s^x (0.8)(0.2)^x \;=\; 0.8 \sum_{x=0}^{\infty} (0.2s)^x$$

$$= \frac{0.8}{1 - 0.2s} \quad \text{for all } s \text{ such that } |0.2s| < 1.$$

This is valid for all s with $|0.2s| < 1$, so it is valid for all s with $|s| < \frac{1}{0.2} = 5$.
(i.e. $-5 < s < 5$.)
The radius of convergence is $R = 5$.

The figure shows the PGF of the Geometric($p = 0.8$) distribution, with its radius of convergence $R = 5$. Note that although the convergence set $(-5, 5)$ is symmetric about 0, the function $G_X(s) = p/(1 - qs) = 4/(5 - s)$ is not.

Geometric(0.8) probability generating function
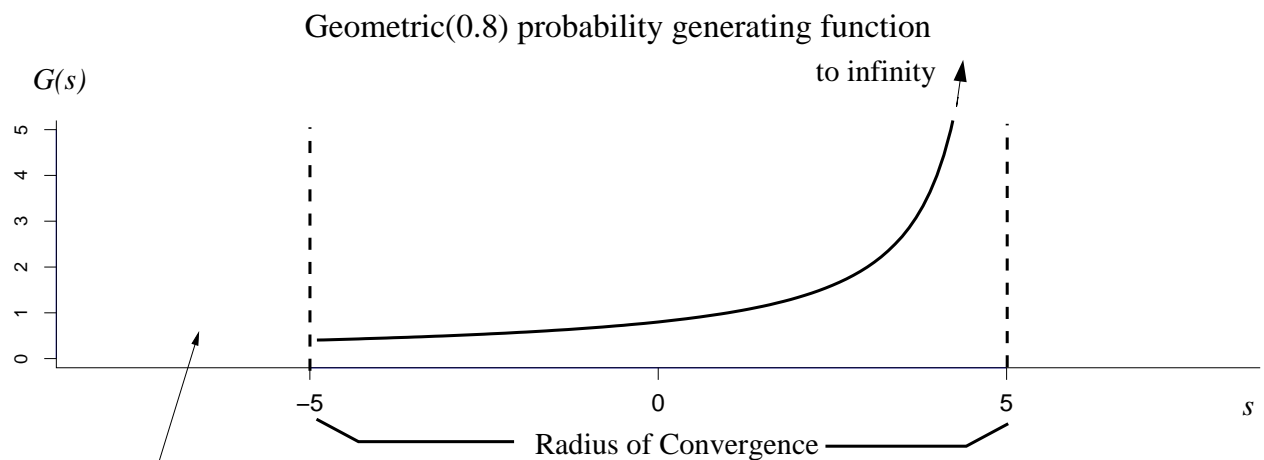
In this region, p/(1−qs) remains finite and well−behaved, but it is no longer equal to E(s^x).

At the limits of convergence, strange things happen:

- At the positive end, as $s \uparrow 5$, both $G_X(s)$ and $p/(1 - qs)$ approach infinity. So the PGF *is* (left)-continuous at $+R$:

$$\lim_{s \uparrow 5} G_X(s) = G_X(5) = \infty.$$

  However, the PGF does *not* converge at $s = +R$.

- At the negative end, as $s \downarrow -5$, the function $p/(1 - qs) = 4/(5 - s)$ is continuous and passes through 0.4 when $s = -5$. However, when $s \leq -5$, this function no longer represents $G_X(s) = 0.8 \sum_{x=0}^{\infty} (0.2s)^x$, because $|0.2s| \geq 1$.

  Additionally, when $s = -5$, $G_X(-5) = 0.8 \sum_{x=0}^{\infty} (-1)^x$ does not exist. Unlike the positive end, this means that $G_X(s)$ is *not* (right)-continuous at $-R$:

  $$\lim_{s \downarrow -5} G_X(s) = 0.4 \neq G_X(-5).$$

  Like the positive end, this PGF does *not* converge at $s = -R$.

Example 2: **Binomial distribution**

Let $X \sim \text{Binomial}(n, p)$. What is the radius of convergence of $G_X(s)$?

As in Section 4.2,

$$\begin{aligned}
G_X(s) &= \sum_{x=0}^{n} s^x \binom{n}{x} p^x q^{n-x} \\
&= \sum_{x=0}^{n} \binom{n}{x} (ps)^x q^{n-x} \\
&= (ps + q)^n \quad \textit{by the Binomial Theorem: true for all } s.
\end{aligned}$$

This is true for all $-\infty < s < \infty$, so the radius of convergence is $R = \infty$.

Abel's Theorem for continuity of power series at $s = 1$

Recall from above that if $X \sim \text{Geometric}(0.8)$, then $G_X(s)$ is not continuous at the negative end of its convergence $(-R)$:

$$\lim_{s \downarrow -5} G_X(s) \neq G_X(-5).$$

Abel's theorem states that this sort of effect can never happen at $s = 1$ (or at $+R$). In particular, $G_X(s)$ is always left-continuous at $s = 1$:

$$\lim_{s \uparrow 1} G_X(s) = G_X(1) \quad \text{always, even if } G_X(1) = \infty.$$

Theorem 4.8: Abel's Theorem.

$$\text{Let} \quad G(s) = \sum_{i=0}^{\infty} p_i s^i \quad \text{for any } p_0, p_1, p_2, \dots \text{ with } p_i \geq 0 \text{ for all } i.$$

Then $G(s)$ is left-continuous at $s = 1$:

$$\lim_{s \uparrow 1} G(s) = \sum_{i=0}^{\infty} p_i \;=\; G(1),$$

whether or not this sum is finite.

Note: Remember that the radius of convergence $R \geq 1$ for any PGF, so Abel's Theorem means that even in the worst-case scenario when $R = 1$, we can still trust that the PGF will be continuous at $s = 1$. (By contrast, we can not be sure that the PGF will be continuous at the the lower limit $-R$).

Abel's Theorem means that for any PGF, we can write $G_X(1)$ as shorthand for $\lim_{s \uparrow 1} G_X(s)$.

It also clarifies our proof that $\mathbb{E}(X) = G'_X(1)$ from Section 4.4. If we assume that term-by-term differentiation is allowed for $G_X(s)$ (see below), then the proof on page 75 gives:

$$G_X(s) \;=\; \sum_{x=0}^{\infty} s^x \, p_x,$$

$$\text{so} \quad G'_X(s) \;=\; \sum_{x=1}^{\infty} x s^{x-1} p_x \qquad \text{(term-by-term differentiation: see below)}.$$

Abel's Theorem establishes that $\mathbb{E}(X)$ is equal to $\lim_{s \uparrow 1} G'_X(s)$:

$$\mathbb{E}(X) \;=\; \sum_{x=1}^{\infty} x p_x$$

$$= \; G'_X(1)$$

$$= \; \lim_{s \uparrow 1} G'_X(s),$$

because Abel's Theorem applies to $G'_X(s) = \sum_{x=1}^{\infty} x s^{x-1} p_x$, establishing that $G'_X(s)$ is left-continuous at $s = 1$. Without Abel's Theorem, we could not be sure that the limit of $G'_X(s)$ as $s \uparrow 1$ would give us the correct answer for $\mathbb{E}(X)$.

Absolute and uniform convergence for term-by-term differentiation

We have stated that the PGF converges for all $|s| < R$ for some R. In fact, the probability generating function converges *absolutely* if $|s| < R$. Absolute convergence is stronger than convergence alone: it means that the sum of absolute values, $\sum_{x=0}^{\infty} |s^x \mathbb{P}(X = x)|$, also converges. When two series both converge absolutely, the product series also converges absolutely. This guarantees that $G_X(s) \times G_Y(s)$ is absolutely convergent for any two random variables X and Y. This is useful because $G_X(s) \times G_Y(s) = G_{X+Y}(s)$ if X and Y are independent.

The PGF also converges *uniformly* on any set $\{s : |s| \leq R'\}$ where $R' < R$. Intuitively, this means that the speed of convergence does not depend upon the value of s. Thus a value n_0 can be found such that for all values of $n \geq n_0$, the *finite* sum $\sum_{x=0}^{n} s^x \mathbb{P}(X = x)$ is *simultaneously* close to the converged value $G_X(s)$, for all s with $|s| \leq R'$. In mathematical notation: $\forall \epsilon > 0$, $\exists n_0 \in \mathbb{Z}$ such that $\forall s$ with $|s| \leq R'$, and $\forall n \geq n_0$,

$$\left| \sum_{x=0}^{n} s^x \mathbb{P}(X = x) - G_X(s) \right| < \epsilon.$$

Uniform convergence allows us to differentiate or integrate the PGF term by term.

Fact: Let $G_X(s) = \mathbb{E}(s^X) = \sum_{x=0}^{\infty} s^x \mathbb{P}(X = x)$, and let $s < R$.

1. $G_X'(s) = \dfrac{d}{ds} \left(\sum_{x=0}^{\infty} s^x \mathbb{P}(X = x) \right) = \sum_{x=0}^{\infty} \dfrac{d}{ds} \left(s^x \mathbb{P}(X = x) \right) = \sum_{x=0}^{\infty} x s^{x-1} \mathbb{P}(X = x).$
   (term by term differentiation).

2. $\displaystyle\int_a^b G_X(s)\, ds = \int_a^b \left(\sum_{x=0}^{\infty} s^x \mathbb{P}(X = x) \right) ds = \sum_{x=0}^{\infty} \left(\int_a^b s^x \mathbb{P}(X = x)\, ds \right)$

   $\displaystyle = \sum_{x=0}^{\infty} \frac{s^{x+1}}{x+1} \mathbb{P}(X = x) \ \ \text{for} \ \ -R < a < b < R.$
   (term by term integration).

4.9   Using PGFs for First Reaching Times in the Random Walk

Remember the Drunkard's Walk from Chapter 1: a drunk person staggers to left and right as he walks. This process is also called the **Random Walk** in stochastic processes. Probability generating functions are particularly useful for processes such as the random walk, because the process is defined as the sum of a single repeating step. The repeating step is a move of one unit, left or right at random. The sum of the first t steps gives the position at time t.

Our interest is to find the distribution of T_{ij}, the *number of steps taken to reach state j, starting at state i.*

T_{ij} is called the *first reaching time from state i to state j.*

We will demonstrate the procedure with the simplest case of a symmetric random walk on the integers.

Let
$$Y_i = \begin{cases} 1 & \text{with probability } 0.5, \\ -1 & \text{with probability } 0.5, \end{cases}$$
and suppose $Y_1, Y_2, \ldots$ are independent.

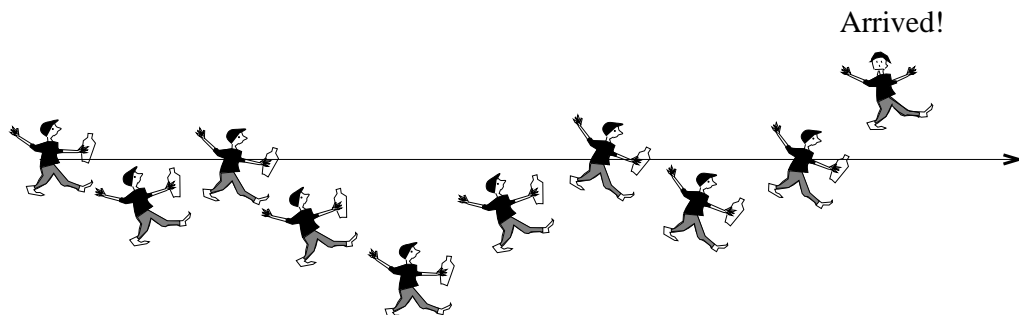Y_i represents the step to left or right taken by the drunkard at time i.

Let $X_0 = 0$, and $X_t = \sum_{i=1}^{t} Y_i$ for $t > 0$, and consider the stochastic process $\{X_0, X_1, X_2, \ldots\}$.

X_t represents the drunkard's *position at time t.*

Define $T_{01} = $ *number of steps (i.e. time taken) to get from state 0 to state 1.*

What is the PGF of T_{01}, $\mathbb{E}\left(s^{T_{01}}\right)$?

Arrived!

Solution:

Define T_{ij} = *number of steps to get from state i to state j for any i, j.*

Let $H(s) = \mathbb{E}\left(s^{T_{01}}\right)$ *be the PGF required.*

Then

$$
\begin{aligned}
H(s) &= \mathbb{E}\left(s^{T_{01}}\right) \\
&= \mathbb{E}\left(s^{T_{01}} \mid Y_1 = 1\right) \mathbb{P}(Y_1 = 1) + \mathbb{E}\left(s^{T_{01}} \mid Y_1 = -1\right) \mathbb{P}(Y_1 = -1) \\
&= \frac{1}{2}\left\{ \mathbb{E}\left(s^{T_{01}} \mid Y_1 = 1\right) + \mathbb{E}\left(s^{T_{01}} \mid Y_1 = -1\right) \right\}. \quad \spadesuit
\end{aligned}
$$

Now if $Y_1 = 1$, then $T_{01} = 1$ definitely, so $\mathbb{E}\left(s^{T_{01}} \mid Y_1 = 1\right) = s^1 = s$.

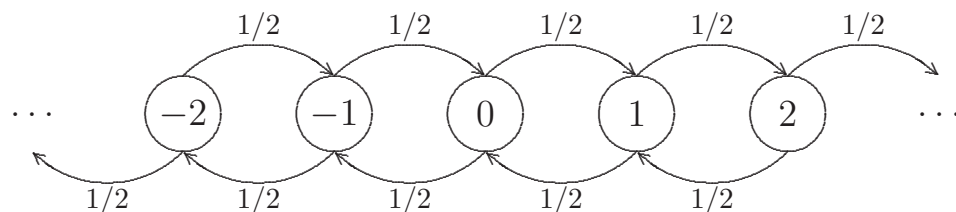*If $Y_1 = -1$, then $T_{01} = 1 + T_{-1,1}$:*

→ *one step from state 0 to state -1,*

→ *then $T_{-1,1}$ steps from state -1 to state 1.*

But $T_{-1,1} = T_{-1,0} + T_{01}$, because the process must pass through 0 to get from -1 to 1.

Now $T_{-1,0}$ and T_{01} are independent (Markov property: see later). Also, they have the same distribution because the process is translation invariant:

Thus

$$
\begin{aligned}
\mathbb{E}\left(s^{T_{01}} \mid Y_1 = -1\right) &= \mathbb{E}\left(s^{1+T_{-1,1}}\right) \\
&= \mathbb{E}\left(s^{1+T_{-1,0}+T_{0,1}}\right) \\
&= s\,\mathbb{E}\left(s^{T_{-1,0}}\right)\mathbb{E}\left(s^{T_{01}}\right) \quad \text{by independence} \\
&= s(H(s))^2 \quad \text{because identically distributed.}
\end{aligned}
$$

Thus

$$
H(s) = \frac{1}{2}\left\{s + s(H(s))^2\right\} \qquad \text{by } \spadesuit.
$$

This is a quadratic in $H(s)$:

$$
\frac{1}{2}s(H(s))^2 - H(s) + \frac{1}{2}s = 0
$$

$$
\Rightarrow H(s) = \frac{1 \pm \sqrt{1 - 4\frac{1}{2}s\frac{1}{2}s}}{s} = \frac{1 \pm \sqrt{1 - s^2}}{s}.
$$

Which root? We know that $\mathbb{P}(T_{01} = 0) = 0$, because it must take at least one step to go from 0 to 1. With the positive root, $\lim_{s\to 0} H(0) = \lim_{s\to 0}\left(\dfrac{2}{s}\right) = \infty$; so we take the negative root instead.

$$
\text{Thus} \quad H(s) = \frac{1 - \sqrt{1 - s^2}}{s}.
$$

Check this has $\lim_{s\to 0} H(s) = 0$ by L'Hospital's Rule:

$$
\begin{aligned}
\lim_{s\to 0}\left(\frac{f(s)}{g(s)}\right) &= \lim_{s\to 0}\left(\frac{f'(s)}{g'(s)}\right) \\
&= \lim_{s\to 0}\left\{\frac{\frac{1}{2}(1-s^2)^{-1/2} \times 2s}{1}\right\} \\
&= 0.
\end{aligned}
$$

4.10 Defective, or improper, random variables

Defective random variables often arise in stochastic processes. They are random variables that *may take the value* ∞.

For example, consider the first reaching time T_{ij} from the random walk in Section 4.9. There might be a chance that *we <u>never</u> reach state j, starting from state i.*

If so, then T_{ij} *can take the value* ∞.

Alternatively, the process might guarantee that *we will <u>always</u> reach state j eventually, starting from state i.*

In that case, T_{ij} *can <u>not</u> take the value* ∞.

Definition: A random variable T is **<u>defective</u>**, or **<u>improper</u>**, if *it can take the value* ∞. *That is,*

$$T \text{ is defective if } \mathbb{P}(T = \infty) > 0.$$

<u>Thinking of $\sum_{t=0}^{\infty} \mathbb{P}(T = t)$ as $1 - \mathbb{P}(T = \infty)$</u>

Although it seems strange, when we write $\sum_{t=0}^{\infty} \mathbb{P}(T = t)$, *we are <u>not</u> including the value $t = \infty$.*

The sum $\sum_{t=0}^{\infty}$ continues without ever stopping: at no point can we say we have 'finished' all the finite values of t so we will now add on $t = \infty$. We simply *never get to $t = \infty$ when we take $\sum_{t=0}^{\infty}$.*

For a defective random variable T, this means that

$$\sum_{t=0}^{\infty} \mathbb{P}(T = t) < 1,$$

because we are missing the positive value of $\mathbb{P}(T = \infty)$.

All probabilities of T must still sum to 1, so we have

$$1 = \sum_{t=0}^{\infty} \mathbb{P}(T = t) + \mathbb{P}(T = \infty),$$

in other words

$$\sum_{t=0}^{\infty} \mathbb{P}(T = t) = 1 - \mathbb{P}(T = \infty).$$

PGFs for defective random variables

When T is defective, the PGF of T is *defined as*

$$G_T(s) = \mathbb{E}(s^T) = \sum_{t=0}^{\infty} \mathbb{P}(T = t) s^t \quad \text{for } |s| < 1.$$

This means that:

- the term for $\mathbb{P}(T = \infty) s^{\infty}$ has been missed out;

- $G_T(s)$ is *only* equal to $\mathbb{E}(s^T)$ for $|s| < 1$, because $s^{\infty} = 0$ when $|s| < 1$ so the missing term doesn't matter;

- for $|s| > 1$, $\mathbb{E}(s^T)$ is infinite or undefined because s^{∞} is infinite or undefined. The function $G_T(s)$ might still exist and be well-behaved for $|s| > 1$, but it is no longer equal to $\mathbb{E}(s^T)$.

- By Abel's Theorem, the function $G_T(s)$ is continuous at $s = 1$, although it no longer equals $\mathbb{E}(s^T)$ when $s = 1$. (Note that the function $\mathbb{E}(s^T)$ is *not* continuous at $s = 1$.) In fact,

$$\lim_{s \uparrow 1} G_T(s) = G_T(1) = \sum_{t=0}^{\infty} \mathbb{P}(T = t) = 1 - \mathbb{P}(T = \infty),$$

from above.

This last fact gives us our **test for defectiveness** in a random variable.

The random variable T is defective if and only if $G_T(1) < 1$.

Summary of defective random variables:

The concept of defective random variables might seem hard to understand. However, you only have to remember the following points.

1. The random variable T is defective if and only if $G_T(1) < 1$.

2. If $G_T(1) < 1$, then the probability that T takes the value ∞ is $\mathbb{P}(T = \infty) = 1 - G_T(1)$.

3. When you are asked to find $\mathbb{E}(T)$ in a context where T might be defective (e.g. $T = T_{ij}$ is a first reaching time in a random walk), you must do the following:

   - First check whether T is defective: *is $G_T(1) < 1$ or $= 1$?*
   - If T is *not* defective $(G_T(1) = 1)$, then $\mathbb{E}(T) = G'_T(1)$ *as usual.*
   - If T *is* defective $(G_T(1) < 1)$, then there is a positive chance that $T = \infty$. This means that $\mathbb{E}(T) = \infty$, *Var*$(T) = \infty$, *and* $\mathbb{E}(T^a) = \infty$ *for any power* a.

     $\mathbb{E}(T)$ and $\text{Var}(T)$ can not be found using the PGF when T is defective: you will get the wrong answer.

Example of defective random variables: finding $\mathbb{E}(T_{01})$ in the random walk

Recall the random walk example in Section 4.9. We defined the first reaching time, T_{01}, as the number of steps taken to get from state 0 to state 1.

In Section 4.9 we found the PGF of T_{01}, which contains all information about the distribution of T_{01}:

$$\text{PGF of } T_{01} = H(s) = \frac{1 - \sqrt{1 - s^2}}{s} \text{ for } |s| < 1.$$

We now wish to use the PGF to find $\mathbb{E}(T_{01})$. We must bear in mind that T_{01} *is a reaching time, so it might be defective: there might be a possibility that we* *never reach state 1, starting from state 0.*

Questions:

a) What is $\mathbb{E}(T_{01})$?

b) What happens if we try to use the Law of Total Expectation instead of the PGF to find $\mathbb{E}(T_{01})$?

a) First check if T_{01} is defective.

T_{01} is defective if and only if $\lim_{s \uparrow 1} H(s) < 1$.

But $\lim_{s \uparrow 1} H(s) = H(1) = \frac{1 - \sqrt{1 - 1^2}}{1} = 1$. So T_{01} is __not__ defective, and we can *proceed with finding $\mathbb{E}(T_{01}) = \lim_{s \to 1} H'(s)$.*

$$H(s) = \frac{1 - \sqrt{1 - s^2}}{s} = s^{-1} - \left(s^{-2} - 1\right)^{1/2}$$

$$\text{So} \quad H'(s) = -s^{-2} - \frac{1}{2}\left(s^{-2} - 1\right)^{-1/2}\left(-2s^{-3}\right)$$

Thus

$$\mathbb{E}(T_{01}) = \lim_{s \uparrow 1} H'(s) = \lim_{s \uparrow 1}\left(-\frac{1}{s^2} + \frac{1}{s^3\sqrt{\frac{1}{s^2} - 1}}\right) = \infty.$$

So (although $\mathbb{P}(T_{01} = \infty) = 0$), the expectation is infinite: $\mathbb{E}(T_{01}) = \infty$.

b) Using the law of total expectation, conditioning on the outcome of the first step Y_1:

$$
\begin{aligned}
\mathbb{E}(T_{01}) &= \mathbb{E}\{\mathbb{E}(T_{01} \mid Y_1)\} \\
&= \mathbb{E}(T_{01} \mid Y_1 = 1)\mathbb{P}(Y_1 = 1) + \mathbb{E}(T_{01} \mid Y_1 = -1)\mathbb{P}(Y_1 = -1) \\
&= 1 \times \tfrac{1}{2} + \{1 + \mathbb{E}(T_{-1,1})\} \times \tfrac{1}{2} \\
&= 1 + \tfrac{1}{2}\mathbb{E}(T_{01} + T_{01}) \\
&= 1 + \tfrac{1}{2} \times 2\mathbb{E}(T_{01}) \\
\Rightarrow \quad \mathbb{E}(T_{01}) &= 1 + \mathbb{E}(T_{01}). \qquad \textit{So 0 = 1 ???}
\end{aligned}
$$

What's wrong? The expectations were not finite, so the law of total expectation was invalid. We can deduce from $\mathbb{E}(T_{01}) = 1 + \mathbb{E}(T_{01})$ that the expectation is infinite.

Note: **(Non-examinable)** If T_{ij} is defective in the random walk, $\mathbb{E}(s^{T_{ij}})$ is not continuous at $s = 1$. In Section 4.9 we saw that we had to solve a quadratic equation to find $H(s) = \mathbb{E}(s^{T_{ij}})$. The negative root solution for $H(s)$ generally represents $\mathbb{E}(s^{T_{ij}})$ for $s < 1$. At $s = 1$, the required solution for $\mathbb{E}(s^{T_{ij}})$ suddenly flips from the $-$ root to the $+$ root of the quadratic. This explains how $\mathbb{E}(s^{T_{ij}})$ can be discontinuous as $s \uparrow 1$, even though the negative root for $H(s)$ is continuous as $s \uparrow 1$ and all the working of Section 4.9 still applies for $s = 1$. The explanation is that we suddenly switch from the $-$ root to the $+$ root at $s = 1$.

When $|s| > 1$, the conditional expectations are not finite so the working of Section 4.9 no longer applies.

The problem with being abstract . . .

- **Each card has a letter on one side and a number on the other.**

- **We wish to test the following rule:**

If the card has a D on one side, then it has a 3 on the other side.

- **Which card or cards should you turn over, and ONLY these cards, in order to test the rule?**

| D | K | 3 | 7 |
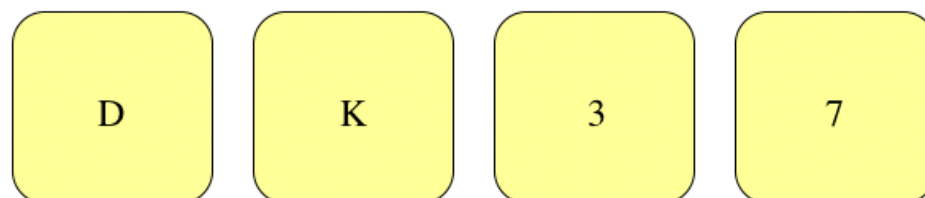
At a party . . .

If you are drinking alcohol, you must be 18 or over.

- **Each card has the person's age on one side, and their drink on the other side.**

- **Which card or cards should you turn over, and ONLY these cards, in order to test the rule?**

| Wine | Juice | 25 | 17 |
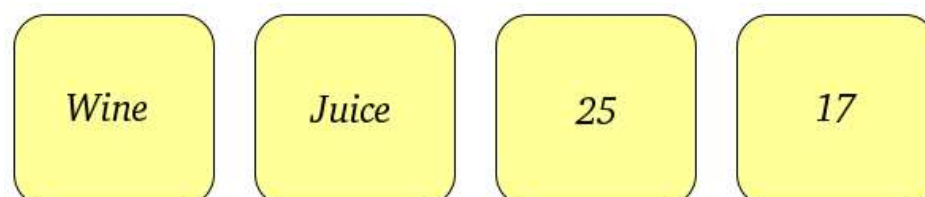
Chapter 5: Mathematical Induction

So far in this course, we have seen some techniques for dealing with stochastic processes: first-step analysis for hitting probabilities (Chapter 2), first-step analysis for expected reaching times (Chapter 3), and probability generating functions (Chapter 4), which are especially useful for dealing with sums of random variables. We now look at one more tool that is often useful for exploring properties of stochastic processes: *proof by mathematical induction.*

5.1 Proving things in mathematics

There are many different ways of constructing a formal proof in mathematics. Some examples are:

- **Proof by counterexample:** a proposition is proved to be *not generally true* because a *particular example* is found for which it is not true.

- **Proof by contradiction:** this can be used either to prove a proposition is true or to prove that it is false. To prove that the proposition is *true* (say), we start by *assuming that it is false.* We then explore the consequences of this assumption until we reach a contradiction, e.g. $0 = 1$. Therefore something must have gone wrong, and the only thing we weren't sure about was our initial assumption that the proposition is false — so our initial assumption must be wrong and the proposition is proved true.

  A famous proof of this sort is the proof that there are infinitely many prime numbers. We start by assuming that there are *finitely* many primes, so they can be listed as $p_1, p_2, \ldots, p_n$, where p_n is the largest prime number. But then the number $p_1 \times p_2 \times \ldots \times p_n + 1$ must also be prime, because it is not divisible by any of the smaller primes. Furthermore this number is definitely bigger than p_n. So we have contradicted the idea that there was a 'biggest' prime called p_n, and therefore there are infinitely many primes.

- **Proof by mathematical induction:** in mathematical induction, we start with a formula that we *suspect* is true. For example, I might *suspect* from

observation that $\sum_{k=1}^{n} k = n(n+1)/2$. I might have tested this formula for many different values of n, but of course I can never test it for *all* values of n. Therefore I need to prove that the formula is *always* true.

The idea of mathematical induction is to say: *suppose* the formula is true for all n up to the value $n = 10$ (say). Can I prove that, *if* it is true for $n = 10$, *then* it will also be true for $n = 11$? And *if* it is true for $n = 11$, then it will also be true for $n = 12$? And so on.

In practice, we usually start lower than $n = 10$. We usually take the very easiest case, $n = 1$, and prove that the formula is true for $n = 1$: LHS $= \sum_{k=1}^{1} k = 1 = 1 \times 2/2 =$ RHS. Then we prove that, *if* the formula is ever true for $n = x$, *then* it will always be true for $n = x + 1$. Because it is true for $n = 1$, it must be true for $n = 2$; and because it is true for $n = 2$, it must be true for $n = 3$; and so on, for all possible n. Thus the formula is proved.

Mathematical induction is therefore a bit like a *first-step analysis for proving things: prove that wherever we are now, the* <u>**next step**</u> *will always be OK. Then if we were OK at the very beginning, we will be OK for ever.*

The method of mathematical induction for proving results is very important in the study of Stochastic Processes. This is because a stochastic process builds up one step at a time, and mathematical induction works on the same principle.

Example: We have already seen examples of inductive-type reasoning in this course. For example, in Chapter 2 for the Gambler's Ruin problem, using the method of repeated substitution to solve for $p_x = \mathbb{P}(\text{Ruin} \,|\, \text{start with \$x})$, we discovered that:

- $p_2 = 2p_1 - 1$
- $p_3 = 3p_1 - 2$
- $p_4 = 4p_1 - 3$

We deduced that $p_x = xp_1 - (x - 1)$ *in general.*

To prove this properly, we should have used the method of mathematical induction.

5.2   Mathematical Induction by example

This example explains the style and steps needed for a proof by induction.

Question: Prove by induction that $\displaystyle\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$ for any integer n. $(\star)$

Approach: follow the steps below.

(i) First verify that the formula is true for a *base case*: usually the smallest appropriate value of n (e.g. $n = 0$ or $n = 1$). Here, the smallest possible value of n is $n = 1$, because we can't have $\sum_{k=1}^{0}$.

First verify $(\star)$ is true when $n = 1$.

$$LHS = \sum_{k=1}^{1} k = 1.$$

$$RHS = \frac{1 \times 2}{2} = 1 = LHS.$$

So $(\star)$ is proved for $n = 1$.

(ii) Next suppose that formula $(\star)$ is true for *all values of n up to and including some value x.* (We have already established that this is the case for $x = 1$).

Using the hypothesis that $(\star)$ is true for all values of n up to and including x, *prove* that it is therefore true for the value $n = x + 1$.

Now suppose that $(\star)$ is true for $n = 1, 2, \ldots, x$ for some x.

Thus we can assume that $\displaystyle\sum_{k=1}^{x} k = \frac{x(x+1)}{2}.$ (a)

                    ((a) for 'allowed' info)

We need to show that if $(\star)$ holds for $n = x$, then it must also hold for $n = x + 1$.

Require to prove that

$$\sum_{k=1}^{x+1} k = \frac{(x+1)(x+2)}{2} \qquad (\star\star)$$

(Obtained by putting $n = x + 1$ in $(\star)$).

$$\text{\textit{LHS of} } (\star\star) = \sum_{k=1}^{x+1} k \;=\; \sum_{k=1}^{x} k \;+\; (x+1) \qquad \textit{by expanding the sum}$$

$$= \frac{x(x+1)}{2} \;+\; (x+1) \qquad \textit{using allowed info (a)}$$

$$= (x+1)\left(\frac{x}{2} + 1\right) \qquad \textit{rearranging}$$

$$= \frac{(x+1)(x+2)}{2}$$

$$= \textit{RHS of } (\star\star).$$

This shows that:

$$\sum_{k=1}^{n} k \;=\; \frac{n(n+1)}{2} \qquad \textit{when } n = x + 1.$$

So, assuming $(\star)$ is true for $n = x$, it is also true for $n = x + 1$.

(iii) Refer back to the base case: if it is true for $n = 1$, then it is true for $n = 1+1 = 2$ by (ii). If it is true for $n = 2$, it is true for $n = 2 + 1 = 3$ by (ii). We could go on forever. This proves that the formula $(\star)$ is true for all n.

We proved $(\star)$ true for $n = 1$, thus $(\star)$ is true for all integers $n \geq 1$. $\square$

General procedure for proof by induction

The procedure above is quite standard. The inductive proof can be summarized like this:

Question: prove that $f(n) = g(n)$ for all integers $n \geq 1$.

Base case: $n = 1$. _Prove that_ $f(1) = g(1)$ _using_

$$
\begin{aligned}
LHS &= f(1) \\
&= \vdots \\
&= g(1) = RHS.
\end{aligned}
$$

General case: _suppose the formula is true for_ $n = x$: _so_ $f(x) = g(x)$. _Prove that the formula is therefore true for_ $n = x + 1$:

$$
\begin{aligned}
LHS &= f(x+1) \\
&= \left\{ \begin{array}{c} \text{some expression breaking down } f(x+1) \\ \text{into } f(x) \text{ and an extra term in } x+1 \end{array} \right\} \\
&= \left\{ \begin{array}{c} \text{formula is true for } n = x, \text{ so substitute} \\ f(x) = g(x) \text{ in the line above} \end{array} \right\} \\
&= \{\text{do some working}\} \\
&= g(x+1) \\
&= RHS.
\end{aligned}
$$

Conclude: the formula is true for $n = 1$, _so it is true for_ $n = 2$, $n = 3$, $n = 4, \ldots$

It is therefore true for all integers $n \geq 1$.   □

5.3 Some harder examples of mathematical induction

Induction problems in stochastic processes are often trickier than usual. Here are some possibilities:

- Backwards induction: start with base case $n = N$ and go backwards, instead of starting at base case $n = 1$ and going forwards.

- Two-step induction, where the proof for $n = x + 1$ relies not only on the formula being true for $n = x$, but also on it being true for $n = x - 1$.

The first example below is hard probably because it is too easy. The second example is an example of a two-step induction.

Example 1: Suppose that $p_0 = 1$ and $p_x = \alpha p_{x+1}$ for all $x = 1, 2, \ldots$. Prove by mathematical induction that $p_n = 1/\alpha^n$ for $n = 0, 1, 2, \ldots$.

Wish to prove
$$p_n = \frac{1}{\alpha^n} \quad \text{for} \ \ n = 0, 1, 2, \ldots \qquad (\star)$$
Information given:
$$p_{x+1} = \frac{1}{\alpha} p_x \qquad (G_1)$$
$$p_0 = 1 \qquad (G_2)$$
Base case: $n = 0$.
$$\text{LHS} = p_0 = 1 \ \text{ by information given } (G_2).$$
$$\text{RHS} = \frac{1}{\alpha^0} = \frac{1}{1} = 1 = \text{LHS}.$$
Therefore $(\star)$ *is true for the base case* $n = 0$.

General case: suppose that $(\star)$ *is true for* $n = x$, *so we can assume*
$$p_x = \frac{1}{\alpha^x}. \qquad (a)$$
Wish to prove that $(\star)$ *is also true for* $n = x + 1$: *i.e.*
$$\text{RTP} \qquad p_{x+1} = \frac{1}{\alpha^{x+1}}. \qquad (\star\star)$$

$$\text{LHS of }(\star\star) = p_{x+1} \;=\; \frac{1}{\alpha} \times p_x \qquad \textit{by given } (G_1)$$

$$= \;\frac{1}{\alpha} \times \frac{1}{\alpha^x} \qquad \textit{by allowed (a)}$$

$$= \;\frac{1}{\alpha^{x+1}}$$

$$= \;\textbf{RHS of }(\star\star).$$

So if formula $(\star)$ is true for $n = x$, it is true for $n = x + 1$. We have shown it is true for $n = 0$, so it is true for all $n = 0, 1, 2, \ldots$. $\square$

Example 2: Gambler's Ruin. In the Gambler's Ruin problem in Section 2.6, we have the following situation:

- $p_x = \mathbb{P}(\text{Ruin} \,|\, \text{start with \$x)$;

- We know from first-step analysis that $p_{x+1} = 2p_x - p_{x-1}$ \qquad (G_1)

- We know from common sense that $p_0 = 1$ \qquad (G_2)

- By direct substitution into (G_1), we obtain:

$$p_2 \;=\; 2p_1 - 1$$
$$p_3 \;=\; 3p_1 - 2$$

- We develop a suspicion that for all $x = 1, 2, 3, \ldots,$

$$p_x = xp_1 - (x - 1) \qquad (\star)$$

- We wish to prove $(\star)$ by mathematical induction.

For this example, *our given information, in (G_1), expresses p_{x+1} in terms of both p_x and p_{x-1}, so we need two base cases. Use $x = 1$ and $x = 2$.*

Wish to prove $p_x = xp_1 - (x-1)$ $(\star)$.
Base case $x = 1$:

$$LHS = p_1.$$

$$RHS = 1 \times p_1 - 0 = p_1 = LHS.$$

$\therefore$ *formula* $(\star)$ *is true for base case* $x = 1$.

Base case $x = 2$:

$$LHS = p_2 = 2p_1 - 1 \quad \text{by information given } (G_1)$$

$$RHS = 2 \times p_1 - 1 = LHS.$$

$\therefore$ *formula* $(\star)$ *is true for base case* $x = 2$.

General case: suppose that $(\star)$ *is true for* <u>*all* x *up to* $x = k$</u>.
So we are allowed:

$$
\begin{array}{llll}
(x = k) & p_k & = & kp_1 - (k-1) & (a_1) \\
(x = k-1) & p_{k-1} & = & (k-1)p_1 - (k-2) & (a_2)
\end{array}
$$

Wish to prove that $(\star)$ *is also true for* $x = k+1$, *i.e.*

$$RTP \quad p_{k+1} = (k+1)p_1 - k. \qquad (\star\star)$$

$$
\begin{aligned}
LHS \text{ of } (\star\star) \;&=\; p_{k+1} \\[4pt]
&=\; 2p_k - p_{k-1} \quad \text{by given information } (G_1) \\[4pt]
&=\; 2\Big\{ kp_1 - (k-1) \Big\} - \Big\{ (k-1)p_1 - (k-2) \Big\} \\
&\quad \text{by allowed } (a_1) \text{ and } (a_2) \\[4pt]
&=\; p_1\Big\{ 2k - (k-1) \Big\} - \Big\{ 2(k-1) - (k-2) \Big\} \\[4pt]
&=\; (k+1)p_1 - k \\[4pt]
&=\; RHS \quad \text{of } (\star\star)
\end{aligned}
$$

So if formula $(\star)$ *is true for* $x = k-1$ *and* $x = k$, *it is true for* $x = k+1$. *We have shown it is true for* $x = 1$ *and* $x = 2$, *so it is true for all* $x = 1, 2, 3, \ldots$. $\qquad \square$
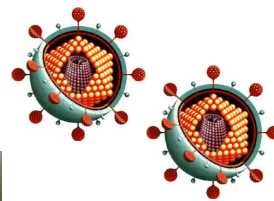
Chapter 6: Branching Processes:
The Theory of Reproduction

Viruses

Aphids

Royalty

DNA

Although the early development of Probability Theory was motivated by problems in gambling, probabilists soon realised that, if they were to continue as a breed, they must also study *reproduction.*

Reproduction is a complicated business, but considerable insights into population growth can be gained from simplified models. The **Branching Process** is a simple but elegant model of population growth. It is also called the **Galton-Watson Process,** because some of the early theoretical results about the process derive from a correspondence between Sir Francis Galton and the Reverend Henry William Watson in 1873. Francis Galton was a cousin of Charles Darwin. In later life, he developed some less elegant ideas about reproduction — namely eugenics, or selective breeding of humans. Luckily he is better remembered for branching processes.

6.1 Branching Processes

Consider some sort of *population* consisting of reproducing individuals.

Examples: living things (animals, plants, bacteria, royal families);
diseases; computer viruses;
rumours, gossip, lies (one lie always leads to another!)

Start conditions: start at time $n = 0$, with a single individual.

Each individual: lives for 1 unit of time. At time $n = 1$, it produces a family of offspring, and immediately dies.

How many offspring? Could be 0, 1, 2, ... This is the family size, Y. ("Y" stands for "number of Young").

Each offspring: lives for 1 unit of time. At time $n = 2$, it produces its own family of offspring, and immediately dies.

and so on...

Assumptions

1. *All individuals reproduce independently of each other.*

2. *The family sizes of different individuals are independent, identically distributed random variables. Denote the family size by Y (number of Young).*

Family size distribution, Y $\mathbb{P}(Y = k) = p_k$.

y	0	1	2	3	4	$\cdots$
$P(Y=y)$	p_0	p_1	p_2	p_3	p_4	$\cdots$
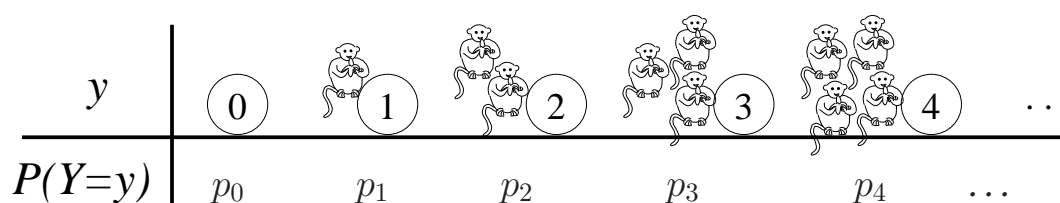
Definition: A **branching process** is defined as follows.

- Single individual at time $n = 0$.

- Every individual lives exactly one unit of time, then produces Y offspring, and dies.

- The number of offspring, Y, takes values $0, 1, 2, \ldots$, and the probability of producing k offspring is $\mathbb{P}(Y = k) = p_k$.

- All individuals reproduce independently. Individuals $1, 2, \ldots, n$ have family sizes $Y_1, Y_2, \ldots Y_n$, where *each Y_i has the same distribution as Y.*

- Let Z_n be the *number of individuals born at time n, for $n = 0, 1, 2, \ldots$. Interpret 'Z_n' as the 'si$\underline{Z}$e' of generation n.*

- Then the branching process is $\{Z_0, Z_1, Z_2, Z_3, \ldots\} = \{Z_n : n \in \mathbb{N}\}$.

Definition: The **state** of the branching process at time n is z_n, *where each z_n can take values $0, 1, 2, 3, \ldots$. Note that $z_0 = 1$ always.*
z_n represents the si$\underline{z}$e of the population at time n.

Note: When we want to say that two random variables X and Y have the same distribution, we write: $X \sim Y$.
For example: $Y_i \sim Y$, *where Y_i is the family size of any individual i.*

Note: The definition of the branching process is easily generalized to start with more than one individual at time $n = 0$.

Branching Process
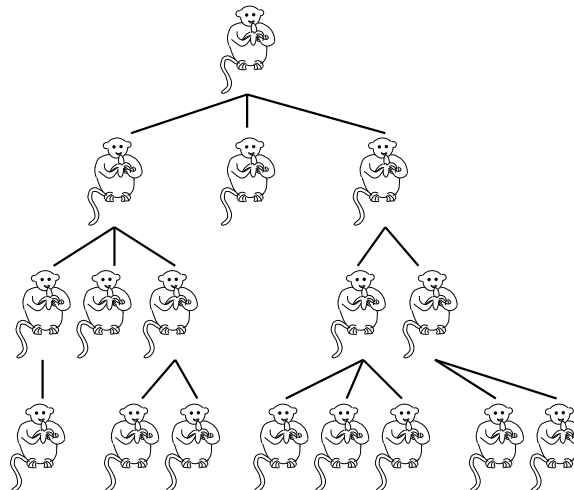
6.2 Questions about the Branching Process

When we have a situation that can be modelled by a branching process, there are several questions we might want to answer.

If the branching process is just beginning, what will happen in the future?

1. What can we find out about the distribution of Z_n (the population siZe at generation n)?

   - can we find the mean and variance of Z_n?
     — *yes, using the probability generating function of family size, Y;*

   - can we find the whole distribution of Z_n?
     — *for special cases of the family size distribution Y, we can find the PGF of Z_n explicitly;*

   - can we find the probability that the population has become **extinct** by generation n, $\quad \mathbb{P}(Z_n = 0) \quad$?
     — *for special cases where we can find the PGF of Z_n (as above).*

2. What can we find out about <u>eventual</u> extinction?

   - can we find the <u>probability</u> of eventual extinction, $\quad \mathbb{P}\left(\lim_{n\to\infty} Z_n = 0\right) \quad$?
     — *yes, always: using the PGF of Y.*

   - can we find general <u>conditions</u> for eventual extinction?
     — *yes: we can find conditions that guarantee that extinction will occur with probability 1.*

   - if eventual extinction is definite, can we find the distribution of the time to extinction?
     — *for special cases where we can find the PGF of Z_n (as above).*

Example: Modelling cancerous growths. Will a colony of cancerous cells become extinct before it is sufficiently large to overgrow the surrounding tissue?

If the branching process is already in progress, what happened in the past?

1. How long has the process been running?

   - *how many generations do we have to go back to get to the single common ancestor?*

2. What has been the distribution of family size over the generations?

3. What is the total number of individuals (over all generations) up to the present day?

Example: It is believed that all humans are descended from a single female ancestor, who lived in Africa. How long ago?
— *estimated at approximately 200,000 years.*
What has been the mean family size over that period?
— *probably very close to 1 female offspring per*
female adult: e.g. estimate = 1.002.

6.3 Analysing the Branching Process

Key Observation: *every individual in every generation starts a new, independent branching process, as if the whole process were starting at the beginning again.*
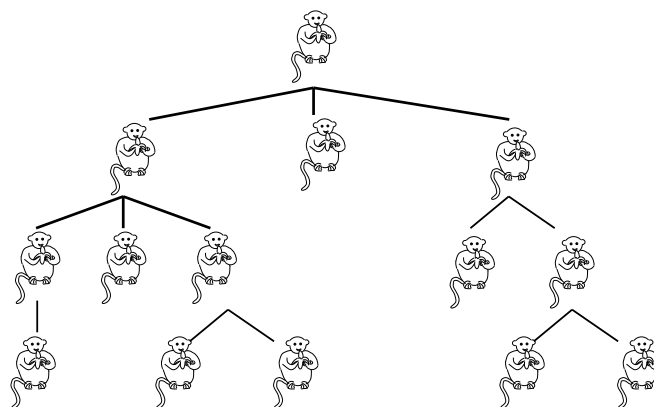
Z_n as a randomly stopped sum

Most of the interesting properties of the branching process centre on the distribution of Z_n (the population size at time n). Using the Key Observation from overleaf, we can find an expression for the probability generating function of Z_n.

Consider the following.

- *The population size at time $n - 1$ is given by Z_{n-1}.*

- *Label the individuals at time $n - 1$ as $1, 2, 3, \ldots, Z_{n-1}$.*

- *Each individual $1, 2, \ldots, Z_{n-1}$ starts a new branching process. Let $Y_1, Y_2, \ldots, Y_{Z_{n-1}}$ be the random family sizes of the individuals $1, 2, \ldots, Z_{n-1}$.*

- *The number of individuals at time n, Z_n, is equal to the total number of offspring of the individuals $1, 2, \ldots, Z_{n-1}$. That is,*

$$Z_n = \sum_{i=1}^{Z_{n-1}} Y_i \,.$$

Thus Z_n is a <u>randomly stopped sum</u>: a sum of $Y_1, Y_2, \ldots$, randomly stopped by the random variable Z_{n-1}.

Note: 1. *Each $Y_i \sim Y$: that is, each individual $i = 1, \ldots, Z_{n-1}$ has the same family size distribution.*

2. *$Y_1, Y_2, \ldots, Y_{Z_{n-1}}$ are independent.*

Probability Generating Function of Z_n

Let $G_Y(s) = \mathbb{E}(s^Y)$ be the probability generating function of Y.
(Recall that Y is the *number of Young of an individual: the family size.*)

Now Z_n is a randomly stopped sum: it is the sum of $Y_1, Y_2, \ldots$, stopped by the random variable Z_{n-1}. So we can use Theorem 4.6 (Chapter 4) to express the PGF of Z_n directly in terms of the PGFs of Y and Z_{n-1}.

By Theorem 4.6, if $Z_n = Y_1 + Y_2 + \ldots + Y_{Z_{n-1}}$, and Z_{n-1} is itself random, then the PGF of Z_n is given by:

$$G_{Z_n}(s) = G_{Z_{n-1}}\Big(G_Y(s)\Big), \qquad (\clubsuit)$$

where $G_{Z_{n-1}}$ is the PGF of the random variable Z_{n-1}.

For ease of notation, we can write:

$$G_{Z_n}(s) = G_n(s), \qquad G_{Z_{n-1}}(s) = G_{n-1}(s), \qquad \text{and so on.}$$

Note that $Z_1 = Y$ *(the number of individuals born at time $n = 1$),*
so we can also write:

$$G_Y(s) = G_1(s) = G(s) \qquad \text{(for simplicity).}$$

Thus, *from $(\clubsuit)$,*

$$\boxed{\; G_n(s) = G_{n-1}\Big(G(s)\Big) \qquad \textit{(Branching Process Recursion Formula.)} \;}$$

Note:

1. $G_n(s) = \mathbb{E}\left(s^{Z_n}\right)$, *the PGF of the population size at time n, Z_n.*

2. $G_{n-1}(s) = \mathbb{E}\left(s^{Z_{n-1}}\right)$, *the PGF of the population size at time $n-1$, Z_{n-1}.*

3. $G(s) = \mathbb{E}\left(s^Y\right) = \mathbb{E}\left(s^{Z_1}\right)$, *the PGF of the family size, Y.*

We are trying to find the PGF of Z_n, the population size at time n.

So far, we have: $G_n(s) = G_{n-1}\Big(G(s)\Big).$ $(\star)$

But by the same argument,

$$G_{n-1}(r) = G_{n-2}\Big(G(r)\Big).$$

(use r instead of s to avoid confusion in the next line.)

Substituting in $(\star)$,

$$
\begin{aligned}
G_n(s) &= G_{n-1}\Big(G(s)\Big) \\[2mm]
&= G_{n-1}(r) \qquad \text{where } r = G(s) \\[2mm]
&= G_{n-2}\Big(G(r)\Big) \\[2mm]
&= G_{n-2}\Big(G\big(G(s)\big)\Big) \qquad \text{replacing } r = G(s).
\end{aligned}
$$

By the same reasoning, we will obtain:

$$G_n(s) = G_{\underbrace{n-3}_{n-3}}\Big(\underbrace{G\big(G\big(G(s)\big)\big)}_{3 \text{ times}}\Big),$$

and so on, until we finally get:

$$
\begin{aligned}
G_n(s) &= G_{n-(n-1)}\Big(\underbrace{G\big(G\big(G\big(\ldots G(s)\ldots\big)\big)\big)}_{n-1 \text{ times}}\Big) \\[3mm]
&= \underbrace{G_1}_{=G}\Big(\underbrace{G\big(G\big(G\big(\ldots G(s)\ldots\big)\big)\big)}_{n-1 \text{ times}}\Big) \\[3mm]
&= \underbrace{G\big(G\big(G\big(\ldots G(s)\ldots\big)\big)\big)}_{n \text{ times}}.
\end{aligned}
$$

We have therefore proved the following Theorem.

Theorem 6.3: Let $G(s) = \mathbb{E}(s^Y) = \sum_{y=0}^{\infty} p_y s^y$ be the PGF of the family size distribution, Y. Let $Z_0 = 1$ (start from a single individual at time 0), and let Z_n be the population size at time n $(n = 0, 1, 2, \ldots)$. Let $G_n(s)$ be the PGF of the random variable Z_n. Then

$$G_n(s) = \underbrace{G\Big(G\Big(G\Big(\ldots G(s) \ldots\Big)\Big)\Big)}_{n \text{ times}}. \qquad \square$$

Note: $G_n(s) = \underbrace{G\Big(G\Big(G\Big(\ldots G(s) \ldots\Big)\Big)\Big)}_{n \text{ times}}$ is called the n-***fold iterate of*** G.

We have therefore found an expression for the PGF of the population size at generation n, although there is no guarantee that it is possible to write it down or manipulate it very easily for large n. For example, if Y has a Poisson(λ) distribution, then $G(s) = e^{\lambda(s-1)}$, and already by generation $n = 3$ we have the following fearsome expression for $G_3(s)$:

$$G_3(s) = e^{\lambda\left(e^{\lambda\left(e^{\lambda(s-1)}-1\right)}-1\right)}. \qquad \text{(Or something like that!)}$$

However, in some circumstances we can find quite reasonable closed-form expressions for $G_n(s)$, notably when Y has a Geometric distribution. In addition, for <u>any</u> distribution of Y we can use the expression $G_n(s) = G_{n-1}\Big(G(s)\Big)$ to derive properties such as the mean and variance of Z_n, and the probability of eventual extinction ($\mathbb{P}(Z_n = 0)$ for some n).

6.4 What does the distribution of Z_n look like?

Before deriving the mean and the variance of Z_n, it is helpful to get some intuitive idea of how the branching process behaves. For example, it seems reasonable to calculate the mean, $\mathbb{E}(Z_n)$, to find out what we expect the population size to be in n generations time, but why are we interested in $\text{Var}(Z_n)$?

The answer is that Z_n usually has a "boom-or-bust" distribution: either the population will take off (boom), and the population size grows quickly, or the population will fail altogether (bust). In fact, if the population fails, it is likely to do so very quickly, within the first few generations. This explains why we are

interested in $\text{Var}(Z_n)$. A huge variance will alert us to the fact that the process does not cluster closely around its mean values. In fact, the mean might be almost useless as a measure of what to expect from the process.
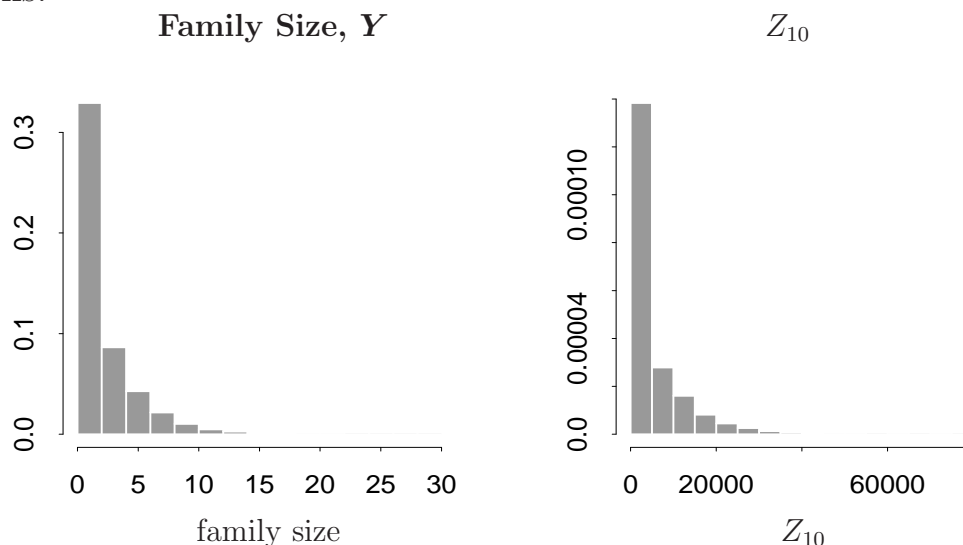
Simulation 1: $Y \sim \text{Geometric}(p = 0.3)$

The following table shows the results from 10 simulations of a branching process, where the family size distribution is $Y \sim \text{Geometric}(p = 0.3)$.

Simulation	Z_0	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
1	1	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	0
3	1	4	19	42	81	181	433	964	2276	5383	12428
4	1	3	3	5	3	15	29	86	207	435	952
5	1	0	0	0	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	0	0	0	0
7	1	2	8	26	68	162	360	845	2039	4746	10941
8	1	1	0	0	0	0	0	0	0	0	0
9	1	1	0	0	0	0	0	0	0	0	0
10	1	1	4	13	18	39	104	294	690	1566	3534

Often, the population is extinct by generation 10. However, when it is not extinct, it can take enormous values (12428, 10941, ...).

The same simulation was repeated 5000 times to find the empirical distribution of the population size at generation 10 (Z_{10}). The figures below show the distribution of family size, Y, and the distribution of Z_{10} from the 5000 simulations.

Family Size, Y Z_{10}

In this example, the family size is rather variable, but the variability in Z_{10} is enormous (note the range on the histogram from 0 to 60,000). Some statistics are:

```
Proportion of samples extinct by generation 10:  0.436
```

```
Summary of Zn:
    Min  1st Qu  Median    Mean  3rd Qu     Max
      0       0    1003    4617    6656   82486
```

```
Mean of Zn:       4617.2
Variance of Zn:  53937785.7
```

So the empirical variance is $\mathrm{Var}(Z_{10}) = 5.39 \times 10^7$. This perhaps contains more useful information than the mean value of 4617. The distribution of Z_n has 43.6% of zeros, but (when it is non-zero) takes values up to *82,486*. Is it really useful to summarize such a distribution by the single mean value 4617?

For interest, out of the 5000 simulations, there were only 35 (0.7%) that had a value for Z_{10} greater than 0 but less than 100. This emphasizes the "boom-or-bust" nature of the distribution of Z_n.

Simulation 2: $Y \sim$ Geometric$(p = 0.5)$

We repeat the simulation above with a different value for p in the Geometric family size distribution: this time, $p = 0.5$. The family size distribution is therefore $Y \sim$ *Geometric*$(p = 0.5)$.

Simulation	Z_0	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
1	1	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0	0	0
6	1	7	9	17	15	20	19	8	7	13	35
7	1	2	5	2	5	8	8	3	3	0	0
8	1	2	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0	0

This time, almost all the populations become extinct. We will see later that this value of p (just) guarantees eventual extinction with probability 1.

The family size distribution, $Y \sim \text{Geometric}(p = 0.5)$, and the results for Z_{10} from 5000 simulations, are shown below. Family sizes are often zero, but families of size 2 and 3 are not uncommon. It seems that this is not enough to save the process from extinction. This time, the maximum population size observed for Z_{10} from 5000 simulations was only 56, and the mean and variance of Z_{10} are much smaller than before.

Family Size, Y Z_{10}
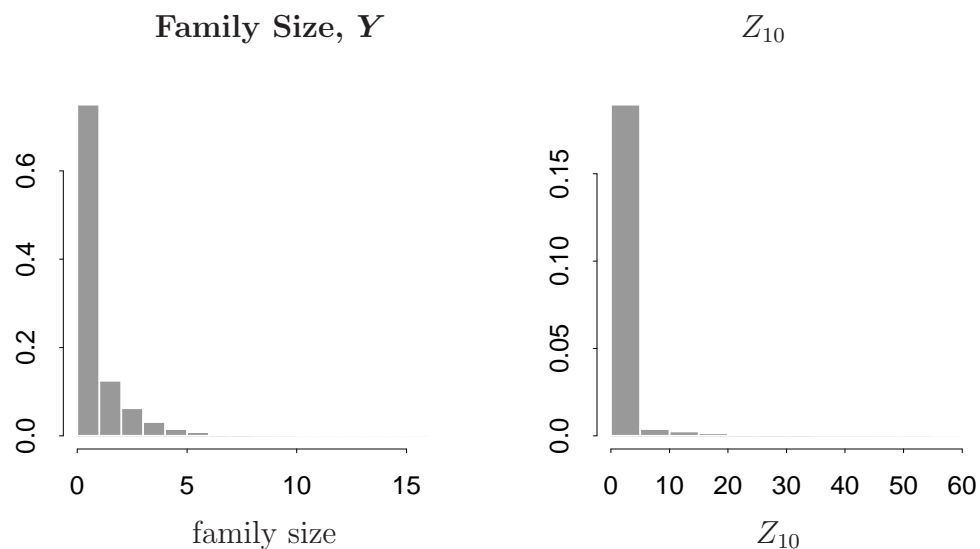
```
Proportion of samples extinct by generation 10:  0.9108


Summary of Zn:
     Min  1st Qu  Median     Mean  3rd Qu      Max
       0       0       0    0.965       0       56


Mean of Zn:       0.965
Variance of Zn:  19.497
```

What happens for larger values of p?

It was mentioned above that $Y \sim \text{Geometric}(p = 0.5)$ just guarantees eventual extinction with probability 1. For $p > 0.5$, extinction is also guaranteed, and tends to happen quickly. For example, when $p = 0.55$, over 97% of simulated populations are already extinct by generation 10.

6.5 Mean and variance of Z_n

The previous section has given us a good idea of the significance and interpretation of $\mathbb{E}(Z_n)$ and $\text{Var}(Z_n)$. We now proceed to calculate them. Both $\mathbb{E}(Z_n)$ and $\text{Var}(Z_n)$ can be expressed in terms of *the mean and variance of the family size distribution, Y.*

Thus, *let $\mathbb{E}(Y) = \mu$ and let $\text{Var}(Y) = \sigma^2$. These are the mean and variance of the number of offspring of a <u>single</u> individual.*

Theorem 6.5: Let $\{Z_0, Z_1, Z_2, \ldots\}$ be a branching process with $Z_0 = 1$ (start with a single individual). Let Y denote the family size distribution, and suppose that $\mathbb{E}(Y) = \mu$. Then

$$\mathbb{E}(Z_n) = \mu^n.$$

Proof:

By page 111, $Z_n = Y_1 + Y_2 + \ldots + Y_{Z_{n-1}}$ is a randomly stopped sum:

$$Z_n = \sum_{i=1}^{Z_{n-1}} Y_i$$

Thus, from Section 3.4 (page 60),

$$
\begin{aligned}
\mathbb{E}(Z_n) &= \mathbb{E}(Y_i) \times \mathbb{E}(Z_{n-1}) \\
&= \mu \times \mathbb{E}(Z_{n-1}) \\
&= \mu\,\{\mu\mathbb{E}(Z_{n-2})\} \\
&= \mu^2\mathbb{E}(Z_{n-2}) \\
&= \vdots \\
&= \mu^{n-1}\mathbb{E}(Z_1) \\
&= \mu^{n-1} \times \mu \\
&= \mu^n. \qquad \square
\end{aligned}
$$

Examples: Consider the simulations of Section 6.4.

1. Family size $Y \sim \text{Geometric}(p = 0.3)$. ***So*** $\mu = \mathbb{E}(Y) = \dfrac{q}{p} = \dfrac{0.7}{0.3} = 2.33$.

   Expected population size by generation $n = 10$ is:

   $$\mathbb{E}(Z_{10}) = \mu^{10} = (2.33)^{10} = 4784.$$

   The theoretical value, *4784, compares well with the sample mean from 5000 simulations, 4617 (page 116).*

2. Family size $Y \sim \text{Geometric}(p = 0.5)$. ***So*** $\mu = \mathbb{E}(Y) = \dfrac{q}{p} = \dfrac{0.5}{0.5} = 1$, *and*

   $$\mathbb{E}(Z_{10}) = \mu^{10} = (1)^{10} = 1.$$

   Compares well with the sample mean of 0.965 (page 117).

Variance of Z_n

Theorem 6.5: Let $\{Z_0, Z_1, Z_2, \ldots\}$ be a branching process with $Z_0 = 1$ (start with a single individual). Let Y denote the family size distribution, and suppose that $\mathbb{E}(Y) = \mu$ and $\text{Var}(Y) = \sigma^2$. Then

$$\text{Var}(Z_n) = \begin{cases} \sigma^2 n & \text{if } \mu = 1, \\[2mm] \sigma^2 \mu^{n-1}\left(\dfrac{1 - \mu^n}{1 - \mu}\right) & \text{if } \mu \neq 1 \quad (> 1 \text{ or } < 1). \end{cases}$$

Proof:

Write $V_n = \text{Var}(Z_n)$. The proof works by finding a recursive formula for V_n.

Using the Law of Total Variance for randomly stopped sums from Section 3.4 (page 60),

$$Z_n = \sum_{i=1}^{Z_{n-1}} Y_i$$

$$\Rightarrow \quad \text{Var}(Z_n) = \{\mathbb{E}(Y_i)\}^2 \times \text{Var}(Z_{n-1}) + \text{Var}(Y_i) \times \mathbb{E}(Z_{n-1})$$

$$\Rightarrow \quad V_n = \mu^2 V_{n-1} + \sigma^2 \mathbb{E}(Z_{n-1})$$

$$\Rightarrow \quad V_n = \mu^2 V_{n-1} + \sigma^2 \mu^{n-1},$$

using $\mathbb{E}(Z_{n-1}) = \mu^{n-1}$ as above.

Also,

$$V_1 = \text{Var}(Z_1) = \text{Var}(Y) = \sigma^2.$$

Find V_n by repeated substitution:

$$V_1 = \sigma^2$$

$$V_2 = \mu^2 V_1 + \sigma^2 \mu = \mu^2 \sigma^2 + \mu \sigma^2 = \mu \sigma^2 (1 + \mu)$$

$$V_3 = \mu^2 V_2 + \sigma^2 \mu^2 = \mu^2 \sigma^2 \left(1 + \mu + \mu^2\right)$$

$$V_4 = \mu^2 V_3 + \sigma^2 \mu^3 = \mu^3 \sigma^2 \left(1 + \mu + \mu^2 + \mu^3\right)$$

$$\vdots \quad \text{etc.}$$

Completing the pattern,

$$V_n = \mu^{n-1} \sigma^2 \left(1 + \mu + \mu^2 + \ldots + \mu^{n-1}\right)$$

$$= \mu^{n-1} \sigma^2 \sum_{r=0}^{n-1} \mu^r$$

$$= \mu^{n-1} \sigma^2 \left(\frac{1 - \mu^n}{1 - \mu}\right). \qquad \text{Valid for } \mu \neq 1.$$

$$\text{(sum of first } n \text{ terms of Geometric series)}$$

When $\mu = 1$:

$$V_n \;=\; 1^{n-1}\sigma^2 \underbrace{\left(1^0 + 1^1 + \ldots + 1^{n-1}\right)}_{n \text{ times}} \;=\; \sigma^2 n.$$

Hence the result:

$$\mathrm{Var}(Z_n) = \begin{cases} \sigma^2\, n & \text{if } \mu = 1, \\[2mm] \sigma^2 \mu^{n-1} \left(\dfrac{1 - \mu^n}{1 - \mu}\right) & \text{if } \mu \neq 1. \end{cases} \qquad \square$$

Examples: Again consider the simulations of Section 6.4.

1. Family size $Y \sim \mathrm{Geometric}(p = 0.3)$. So $\mu = \mathbb{E}(Y) = \dfrac{q}{p} = \dfrac{0.7}{0.3} = 2.33$.

    $\sigma^2 = \mathit{Var}(Y) = \dfrac{q}{p^2} = \dfrac{0.7}{(0.3)^2} = 7.78$.

    $$\mathit{Var}(Z_{10}) = \sigma^2 \mu^9 \left(\dfrac{1 - \mu^{10}}{1 - \mu}\right) = 5.72 \times 10^7.$$

    Compares well with the sample variance from 5000 simulations, 5.39×10^7 *(page 116).*

2. Family size $Y \sim \mathrm{Geometric}(p = 0.5)$. So $\mu = \mathbb{E}(Y) = \dfrac{q}{p} = \dfrac{0.5}{0.5} = 1$.

    $\sigma^2 = \mathit{Var}(Y) = \dfrac{q}{p^2} = \dfrac{0.5}{(0.5)^2} = 2$. *Using the formula for* $\mathit{Var}(Z_n)$ *when* $\mu = 1$, *we have:*

    $$\mathit{Var}(Z_{10}) = \sigma^2 n = 2 \times 10 = 20.$$

    Compares well with the sample variance of 19.5 (page 117).

Chapter 7:
Extinction in Branching Processes

Extinction is Forever
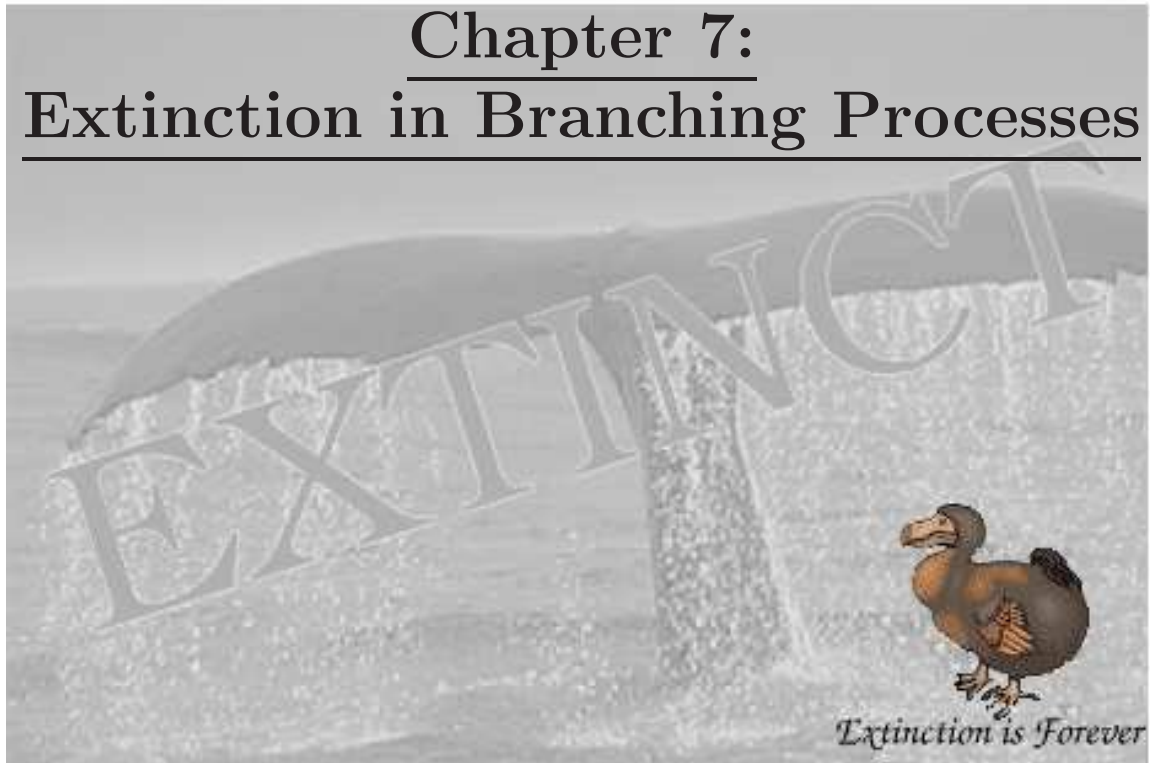
Revision: a branching process consists of reproducing individuals.

- All individuals are independent.
- Start with a single individual at time 0: $Z_0 = 1$.
- Each individual lives a single unit of time, then has Y offspring and dies.
- Let Z_n be the si<u>Z</u>e of generation n: the number of individuals born at time n.
- The branching process is $\{Z_0 = 1, Z_1, Z_2, \ldots\}$.

Branching Process Recursion Formula

This is the fundamental formula for branching processes. Let $G_n(s) = \mathbb{E}(s^{Z_n})$ be the PGF of Z_n, the population size at time n. Let $G(s) = G_1(s)$, the PGF of the family size distribution Y, or equivalently, of Z_1. Then:

$$G_n(s) = G_{n-1}\big(G(s)\big) = \underbrace{G\Big(G\big(G\big(\ldots G(s)\ldots\big)\big)\Big)}_{n \text{ times}} = G\big(G_{n-1}(s)\big).$$

7.1 Extinction Probability

One of the most interesting applications of branching processes is calculating the probability of eventual extinction. For example, what is the probability that a colony of cancerous cells becomes extinct before it overgrows the surrounding tissue? What is the probability that an infectious disease dies out before reaching an epidemic? What is the probability that a family line (e.g. for royal families) becomes extinct?

It is possible to find several results about the probability of eventual extinction.

Extinction by generation n

The population is extinct by generation n if $Z_n = 0$
(no individuals at time n).

If $Z_n = 0$, then *the population is extinct
for ever: $Z_t = 0$ for all $t \geq n$.*

Extinction is Forever

Definition: Define event E_n to be the event
$E_n = \{Z_n = 0\}$ *(event that the population is extinct by generation n).*

Note: $E_0 \subseteq E_1 \subseteq E_2 \subseteq E_3 \subseteq E_4 \subseteq \dots$
This is because event E_i forces E_j to be true for all $j \geq i$, so E_i is a 'part' or subset of E_j for $j \geq i$.

Ultimate extinction

At the start of the branching process, we are interested in the probability of *ultimate extinction:* **the probability that the population will be extinct by generation n, for <u>any</u> value of n.**

We can express this probability in different ways:

$$\mathbb{P}(\text{ultimate extinction}) = \mathbb{P}\left(\bigcup_{n=0}^{\infty} E_n\right) \quad \begin{pmatrix} \textit{i.e. extinct by generation 0 \underline{or}} \\ \textit{extinct by generation 1 \underline{or}} \\ \textit{extinct by generation 2 \underline{or}\dots} \end{pmatrix}$$

Or: $\mathbb{P}(\text{ultimate extinction}) = \mathbb{P}\left(\lim_{n\to\infty} E_n\right).$ *(i.e. $\mathbb{P}$(extinct by generation ∞)).*

Note: By the Continuity Theorem (Chapter 2), and because $E_0 \subseteq E_1 \subseteq E_2 \subseteq \ldots$, we have:

$$\mathbb{P}(\text{ultimate extinction}) = \mathbb{P}\left(\lim_{n\to\infty} E_n\right) = \lim_{n\to\infty} \mathbb{P}(E_n).$$

Thus the probability of eventual extinction is the limit as $n \to \infty$ of the probability of extinction by generation n.

We will use the Greek letter Gamma (γ) for the probability of extinction: think of <u>G</u>amma for 'all <u>G</u>one'!

$$\gamma_n = \mathbb{P}(E_n) \;=\; \mathbb{P}(\textbf{\textit{extinct by generation }} n).$$

$$\gamma \;=\; \mathbb{P}(\textbf{\textit{ultimate extinction}}).$$

By the Note above, we have established that we are looking for:

$$\boxed{\mathbb{P}(\textbf{\textit{ultimate extinction}}) = \gamma = \lim_{n\to\infty} \gamma_n.}$$

γ

Extinction is Forever

<u>**Theorem 7.1:**</u> Let γ be the probability of ultimate extinction. Then

γ is the smallest non-negative solution of the equation $G(s) = s$, where $G(s)$ is the PGF of the family size distribution, Y.

To find the probability of ultimate extinction, we therefore:

- *find the PGF of family size, Y: $G(s) = \mathbb{E}(s^Y)$;*

- *find values of s that satisfy $G(s) = s$;*

- *find the <u>smallest</u> of these values that is ≥ 0. **This is the required value** γ.*

$$\boxed{G(\gamma) = \gamma, \text{ and } \gamma \text{ is the smallest value} \geq 0 \text{ for which this holds.}}$$

Note: Recall that, for any (non-defective) random variable Y with PGF $G(s)$,

$$G(1) = \mathbb{E}(1^Y) = \sum_y 1^y \mathbb{P}(Y = y) = \sum_y \mathbb{P}(Y = y) = 1.$$

So $G(1) = 1$ always, and therefore ***there always exists a solution for $G(s) = s$ in $[0, 1]$.***

The required value γ is the smallest such solution ≥ 0.

Before proving Theorem 7.1 we prove the following Lemma.

Lemma: Let $\gamma_n = \mathbb{P}(Z_n = 0)$.   Then $\gamma_n = G(\gamma_{n-1})$.

Proof: *If $G_n(s)$ is the PGF of Z_n, then $\mathbb{P}(Z_n = 0) = G_n(0)$. (Chapter 4.)*

So $\gamma_n = G_n(0)$. Similarly, $\gamma_{n-1} = G_{n-1}(0)$.

Now $G_n(0) = \underbrace{G\Big(G\Big(G\Big(\ldots G(0)\ldots\Big)\Big)\Big)}_{n \text{ times}} = G\Big(G_{n-1}(0)\Big).$

So $\gamma_n = G\Big(G_{n-1}(0)\Big) = G\Big(\gamma_{n-1}\Big).$ □

Proof of Theorem 7.1: We need to prove:

(i) $G(\gamma) = \gamma$;
(ii) γ is the smallest non-negative value for which $G(\gamma) = \gamma$.
That is, if $s \geq 0$ and $G(s) = s$, then $\gamma \leq s$.

Proof of (i):

From ❀ overleaf, $\gamma = \lim_{n \to \infty} \gamma_n = \lim_{n \to \infty} G\Big(\gamma_{n-1}\Big)$ *(by Lemma)*

$$= G\Big(\lim_{n \to \infty} \gamma_{n-1}\Big) \qquad \text{(G is continuous)}$$

$$= G(\gamma).$$

So $G(\gamma) = \gamma$, as required.

Proof of (ii):

First note that $G(s)$ is an increasing function on $[0, 1]$:

$$G(s) = \mathbb{E}(s^Y) \ = \ \sum_{y=0}^{\infty} s^y \mathbb{P}(Y = y)$$

$$\Rightarrow \quad G'(s) \ = \ \sum_{y=0}^{\infty} y s^{y-1} \mathbb{P}(Y = y)$$

$$\Rightarrow \quad G'(s) \ \geq \ 0 \quad \text{for } 0 \leq s \leq 1, \quad \text{so } G \text{ is increasing on } [0, 1].$$

$G(s)$ is increasing on $[0, 1]$ means that:

$$s_1 \leq s_2 \quad \Rightarrow \quad G(s_1) \leq G(s_2) \quad \text{for any } s_1, s_2 \in [0, 1]. \qquad \clubsuit$$

The branching process begins with $Z_0 = 1$, so

$$\mathbb{P}(\text{extinct by generation } 0) = \gamma_0 = 0.$$

At any later generation, $\gamma_n = G(\gamma_{n-1})$ by Lemma.

Now suppose that $s \geq 0$ and $G(s) = s$. Then we have:

$$0 \leq s \quad \Rightarrow \quad \gamma_0 \leq s \qquad (\text{because } \gamma_0 = 0)$$

$$\Rightarrow \quad G(\gamma_0) \leq G(s) \qquad (\text{by } \clubsuit)$$

$$\text{i.e.} \qquad \gamma_1 \leq s$$

$$\Rightarrow \quad G(\gamma_1) \leq G(s) \qquad (\text{by } \clubsuit)$$

$$\text{i.e.} \qquad \gamma_2 \leq s$$

$$\vdots$$

Thus $$\gamma_n \leq s \qquad \text{for all } n.$$

So if $s \geq 0$ and $G(s) = s$, then $\gamma = \lim_{n \to \infty} \gamma_n \leq s.$ $\qquad\square$

Example 1: Let $\{Z_0 = 1, Z_1, Z_2, \ldots\}$ be a branching process with family size distribution $Y \sim \text{Binomial}(2, \frac{1}{4})$. Find the probability that the process will eventually die out.

Solution:

Let $G(s) = \mathbb{E}(s^Y)$. *The probability of ultimate extinction is* γ, *where* γ *is the smallest solution* ≥ 0 *to the equation* $G(s) = s$.

For $Y \sim \text{Binomial}(n, p)$, *the PGF is* $G(s) = (ps + q)^n$ *(Chapter 4).*

So if $Y \sim \text{Binomial}(2, \frac{1}{4})$ *then* $G(s) = (\frac{1}{4}s + \frac{3}{4})^2$.

We need to solve $G(s) = s$:

$$G(s) = (\tfrac{1}{4}s + \tfrac{3}{4})^2 = s$$

$$\tfrac{1}{16}s^2 + \tfrac{6}{16}s + \tfrac{9}{16} = s$$

$$\tfrac{1}{16}s^2 - \tfrac{10}{16}s + \tfrac{9}{16} = 0$$
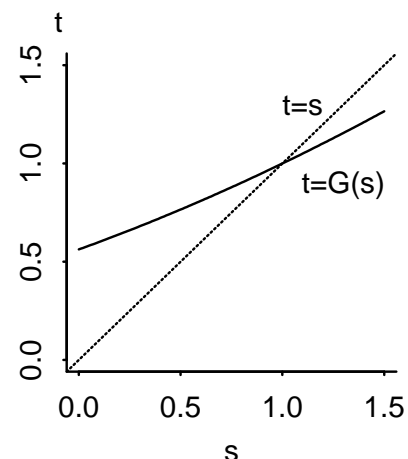
Trick: we know that $G(1) = 1$, *so* $s = 1$ *has got to be a solution. Use this for a quick factorization.*

$$(s - 1)\left(\tfrac{1}{16}s - \tfrac{9}{16}\right) = 0.$$

Thus

$$s = 1$$

or

$$\tfrac{1}{16}s = \tfrac{9}{16} \quad \Rightarrow \quad s = 9.$$

The smallest solution ≥ 0 *is* $s = 1$.

Thus the probability of ultimate extinction is $\gamma = 1$.

Extinction is <u>*definite*</u> *when the family size distribution is* $Y \sim \text{Binomial}(2, \frac{1}{4})$.

Example 2: Let $\{Z_0 = 1, Z_1, Z_2, \ldots\}$ be a branching process with family size distribution $Y \sim$ Geometric$(\frac{1}{4})$. Find the probability that the process will eventually die out.

Solution:

Let $G(s) = \mathbb{E}(s^Y)$. Then $\mathbb{P}(\text{ultimate extinction}) = \gamma$, where γ is the smallest solution ≥ 0 to the equation $G(s) = s$.

For $Y \sim$ Geometric(p), the PGF is $G(s) = \frac{p}{1-qs}$ (Chapter 4).

So if $Y \sim$ Geometric$(\frac{1}{4})$ then $G(s) = \dfrac{1/4}{1 - (3/4)s} = \dfrac{1}{4 - 3s}$.

We need to solve $G(s) = s$:

$$G(s) = \tfrac{1}{4-3s} = s$$

$$4s - 3s^2 = 1$$

$$3s^2 - 4s + 1 = 0$$

Trick: know that $s = 1$ is a solution.

$$(s-1)(3s-1) = 0.$$

Thus

$$s = 1$$

or

$$3s = 1 \quad \Rightarrow \quad s = \tfrac{1}{3}.$$

The smallest solution ≥ 0 is $s = \frac{1}{3}$.

Thus the probability of ultimate extinction is $\gamma = \frac{1}{3}$.

Extinction is possible but not definite when the family size distribution is $Y \sim$ Geometric$(\frac{1}{4})$.
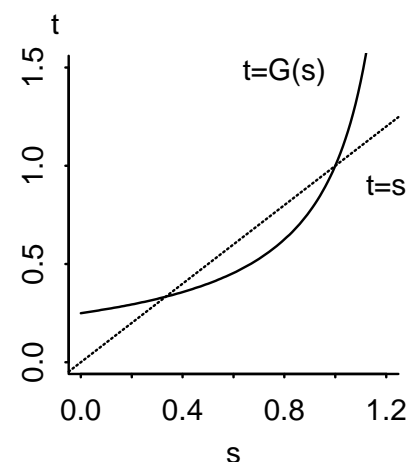
7.2 Conditions for ultimate extinction

It turns out that the probability of extinction depends crucially on the value of μ, *the mean of the family size distribution Y.*
Some values of μ *guarantee* that the branching process will die out with probability 1. Other values guarantee that the probability of extinction will be strictly less than 1. We will see below that the threshold value is $\mu = 1$.

If the mean number of offspring per individual μ is more than 1 (so on average, individuals replace themselves plus a bit extra), then the branching process is not *guaranteed* to die out — although it might do. However, if the mean number of offspring per individual μ is 1 or less, the process is *guaranteed* to become extinct (unless $Y = 1$ with probability 1). The result is not too surprising for $\mu > 1$ or $\mu < 1$, but it is a little surprising that extinction is generally guaranteed if $\mu = 1$.

Theorem 7.2: Let $\{Z_0 = 1, Z_1, Z_2, \ldots\}$ be a branching process with family size distribution Y. Let $\mu = \mathbb{E}(Y)$ be the mean family size distribution, and let γ be the probability of ultimate extinction. Then

(i) *If $\mu > 1$, then $\gamma < 1$: extinction is <u>not</u> guaranteed if $\mu > 1$.*

(ii) *If $\mu < 1$, then $\gamma = 1$: extinction <u>is</u> guaranteed if $\mu < 1$.*

(iii) *If $\mu = 1$, then $\gamma = 1$ <u>unless</u> the family size is always constant at $Y = 1$.*

Lemma: Let $G(s)$ be the PGF of family size Y. Then $G(s)$ and $G'(s)$ are strictly increasing for $0 < s < 1$, as long as Y can take values ≥ 2.

Proof: $G(s) = \mathbb{E}(s^Y) = \sum_{y=0}^{\infty} s^y \mathbb{P}(Y = y)$.

So $G'(s) = \sum_{y=1}^{\infty} y s^{y-1} \mathbb{P}(Y = y) > 0$ for $0 < s < 1$,

because all terms are ≥ 0 and at least 1 term is > 0 (if $\mathbb{P}(Y \geq 2) > 0$).

Similarly, $G''(s) = \sum_{y=2}^{\infty} y(y-1)s^{y-2}\mathbb{P}(Y = y) > 0$ for $0 < s < 1$.

So $G(s)$ and $G'(s)$ are strictly increasing for $0 < s < 1$. $\square$

***Note:*** When $G''(s) > 0$ for $0 < s < 1$, the function G is said to be **convex** on that interval.

Convex: G''(s) > 0   Concave: G''(s) < 0
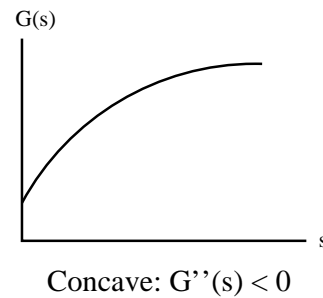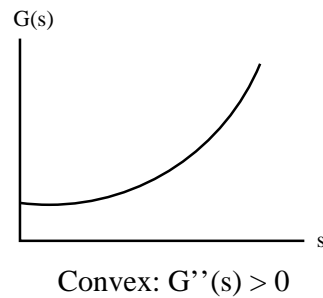
$G''(s) > 0$ means that *the gradient of G is constantly increasing for $0 < s < 1$.*

Proof of Theorem 7.2: This is usually done graphically.

The graph of $G(s)$ satisfies the following conditions:

1. *$G(s)$ is increasing and strictly convex (as long as Y can be ≥ 2).*

2. *$G(0) = \mathbb{P}(Y = 0) \geq 0$.*

3. *$G(1) = 1$.*

4. *$G'(1) = \mu$, so the slope of $G(s)$ at $s = 1$ gives the value μ.*

5. *The extinction probability γ is the smallest value ≥ 0 for which $G(s) = s$.*
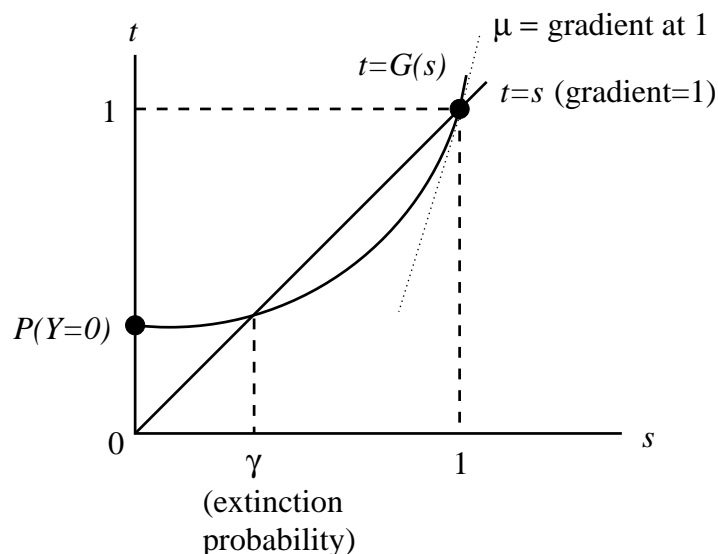
Case (i): $\mu > 1$

When $\mu > 1$, the curve $G(s)$ is forced beneath the line $t = s$ at $s = 1$. The curve $G(s)$ has to cross the line $t = s$ again to meet the t-axis at $\mathbb{P}(Y = 0)$.
Thus there must be a solution $\gamma < 1$ to the equation $G(s) = s$.
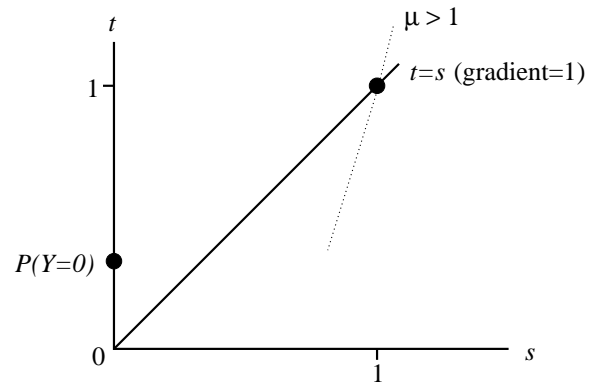
Case (ii): $\mu < 1$

When $\mu < 1$, the curve $G(s)$ is forced above the line $t = s$ for $s < 1$. There is no possibility for the curve $G(s)$ to cross the line $t = s$ again before meeting the t-axis.
Thus there can be no solution < 1 to the equation $G(s) = s$, so $\gamma = 1$.

The exception is where Y can take only values 0 and 1, so $G(s)$ is not strictly convex (see Lemma). However, in that case $G(s) = p_0 + p_1 s$ is a straight line, giving the same result $\gamma = 1$.
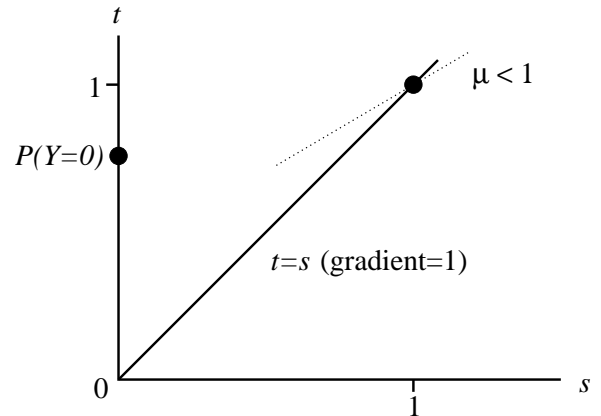
Case (iii): $\mu = 1$

When $\mu = 1$, the situation is the same as for $\mu < 1$.

The exception is where Y takes only the value 1. Then $G(s) = s$ for all $0 \leq s \leq 1$, so the smallest solution ≥ 0 is $\gamma = 0$.

Thus extinction is guaranteed for $\mu = 1$, **unless** $Y = 1$ with probability 1.
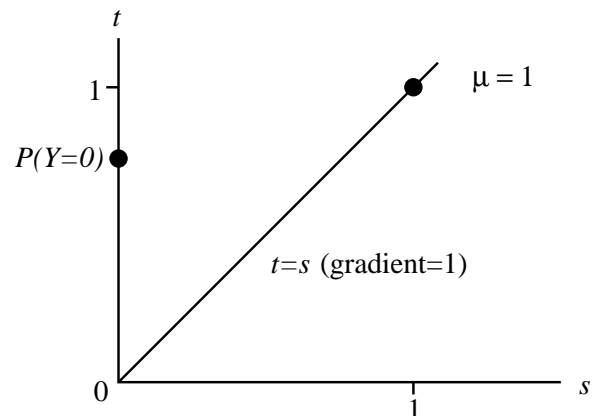
Example 1: Let $\{Z_0 = 1, Z_1, Z_2, \ldots\}$ be a branching process with family size distribution $Y \sim \text{Binomial}(2, \frac{1}{4})$, as in Section 7.1. Find the probability of eventual extinction.

Solution:

Consider $Y \sim \text{Binomial}(2, \frac{1}{4}.)$ *The mean of Y is $\mu = 2 \times \frac{1}{4} = \frac{1}{2} < 1$. Thus, by Theorem 7.2,*

$$\gamma = \mathbb{P}(\text{ultimate extinction}) = 1.$$

(The longer calculation in Section 7.1 was not necessary.)

Example 2: Let $\{Z_0 = 1, Z_1, Z_2, \ldots\}$ be a branching process with family size distribution $Y \sim \text{Geometric}(\frac{1}{4})$, as in Section 7.1. Find the probability of eventual extinction.

Solution:

Consider $Y \sim \text{Geometric}(\frac{1}{4}.)$ *The mean of Y is $\mu = \frac{1-1/4}{1/4} = 3 > 1$. Thus, by Theorem 7.2,*

$$\gamma = \mathbb{P}(\text{ultimate extinction}) < 1.$$

To find the value of γ, we still need to go through the calculation presented in Section 7.1. (Answer: $\gamma = \frac{1}{3}$.)

Note: The mean μ of the offspring distribution Y is known as the ***criticality parameter.***

- If $\mu < 1$, extinction is definite ($\gamma = 1$). The process is called <u>subcritical</u>. Note that $\mathbb{E}(Z_n) = \mu^n \to 0$ as $n \to \infty$.

- If $\mu = 1$, extinction is definite unless $Y \equiv 1$. The process is called <u>critical</u>. Note that $\mathbb{E}(Z_n) = \mu^n = 1 \; \forall n$, even though extinction is definite.

- If $\mu > 1$, extinction is not definite ($\gamma < 1$). The process is called <u>supercritical</u>. Note that $\mathbb{E}(Z_n) = \mu^n \to \infty$ as $n \to \infty$.

BAD Luck ! your population is SCHEDULED to become Extinct!

7.3 Time to Extinction

Suppose the population is doomed to extinction — or maybe it isn't. Either way, it is useful to know how long it will take for the population to become extinct. This is the distribution of T, the number of generations before extinction. For example, how long do we expect a disease epidemic like SARS to continue? How long have we got to organize ourselves to save the kakapo or the tuatara before they become extinct before our very eyes?

1. Extinction (by) time n

The branching process is extinct **by** time n if $Z_n = 0$.

Thus the probability that the process has become extinct **by** time n is:

$$\mathbb{P}(Z_n = 0) = G_n(0) = \gamma_n.$$

Note: Recall that $G_n(s) = \mathbb{E}(s^{Z_n}) = \underbrace{G\Big(G\big(G\big(\ldots G(s)\ldots\big)\big)\Big)}_{n \text{ times}}.$

There is no guarantee that the PGF $G_n(s)$ or the value $G_n(0)$ can be calculated easily. However, we can build up $G_n(0)$ in steps:

e.g. $G_2(0) = G(G(0))$*; then* $G_3(0) = G(G_2(0))$*, or even* $G_4(0) = G_2(G_2(0))$*.*

2. Extinction (at) time n

Let T be the **exact time of extinction**. That is, $T = n$ if generation n is the first generation with no individuals:

$$T = n \iff Z_n = 0 \quad AND \quad Z_{n-1} > 0.$$

Now by the Partition Rule,

$$\mathbb{P}(Z_n = 0 \cap Z_{n-1} > 0) + \mathbb{P}(Z_n = 0 \cap Z_{n-1} = 0) = \mathbb{P}(Z_n = 0). \qquad (\star)$$

But the event $\{Z_n = 0 \cap Z_{n-1} = 0\}$ is the event that the process is extinct by generation $n - 1$ AND it is extinct by generation n. However, we know it will always be extinct by generation n if it is extinct by generation $n - 1$, so the $Z_n = 0$ part is redundant. So

$$\mathbb{P}(Z_n = 0 \cap Z_{n-1} = 0) = \mathbb{P}(Z_{n-1} = 0) = G_{n-1}(0).$$

Similarly,

$$\mathbb{P}(Z_n = 0) = G_n(0).$$

So $(\star)$ gives:

$$\boxed{\mathbb{P}(T = n) = \mathbb{P}(Z_n = 0 \cap Z_{n-1} > 0) = G_n(0) - G_{n-1}(0) = \gamma_n - \gamma_{n-1}.}$$

This gives the distribution of T, the exact time at which extinction occurs.

Example: Binary splitting. Suppose that the family size distribution is

$$Y = \begin{cases} 0 & \text{with probability } q = 1 - p, \\ 1 & \text{with probability } p. \end{cases}$$

Find the distribution of the time to extinction.

Solution:

Consider

$$G(s) = \mathbb{E}(s^Y) = qs^0 + ps^1 \;=\; q + ps.$$

$$G_2(s) = G\Big(G(s)\Big) = q + p(q + ps) \;=\; q(1 + p) + p^2 s.$$

$$G_3(s) = G\Big(G_2(s)\Big) = q + p(q + pq + p^2 s) \;=\; q(1 + p + p^2) + p^3 s.$$

$$\vdots$$

$$G_n(s) \;=\; q(1 + p + p^2 + \ldots + p^{n-1}) + p^n s.$$

Thus time to extinction, T, satisfies

$$\mathbb{P}(T = n) \;=\; G_n(0) - G_{n-1}(0)$$

$$\;=\; q(1 + p + p^2 + \ldots + p^{n-1}) - q(1 + p + p^2 + \ldots + p^{n-2})$$

$$\;=\; qp^{n-1} \qquad \text{for } n = 1, 2, \ldots$$

Thus

$$T - 1 \sim \text{Geometric}(q).$$

It follows that $\mathbb{E}(T - 1) = \frac{p}{q}$, so

$$\mathbb{E}(T) = 1 + \frac{p}{q} = \frac{1 - p + p}{q} = \frac{1}{q}.$$

Note: The **expected time to extinction**, $\mathbb{E}(T)$, is:

- *finite* if $\mu < 1$;

- *infinite* if $\mu = 1$ *(despite extinction being definite), if σ^2 is finite;*

- *infinite* if $\mu > 1$ *(because with positive probability, extinction never happens).*

(Results not proved here.)

7.4 Case Study: Geometric Branching Processes

Recall that $G_n(s) = \mathbb{E}(s^{Z_n}) = \underbrace{G\Big(G\Big(G\Big(\ldots G(s)\ldots\Big)\Big)\Big)}_{n \text{ times}}.$

In general, it is not possible to find a closed-form expression for $G_n(s)$. We achieved a closed-form $G_n(s)$ in the Binary Splitting example (page 134), but binary splitting only allows family size Y to be 0 or 1, which is a very restrictive model.

The only non-trivial family size distribution that allows us to find a closed-form expression for $G_n(s)$ is the **Geometric distribution.**

When family size $Y \sim \text{Geometric}(p)$, we can do the following:

- Derive a closed-form expression for $G_n(s)$, the PGF of Z_n.

- Find the probability distribution of the **exact time of extinction, T:** not just the probability that extinction will occur at some unspecified time (γ).

- Find the **full probability distribution of Z_n:** probabilities $\mathbb{P}(Z_n = 0)$, $\mathbb{P}(Z_n = 1)$, $\mathbb{P}(Z_n = 2)$, ...

With $Y \sim \text{Geometric}(p)$, we can therefore calculate just about every quantity we might be interested in for the branching process.

1. Closed form expression for $G_n(s)$

Theorem 7.4: Let $\{Z_0 = 1, Z_1, Z_2, \ldots\}$ be a branching process with family size distribution $Y \sim \text{Geometric}(p)$. The PGF of Z_n is given by:

$$G_n(s) = \mathbb{E}\left(s^{Z_n}\right) = \begin{cases} \dfrac{n - (n-1)s}{n + 1 - ns} & \text{if } \quad p = q = 0.5, \\[3em] \dfrac{(\mu^n - 1) - \mu(\mu^{n-1} - 1)s}{(\mu^{n+1} - 1) - \mu(\mu^n - 1)s} & \text{if } \quad p \neq q, \text{ where } \mu = \frac{q}{p}. \end{cases}$$

Proof (sketch):

The proof for both $p = q$ and $p \neq q$ proceed by mathematical induction. We will give a sketch of the proof when $p = q = 0.5$. The proof for $p \neq q$ works in the same way but is trickier.

Consider $p = q = \frac{1}{2}$. Then

$$G(s) = \frac{p}{1 - qs} = \frac{\frac{1}{2}}{1 - \frac{s}{2}} = \frac{1}{2 - s}.$$

Using the Branching Process Recursion Formula (Chapter 6),

$$G_2(s) = G\big(G(s)\big) = \frac{1}{2 - G(s)} = \frac{1}{2 - \frac{1}{2-s}} = \frac{2 - s}{2(2 - s) - 1} = \frac{2 - s}{3 - 2s}.$$

The inductive hypothesis is that $G_n(s) = \dfrac{n - (n-1)s}{n + 1 - ns}$, and it holds for $n = 1$ and $n = 2$. Suppose it holds for n. Then

$$G_{n+1}(s) = G_n\big(G(s)\big) = \frac{n - (n-1)G(s)}{n + 1 - nG(s)} = \frac{n - (n-1)\left(\frac{1}{2-s}\right)}{n + 1 - n\left(\frac{1}{2-s}\right)}$$

$$= \frac{(2 - s)n - (n - 1)}{(2 - s)(n + 1) - n}$$

$$= \frac{n + 1 - ns}{n + 2 - (n+1)s}.$$

Therefore, if the hypothesis holds for n, it also holds for $n + 1$. Thus the hypothesis is proved for all n. $\qquad\square$

---

2. Exact time of extinction, T

Let $Y \sim \text{Geometric}(p)$, and let T be the exact generation of extinction.

From Section 7.3,

$$\mathbb{P}(T = n) = \mathbb{P}(Z_n = 0) - \mathbb{P}(Z_{n-1} = 0) = G_n(0) - G_{n-1}(0).$$

By using the closed-form expressions overleaf for $G_n(0)$ and $G_{n-1}(0)$, we can find $\mathbb{P}(T = n)$ for any n.

3. Whole distribution of Z_n

From Chapter 4, $\mathbb{P}(Z_n = r) = \dfrac{1}{r!}G_n^{(r)}(0)$.

Now our closed-form expression for $G_n(s)$ has the same format regardless of whether $\mu = 1$ ($p = 0.5$), or $\mu \neq 1$ ($p \neq 0.5$):

$$G_n(s) = \frac{A - Bs}{C - Ds}\,.$$

(For example, when $\mu = 1$, we have $A = D = n$, $B = n - 1$, $C = n + 1$.) Thus:

$$\mathbb{P}(Z_n = 0) = G_n(0) = \frac{A}{C}$$

$$G_n'(s) = \frac{(C - Ds)(-B) + (A - Bs)D}{(C - Ds)^2} = \frac{AD - BC}{(C - Ds)^2}$$

$$\Rightarrow \quad \mathbb{P}(Z_n = 1) = \frac{1}{1!}G_n'(0) = \frac{AD - BC}{C^2}$$

$$G_n''(s) = \frac{(-2)(-D)(AD - BC)}{(C - Ds)^3} = \frac{2D(AD - BC)}{(C - Ds)^3}$$

$$\Rightarrow \quad \mathbb{P}(Z_n = 2) = \frac{1}{2!}G_n''(0) = \left(\frac{AD - BC}{CD}\right)\left(\frac{D}{C}\right)^2$$

$$\vdots$$

$$\Rightarrow \quad \mathbb{P}(Z_n = r) = \frac{1}{r!}G_n^{(r)}(0) = \left(\frac{AD - BC}{CD}\right)\left(\frac{D}{C}\right)^r \quad \text{for } r = 1, 2, \ldots$$

(Exercise)

This is very simple and powerful: we can substitute the values of A, B, C, and D to find $\mathbb{P}(Z_n = r)$ or $\mathbb{P}(Z_n \leq r)$ for any r and n.

Note: A Java applet that simulates branching processes can be found at:
`http://www.dartmouth.edu/~chance/teaching_aids/books_articles/`
`probability_book/bookapplets/chapter10/Branch/Branch.html`

Chapter 8: Markov Chains

it only matters where you are, not where you've been...

8.1 Introduction

So far, we have looked at stochastic processes such as the
random walk, $\{X_0, X_1, X_2, \ldots\}$ where X_t is the position
at time t, and the branching process, $\{Z_0, Z_1, Z_2, \ldots\}$,
where Z_t is the number of individuals in generation t.
We have also examined many stochastic processes
using First-Step Analysis on the transition diagrams.

A.A.Markov
1856-1922

Most of these processes have one thing in common:
X_{t+1} *depends only on* X_t.
It does <u>not</u> depend upon $X_0, X_1, \ldots, X_{t-1}$.

Example: Random Walk.

none of these steps matter for time t+1

? time t+1

time t

?
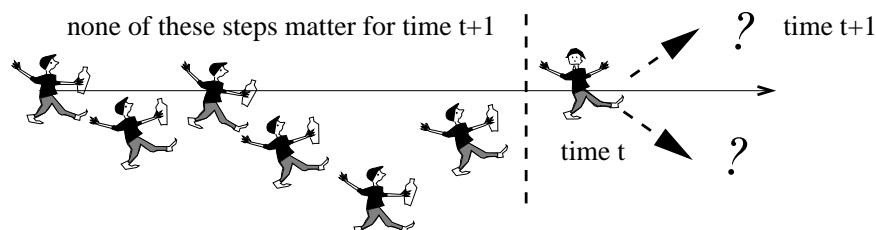
Example: Branching Process.

none of these
steps matter
for time t+1

time t

Processes like this are called
Markov Chains.

? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?

time t+1

In a Markov chain, the

future depends only

upon the present.

Meet... _the_ _Markov fleas!!_

Glee-flea   _Forget-flea_   _Purpose-flea_   _Skill-flea_
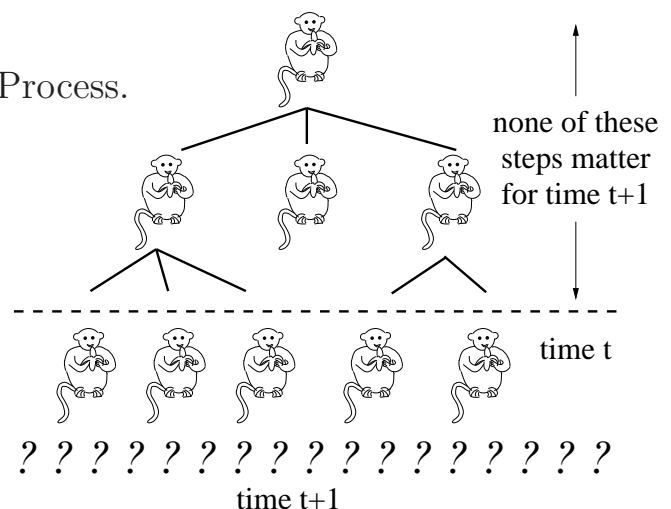
The text-book image
of a Markov chain has
a flea hopping about at
random on the vertices
of the transition diagram,
according to the probabilities shown.
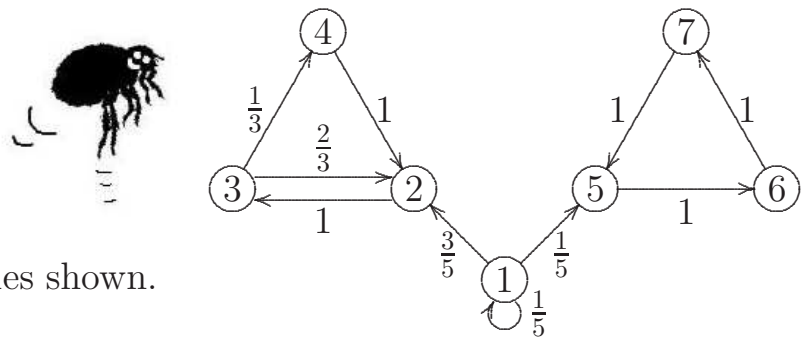
The transition diagram above shows a system with 7 possible states:

$$\text{state space } S = \{1, 2, 3, 4, 5, 6, 7\}.$$

Questions of interest

- Starting from state 1, what is the probability of ever reaching state 7?
- Starting from state 2, what is the expected time taken to reach state 4?
- Starting from state 2, what is the long-run proportion of time spent in state 3?
- Starting from state 1, what is the probability of being in state 2 at time t? Does the probability converge as $t \to \infty$, and if so, to what?

We have been answering questions like the first two using first-step analysis since the start of STATS 325. In this chapter we develop a unified approach to all these questions using the matrix of transition probabilities, called the **transition matrix**.

8.2 Definitions

The Markov chain is the process $X_0, X_1, X_2, \ldots$.

Definition: The **state** of a Markov chain at time t is the *value of X_t.*

For example, if $X_t = 6$, we say *the process is in state* 6 *at time* t.

Definition: The **state space** of a Markov chain, S, is the set of values that each X_t can take. For example, $S = \{1, 2, 3, 4, 5, 6, 7\}$.

Let S have size N (possibly infinite).

Definition: A **trajectory** of a Markov chain is *a particular set of values for* $X_0, X_1, X_2, \ldots$.

For example, if $X_0 = 1$, $X_1 = 5$, and $X_2 = 6$, then the trajectory up to time $t = 2$ is $1, 5, 6$.

More generally, if we refer to the trajectory $s_0, s_1, s_2, s_3, \ldots$, we mean that $X_0 = s_0$, $X_1 = s_1$, $X_2 = s_2$, $X_3 = s_3$, $\ldots$

'Trajectory' is just a word meaning *'path'.*

Markov Property

The basic property of a Markov chain is that *only the most recent point in the trajectory affects what happens next.*

This is called the *Markov Property.*
It means that X_{t+1} *depends upon X_t, but it does not depend upon* $X_{t-1}, \ldots, X_1, X_0$.

We formulate the Markov Property in mathematical notation as follows:

$$\mathbb{P}(X_{t+1} = s \,|\, X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) \;=\; \mathbb{P}(X_{t+1} = s \,|\, X_t = s_t),$$

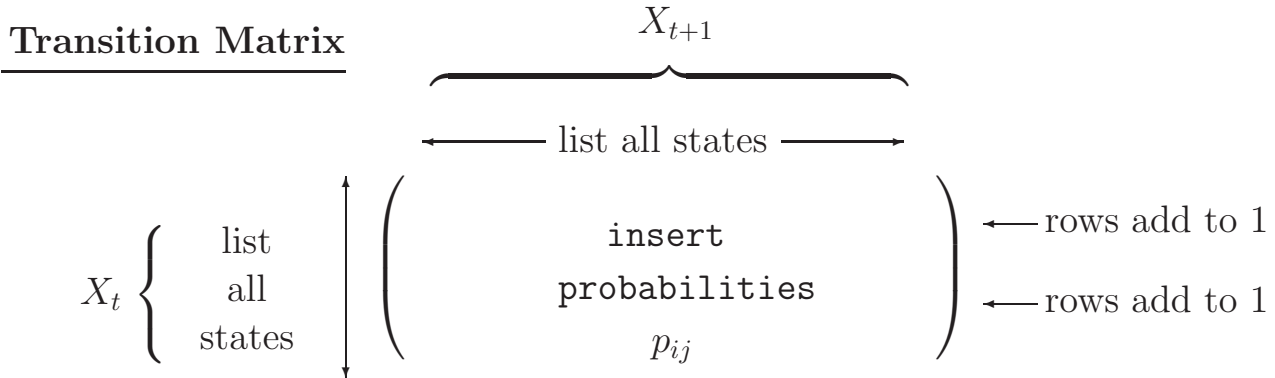for all $t = 1, 2, 3, \ldots$ and for all states $s_0, s_1, \ldots, s_t, s$.

Explanation:

$$\mathbb{P}(X_{t+1} = s \;\;|\;\; X_t = s_t, \;\; \overbrace{X_{t-1} = s_{t-1}, X_{t-2} = s_{t-2}, \ldots, X_1 = s_1, X_0 = s_0})$$

$\uparrow$

distribution $\uparrow$

    of X_{t+1}   *depends* $\uparrow$

         on X_t   *but whatever happened before time t*

doesn't matter.

Definition: Let $\{X_0, X_1, X_2, \ldots\}$ be a sequence of discrete random variables. Then $\{X_0, X_1, X_2, \ldots\}$ is a **Markov chain** if *it satisfies the Markov property:*

$$\mathbb{P}(X_{t+1} = s \,|\, X_t = s_t, \ldots, X_0 = s_0) \;=\; \mathbb{P}(X_{t+1} = s \,|\, X_t = s_t),$$

for all $t = 1, 2, 3, \ldots$ and for all states $s_0, s_1, \ldots, s_t, s$.

8.3 The Transition Matrix

We have seen many examples of **transition diagrams** to describe Markov chains. The transition diagram is so-called because it shows the <u>transitions</u> between different states.
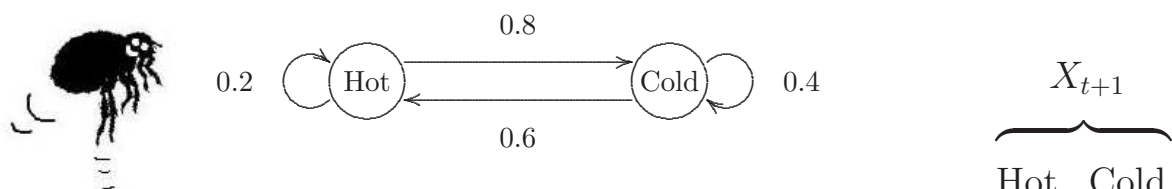
We can also summarize the probabilities in a **matrix:**

$$X_t \left\{ \begin{array}{c} \text{Hot} \\ \text{Cold} \end{array} \overbrace{\left(\begin{array}{cc} 0.2 & 0.8 \\ 0.6 & 0.4 \end{array} \right)}^{\begin{array}{cc} \text{Hot} & \text{Cold} \end{array}} \right.$$

The matrix describing the Markov chain is called the ***transition matrix.***
It is the most important tool for analysing Markov chains.

Transition Matrix

$$X_{t+1}$$

$$\overbrace{\longleftarrow \text{list all states} \longrightarrow}$$

$$X_t \left\{ \begin{array}{c} \text{list} \\ \text{all} \\ \text{states} \end{array} \right. \Bigg\downarrow \left(\begin{array}{c} \texttt{insert} \\ \texttt{probabilities} \\ p_{ij} \end{array} \right) \begin{array}{l} \longleftarrow \text{rows add to 1} \\ \\ \longleftarrow \text{rows add to 1} \end{array}$$

The transition matrix is usually given the symbol $P = (p_{ij})$.

In the transition matrix P:

- *the R<u>OWS</u> represent N<u>OW</u>, or FROM (X_t);*

- *the COLUMNS represent NEXT, or TO (X_{t+1});*

- *entry (i, j) is the CONDITIONAL probability that NEXT $= j$, given that NOW $= i$: the probability of going FROM state i TO state j.*

$$\boxed{p_{ij} = \mathbb{P}(X_{t+1} = j \,|\, X_t = i).}$$

Notes: 1. The transition matrix P must list *all* possible states in the state space S.

2. P is a *square matrix* $(N \times N)$, because X_{t+1} and X_t both take values in the same state space S (of size N).

3. The **rows** of P should each *sum to 1:*

$$\sum_{j=1}^{N} p_{ij} = \sum_{j=1}^{N} \mathbb{P}(X_{t+1} = j \,|\, X_t = i) = \sum_{j=1}^{N} \mathbb{P}_{\{X_t = i\}}(X_{t+1} = j) = 1.$$

This simply states that X_{t+1} *must* take one of the listed values.

4. The **columns** of P do **not** in general sum to 1.

Definition: Let $\{X_0, X_1, X_2, \ldots\}$ be a Markov chain with state space S, where S has size N (possibly infinite). The **transition probabilities** of the Markov chain are

$$p_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i) \quad \textit{for } i, j \in S, \quad t = 0, 1, 2, \ldots$$

Definition: The **transition matrix** of the Markov chain is $P = (p_{ij})$.

8.4 Example: setting up the transition matrix

We can create a transition matrix for any of the transition diagrams we have seen in problems throughout the course. For example, check the matrix below.

Example: Tennis game at Deuce.
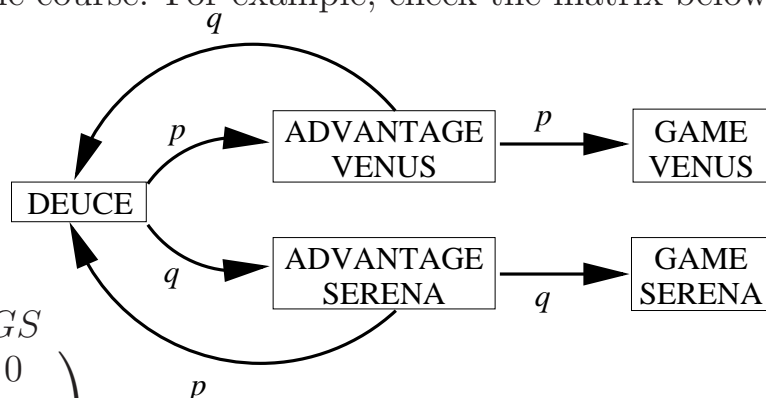
$$
\begin{array}{c c}
 & \begin{array}{ccccc} D & AV & AS & GV & GS \end{array} \\
\begin{array}{c} D \\ AV \\ AS \\ GV \\ GS \end{array} &
\left(\begin{array}{ccccc}
0 & p & q & 0 & 0 \\
q & 0 & 0 & p & 0 \\
p & 0 & 0 & 0 & q \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{array} \right)
\end{array}
$$

8.5 Matrix Revision

Notation

Let A be an $N \times N$ matrix.
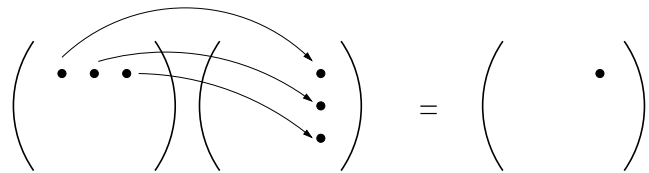
We write $A = (a_{ij})$,
i.e. A comprises elements a_{ij}.

The (i, j) element of A is written both as a_{ij} and $(A)_{ij}$:
e.g. for matrix A^2 we might write $(A^2)_{ij}$.
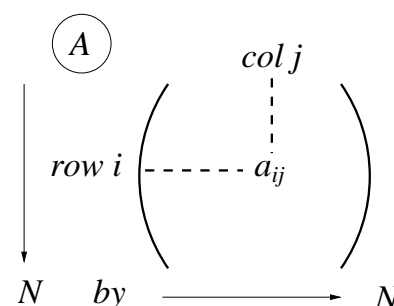
Matrix multiplication

Let $A = (a_{ij})$ and $B = (b_{ij})$
be $N \times N$ matrices.

The product matrix is $A \times B = AB$, with elements $(AB)_{ij} = \sum_{k=1}^{N} a_{ik}b_{kj}$.

Summation notation for a matrix squared

Let A be an $N \times N$ matrix. Then

$$(A^2)_{ij} = \sum_{k=1}^{N}(A)_{ik}(A)_{kj} = \sum_{k=1}^{N} a_{ik}a_{kj}.$$

Pre-multiplication of a matrix by a vector

Let A be an $N \times N$ matrix, and let $\boldsymbol{\pi}$ be an $N \times 1$ column vector: $\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_N \end{pmatrix}$.

We can pre-multiply A by $\boldsymbol{\pi}^T$ to get a $1 \times N$ row vector,
$\boldsymbol{\pi}^T A = \big((\boldsymbol{\pi}^T A)_1, \ldots, (\boldsymbol{\pi}^T A)_N\big)$, with elements

$$(\boldsymbol{\pi}^T A)_j = \sum_{i=1}^{N} \pi_i a_{ij}.$$

8.6 The t-step transition probabilities

Let $\{X_0, X_1, X_2, \ldots\}$ be a Markov chain with state space $S = \{1, 2, \ldots, N\}$.

Recall that the elements of the transition matrix P are defined as:

$$(P)_{ij} = p_{ij} = \mathbb{P}(X_1 = j \mid X_0 = i) = \mathbb{P}(X_{n+1} = j \mid X_n = i) \quad \text{for any } n.$$

p_{ij} is the probability of making a transition FROM state i TO state j in a SINGLE step.

Question: what is the probability of making a transition from state i to state j over **two** steps?   *I.e. what is $\mathbb{P}(X_2 = j \mid X_0 = i)$?*

We are seeking $\mathbb{P}(X_2 = j \mid X_0 = i)$. Use the *Partition Theorem:*

$$
\begin{aligned}
\mathbb{P}(X_2 = j \mid X_0 = i) &= \mathbb{P}_i(X_2 = j) \quad \textit{(notation of Ch 2)} \\[2mm]
&= \sum_{k=1}^{N} \mathbb{P}_i(X_2 = j \mid X_1 = k)\mathbb{P}_i(X_1 = k) \quad \textit{(Partition Thm)} \\[2mm]
&= \sum_{k=1}^{N} \mathbb{P}(X_2 = j \mid X_1 = k,\ X_0 = i)\mathbb{P}(X_1 = k \mid X_0 = i) \\[2mm]
&= \sum_{k=1}^{N} \mathbb{P}(X_2 = j \mid X_1 = k)\mathbb{P}(X_1 = k \mid X_0 = i) \\[1mm]
&\hphantom{= \sum_{k=1}^{N}} \textit{(Markov Property)} \\[2mm]
&= \sum_{k=1}^{N} p_{kj}p_{ik} \quad \textit{(by definitions)} \\[2mm]
&= \sum_{k=1}^{N} p_{ik}p_{kj} \quad \textit{(rearranging)} \\[2mm]
&= (P^2)_{ij}. \quad \textit{(see Matrix Revision)}
\end{aligned}
$$

The two-step transition probabilities are therefore given by *the matrix P^2:*

$$
\boxed{\; \mathbb{P}(X_2 = j \mid X_0 = i) = \mathbb{P}(X_{n+2} = j \mid X_n = i) = \left(P^2\right)_{ij} \quad \textit{for any } n. \;}
$$

3-step transitions: We can find $\mathbb{P}(X_3 = j \mid X_0 = i)$ similarly, but conditioning on the state at time 2:

$$
\begin{aligned}
\mathbb{P}(X_3 = j \mid X_0 = i) &= \sum_{k=1}^{N} \mathbb{P}(X_3 = j \mid X_2 = k)\mathbb{P}(X_2 = k \mid X_0 = i) \\[2mm]
&= \sum_{k=1}^{N} p_{kj}\left(P^2\right)_{ik} \\[2mm]
&= (P^3)_{ij}.
\end{aligned}
$$

The three-step transition probabilities are therefore given by the matrix P^3:

$$\mathbb{P}(X_3 = j \mid X_0 = i) = \mathbb{P}(X_{n+3} = j \mid X_n = i) = \left(P^3\right)_{ij} \quad \text{for any } n.$$

General case: t-step transitions

The above working extends to show that the t-step transition probabilities are given by the matrix P^t for any t:

$$\mathbb{P}(X_t = j \mid X_0 = i) = \mathbb{P}(X_{n+t} = j \mid X_n = i) = \left(P^t\right)_{ij} \quad \textit{for any } n.$$

We have proved the following Theorem.

Theorem 8.6: Let $\{X_0, X_1, X_2, \ldots\}$ be a Markov chain with $N \times N$ transition matrix P. Then the t-step transition probabilities are given by the matrix P^t. That is,

$$\mathbb{P}(X_t = j \mid X_0 = i) = \left(P^t\right)_{ij}.$$

It also follows that

$$\mathbb{P}(X_{n+t} = j \mid X_n = i) = \left(P^t\right)_{ij} \text{ for any } n. \qquad \square$$

8.7 Distribution of X_t

Let $\{X_0, X_1, X_2, \ldots\}$ be a Markov chain with state space $S = \{1, 2, \ldots, N\}$.

Now each X_t is a random variable, so it has a *probability distribution.*

We can write the probability distribution of X_t as an $N \times 1$ *vector.*

For example, consider X_0. Let $\boldsymbol{\pi}$ be an $N \times 1$ vector denoting the probability distribution of X_0:

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{pmatrix} = \begin{pmatrix} \mathbb{P}(X_0 = 1) \\ \mathbb{P}(X_0 = 2) \\ \vdots \\ \mathbb{P}(X_0 = N) \end{pmatrix}$$

In the flea model, this corresponds to *the flea choosing at random which vertex it starts off from at time 0, such that*

$$\mathbb{P}(\textit{flea chooses vertex } i \textit{ to start}) = \pi_i.$$

Notation: we will write $X_0 \sim \boldsymbol{\pi}^T$ to denote that the row vector of probabilities is given by the row vector $\boldsymbol{\pi}^T$.

Probability distribution of X_1

Use the Partition Rule, conditioning on X_0:

$$
\begin{aligned}
\mathbb{P}(X_1 = j) &= \sum_{i=1}^{N} \mathbb{P}(X_1 = j \mid X_0 = i)\mathbb{P}(X_0 = i) \\
&= \sum_{i=1}^{N} p_{ij}\pi_i \quad \textit{by definitions} \\
&= \sum_{i=1}^{N} \pi_i p_{ij} \\
&= \left(\boldsymbol{\pi}^T P\right)_j.
\end{aligned}
$$

(pre-multiplication by a vector from Section 8.5).

This shows that $\mathbb{P}(X_1 = j) = \left(\boldsymbol{\pi}^T P\right)_j$ *for all j.*
The row vector $\boldsymbol{\pi}^T P$ is therefore *the probability distribution of X_1:*

$$
\boxed{
\begin{aligned}
X_0 &\sim \boldsymbol{\pi}^T \\
X_1 &\sim \boldsymbol{\pi}^T P.
\end{aligned}
}
$$

Probability distribution of X_2

Using the Partition Rule as before, conditioning again on X_0:

$$\mathbb{P}(X_2 = j) = \sum_{i=1}^{N} \mathbb{P}(X_2 = j \mid X_0 = i)\mathbb{P}(X_0 = i) = \sum_{i=1}^{N} \left(P^2\right)_{ij} \pi_i = \left(\boldsymbol{\pi}^T P^2\right)_j.$$

The row vector $\boldsymbol{\pi}^T P^2$ is therefore the probability distribution of X_2:

$$
\begin{aligned}
X_0 &\sim \boldsymbol{\pi}^T \\
X_1 &\sim \boldsymbol{\pi}^T P \\
X_2 &\sim \boldsymbol{\pi}^T P^2 \\
&\vdots \\
X_t &\sim \boldsymbol{\pi}^T P^t.
\end{aligned}
$$

These results are summarized in the following Theorem.

Theorem 8.7: Let $\{X_0, X_1, X_2, \ldots\}$ be a Markov chain with $N \times N$ transition matrix P. If the probability distribution of X_0 is given by the $1 \times N$ row vector $\boldsymbol{\pi}^T$, then the probability distribution of X_t is given by the $1 \times N$ row vector $\boldsymbol{\pi}^T P^t$. That is,

$$
X_0 \sim \boldsymbol{\pi}^T \quad \Rightarrow \quad X_t \sim \boldsymbol{\pi}^T P^t.
$$

Note: The distribution of X_t is $X_t \sim \boldsymbol{\pi}^T P^t$.

The distribution of X_{t+1} is $X_{t+1} \sim \boldsymbol{\pi}^T P^{t+1}$.

Taking one step in the Markov chain corresponds to *multiplying by P on the right.*

Note: The t-step transition matrix is P^t *(Theorem 8.6).*

The $(t+1)$-step transition matrix is P^{t+1}.

Again, taking one step in the Markov chain corresponds to *multiplying by P on the right.*

take 1 step...

$\longleftarrow P =$ *...multiply by P
on the right*

8.8 Trajectory Probability

Recall that a trajectory is a sequence of values for $X_0, X_1, \ldots, X_t$.

Because of the Markov Property, we can find the probability of any trajectory by multiplying together the starting probability and all subsequent single-step probabilities.
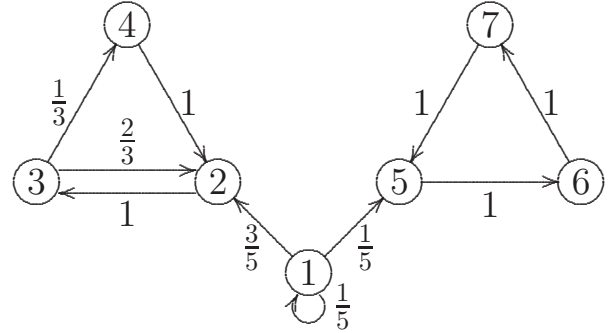
Example: Let $X_0 \sim (\frac{3}{4}, 0, \frac{1}{4}, 0, 0, 0, 0)$. What is the probability of the trajectory 1, 2, 3, 2, 3, 4?

$$
\begin{aligned}
\mathbb{P}(1, 2, 3, 2, 3, 4) &= \mathbb{P}(X_0 = 1) \times p_{12} \times p_{23} \times p_{32} \times p_{23} \times p_{34} \\
&= \tfrac{3}{4} \times \tfrac{3}{5} \times 1 \times \tfrac{2}{3} \times 1 \times \tfrac{1}{3} \\
&= \tfrac{1}{10}.
\end{aligned}
$$

Proof in formal notation using the Markov Property:

Let $X_0 \sim \boldsymbol{\pi}^T$. We wish to find the probability of the trajectory $s_0, s_1, s_2, \ldots, s_t$.

$\mathbb{P}(X_0 = s_0, X_1 = s_1, \ldots, X_t = s_t)$

$= \mathbb{P}(X_t = s_t \mid X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) \times \mathbb{P}(X_{t-1} = s_{t-1}, \ldots, X_0 = s_0)$

$= \mathbb{P}(X_t = s_t \mid X_{t-1} = s_{t-1}) \times \mathbb{P}(X_{t-1} = s_{t-1}, \ldots, X_0 = s_0)$ (Markov Property)

$= p_{s_{t-1}, s_t} \mathbb{P}(X_{t-1} = s_{t-1} \mid X_{t-2} = s_{t-2}, \ldots, X_0 = s_0) \times \mathbb{P}(X_{t-2} = s_{t-2}, \ldots, X_0 = s_0)$

$\vdots$

$= p_{s_{t-1}, s_t} \times p_{s_{t-2}, s_{t-1}} \times \ldots \times p_{s_0, s_1} \times \mathbb{P}(X_0 = s_0)$

$= p_{s_{t-1}, s_t} \times p_{s_{t-2}, s_{t-1}} \times \ldots \times p_{s_0, s_1} \times \pi_{s_0}.$

8.9 Worked Example: distribution of X_t and trajectory probabilities

Purpose-flea zooms around the vertices of the transition diagram opposite. Let X_t be Purpose-flea's state at time t ($t = 0, 1, \ldots$).
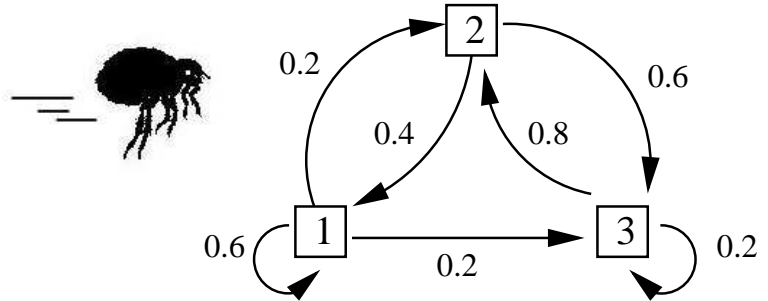
(a) Find the transition matrix, P.

Answer: $P = \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.4 & 0 & 0.6 \\ 0 & 0.8 & 0.2 \end{pmatrix}$

(b) Find $\mathbb{P}(X_2 = 3 \,|\, X_0 = 1)$.

$$\mathbb{P}(X_2 = 3 \,|\, X_0 = 1) = \left(P^2\right)_{13} = \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} \cdot & \cdot & 0.2 \\ \cdot & \cdot & 0.6 \\ \cdot & \cdot & 0.2 \end{pmatrix}$$

$$= 0.6 \times 0.2 + 0.2 \times 0.6 + 0.2 \times 0.2$$

$$= 0.28.$$

Note: we only need one element of the matrix P^2, so don't lose exam time by finding the whole matrix.

(c) Suppose that Purpose-flea is equally likely to start on any vertex at time 0. Find the probability distribution of X_1.

From this info, the distribution of X_0 is $\boldsymbol{\pi}^T = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. We need $X_1 \sim \boldsymbol{\pi}^T P$.

$$\boldsymbol{\pi}^T P = \begin{matrix} \left(\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}\right) \end{matrix} \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.4 & 0 & 0.6 \\ 0 & 0.8 & 0.2 \end{pmatrix} = \left(\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}\right).$$

Thus $X_1 \sim \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ and therefore X_1 is also equally likely to be 1, 2, or 3.

(d) Suppose that Purpose-flea begins at vertex 1 at time 0. Find the probability distribution of X_2.

The distribution of X_0 is now $\boldsymbol{\pi}^T = (1, 0, 0)$. We need $X_2 \sim \boldsymbol{\pi}^T P^2$.

$$\boldsymbol{\pi}^T P^2 \quad = \quad (1 \ 0 \ 0) \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.4 & 0 & 0.6 \\ 0 & 0.8 & 0.2 \end{pmatrix} \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.4 & 0 & 0.6 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

$$= \quad (0.6 \ 0.2 \ 0.2) \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.4 & 0 & 0.6 \\ 0 & 0.8 & 0.2 \end{pmatrix}$$

$$= \quad (0.44 \quad 0.28 \quad 0.28).$$

Thus $\mathbb{P}(X_2 = 1) = 0.44, \ \mathbb{P}(X_2 = 2) = 0.28, \ \mathbb{P}(X_2 = 3) = 0.28.$

Note that it is quickest to multiply the vector by the matrix first: we don't need to compute P^2 in entirety.

(e) Suppose that Purpose-flea is equally likely to start on any vertex at time 0. Find the probability of obtaining the trajectory (3, 2, 1, 1, 3).

$$\begin{aligned} \mathbb{P}(3, 2, 1, 1, 3) &= \mathbb{P}(X_0 = 3) \times p_{32} \times p_{21} \times p_{11} \times p_{13} \quad \textit{(Section 8.8)} \\ &= \tfrac{1}{3} \times 0.8 \times 0.4 \times 0.6 \times 0.2 \\ &= 0.0128. \end{aligned}$$

8.10 Class Structure

The state space of a Markov chain can be partitioned into a set of non-overlapping *communicating classes.*

States i and j are in the same communicating class if there is some way of getting from state i to state j, AND there is some way of getting from state j to state i. It needn't be possible to get between i and j in a ***single*** step, but it must be possible over some number of steps to travel between them both ways.

We write $i \leftrightarrow j$.

Definition: Consider a Markov chain with state space S and transition matrix P, and consider states $i, j \in S$. Then **state i communicates with state j** if:

1. there exists some t such that $(P^t)_{ij} > 0$, AND

2. there exists some u such that $(P^u)_{ji} > 0$.

Mathematically, it is easy to show that the communicating relation $\leftrightarrow$ is an **equivalence relation,** which means that it partitions the sample space S into non-overlapping equivalence classes.

Definition: States i and j are in the same **communicating class** if $i \leftrightarrow j$: *i.e. if each state is accessible from the other.*

Every state is a member of *exactly one communicating class.*

Example: Find the communicating classes associated with the transition diagram shown.
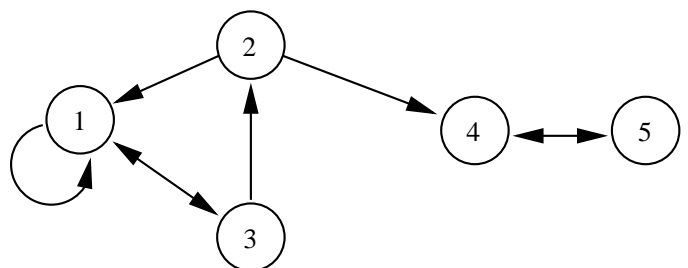
Solution:

$\{1, 2, 3\}, \quad \{4, 5\}.$

State 2 leads to state 4, but state 4 does not lead back to state 2, so they are in different communicating classes.

Definition: A communicating class of states is **closed** if *it is not possible to leave that class.*

That is, the communicating class C is closed if $p_{ij} = 0$ whenever $i \in C$ and $j \notin C$.

Example: In the transition diagram above:

- Class $\{1, 2, 3\}$ is *not* *closed: it is possible to escape to class $\{4, 5\}$.*

- Class $\{4, 5\}$ is *closed: it is not possible to escape.*

Definition: A state i is said to be **absorbing** if *the set $\{i\}$ is a closed class.*
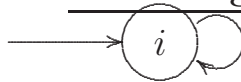
Definition: A Markov chain or transition matrix P is said to be **irreducible** if $i \leftrightarrow j$ *for all $i, j \in S$.* *That is, the chain is irreducible if the state space S is a single communicating class.*

8.11 Hitting Probabilities

We have been calculating hitting probabilities for Markov chains since Chapter 2, using First-Step Analysis. The hitting probability describes the probability that the Markov chain will *ever* reach some state or set of states.

In this section we show how hitting probabilities can be written in a single vector. We also see a general formula for calculating the hitting probabilities. In general it is easier to continue using our own common sense, but occasionally the formula becomes more necessary.

Vector of hitting probabilities

Let A be some subset of the state space S. (A need not be a communicating class: it can be any subset required, including the subset consisting of a single state: e.g. $A = \{4\}$.)

The **hitting probability** from state i to set A is the probability of <u>ever</u> reaching the set A, starting from initial state i. We write this probability as h_{iA}. Thus

$$h_{iA} = \mathbb{P}(X_t \in A \text{ for some } t \geq 0 \mid X_0 = i).$$

Example: Let set $A = \{1, 3\}$ as shown.

The hitting probability for set A is:
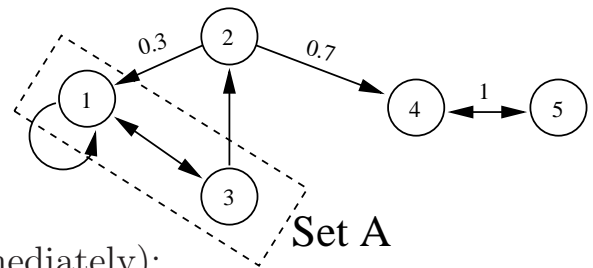
- *1 starting from states 1 or 3*
  (We are starting in set A, so we hit it immediately);

- *0 starting from states 4 or 5*
  (The set $\{4, 5\}$ is a closed class, so we can never escape out to set A);

- *0.3 starting from state 2*
  (We could hit A at the first step (probability 0.3), but otherwise we move to state 4 and get stuck in the closed class $\{4, 5\}$ (probability 0.7).)

We can summarize all the information from the example above in a *vector of hitting probabilities:*

$$\boldsymbol{h_A} = \begin{pmatrix} h_{1A} \\ h_{2A} \\ h_{3A} \\ h_{4A} \\ h_{5A} \end{pmatrix} = \begin{pmatrix} 1 \\ 0.3 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

Note: When A is a closed class, the hitting probability h_{iA} is called the *absorption probability*.

In general, if there are N possible states, the vector of hitting probabilities is

$$\boldsymbol{h_A} = \begin{pmatrix} h_{1A} \\ h_{2A} \\ \vdots \\ h_{NA} \end{pmatrix} = \begin{pmatrix} \mathbb{P}(\textit{hit A starting from state 1}) \\ \mathbb{P}(\textit{hit A starting from state 2}) \\ \vdots \\ \mathbb{P}(\textit{hit A starting from state N}) \end{pmatrix}.$$

Example: finding the hitting probability vector using First-Step Analysis

Suppose $\{X_t : t \geq 0\}$ has the following transition diagram:
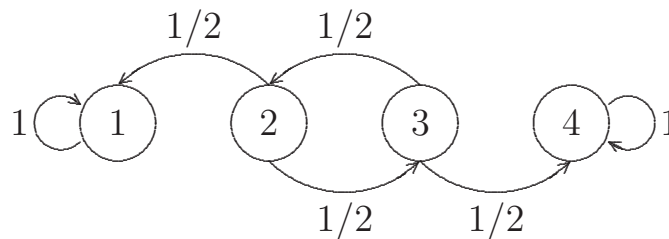
Find the vector of hitting probabilities for state 4.

Solution:

Let $h_{i4} = \mathbb{P}(\textit{hit state 4, starting from state } i)$. *Clearly,*

$$h_{14} = 0$$
$$h_{44} = 1$$

Using first-step analysis, we also have:

$$h_{24} = \tfrac{1}{2}h_{34} + \tfrac{1}{2} \times 0$$
$$h_{34} = \tfrac{1}{2} + \tfrac{1}{2}h_{24}$$

Solving,

$$h_{34} = \tfrac{1}{2} + \tfrac{1}{2}\left(\tfrac{1}{2}h_{34}\right) \quad \Rightarrow \quad h_{34} = \tfrac{2}{3}. \qquad \textbf{\textit{So also, }} h_{24} = \tfrac{1}{2}h_{34} = \tfrac{1}{3}.$$

So the vector of hitting probabilities is

$$\boldsymbol{h_A} = \left(0, \ \tfrac{1}{3}, \ \tfrac{2}{3}, \ 1\right).$$

Formula for hitting probabilities

In the previous example, we used our common sense to state that $h_{14} = 0$. While this is easy for a human brain, it is harder to explain a general rule that would describe this 'common sense' mathematically, or that could be used to write computer code that will work for all problems.

Although it is usually best to continue to use common sense when solving problems, this section provides a general formula that will *always* work to find a vector of hitting probabilities $\boldsymbol{h_A}$.

__Theorem 8.11:__ The vector of hitting probabilities $\boldsymbol{h_A} = (h_{iA} : i \in S)$ is the minimal non-negative solution to the following equations:

$$h_{iA} = \begin{cases} 1 & \text{for} \quad i \in A, \\ \displaystyle\sum_{j \in S} p_{ij} h_{jA} & \text{for} \quad i \notin A. \end{cases}$$

The 'minimal non-negative solution' means that:

1. the values $\{h_{iA}\}$ collectively satisfy the equations above;
2. each value h_{iA} is ≥ 0 (non-negative);
3. given any other non-negative solution to the equations above, say $\{g_{iA}\}$ where $g_{iA} \geq 0$ for all i, then $h_{iA} \leq g_{iA}$ for all i (minimal solution).

Example: How would this formula be used to substitute for 'common sense' in the previous example?

The equations give:

$$h_{i4} = \begin{cases} 1 & \text{if} \quad i = 4, \\ \displaystyle\sum_{j \in S} p_{ij} h_{j4} & \text{if} \quad i \neq 4. \end{cases}$$
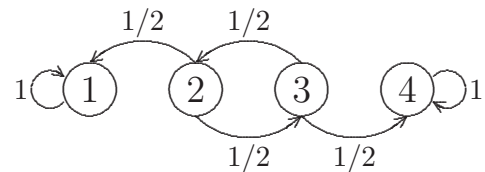
Thus,

$$
\begin{aligned}
h_{44} &= 1 \\
h_{14} &= h_{14} \quad \textit{unspecified! Could be anything!} \\
h_{24} &= \tfrac{1}{2}h_{14} + \tfrac{1}{2}h_{34} \\
h_{34} &= \tfrac{1}{2}h_{24} + \tfrac{1}{2}h_{44} = \tfrac{1}{2}h_{24} + \tfrac{1}{2}
\end{aligned}
$$

*Because h_{14} could be anything, we have to use the <u>minimal non-negative value,</u>
which is $h_{14} = 0$.*
(Need to check $h_{14} = 0$ does not force $h_{i4<0}$ for any other i: OK.)

The other equations can then be solved to give the same answers as before. $\square$

Proof of Theorem 8.11 (non-examinable):

Consider the equations $\quad h_{iA} = \begin{cases} 1 & \text{for} \quad i \in A, \\ \sum_{j \in S} p_{ij} h_{jA} & \text{for} \quad i \notin A. \end{cases}$ $\qquad (\star)$

We need to show that:

  (i) the hitting probabilities $\{h_{iA}\}$ collectively satisfy the equations $(\star)$;

  (ii) if $\{g_{iA}\}$ is any other non-negative solution to $(\star)$, then the hitting probabilities $\{h_{iA}\}$ satisfy $h_{iA} \le g_{iA}$ for all i (minimal solution).

<u>Proof of (i):</u> Clearly, $h_{iA} = 1$ if $i \in A$ (as the chain hits A immediately).

Suppose that $i \notin A$. Then

$$
\begin{aligned}
h_{iA} &= \mathbb{P}(X_t \in A \text{ for some } t \ge 1 \,|\, X_0 = i) \\
&= \sum_{j \in S} \mathbb{P}(X_t \in A \text{ for some } t \ge 1 \,|\, X_1 = j)\mathbb{P}(X_1 = j \,|\, X_0 = i) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(Partition Rule)} \\
&= \sum_{j \in S} h_{jA}\, p_{ij} \qquad \text{(by definitions)}.
\end{aligned}
$$

Thus the hitting probabilities $\{h_{iA}\}$ must satisfy the equations $(\star)$.

<u>Proof of (ii):</u> Let $h_{iA}^{(t)} = \mathbb{P}(\text{hit } A \text{ at or before time } t \,|\, X_0 = i)$.

We use mathematical induction to show that $h_{iA}^{(t)} \le g_{iA}$ for all t, and therefore $h_{iA} = \lim_{t \to \infty} h_{iA}^{(t)}$ must also be $\le g_{iA}$.

<u>Time $t = 0$:</u> $h_{iA}^{(0)} = \begin{cases} 1 & \text{if} \quad i \in A, \\ 0 & \text{if} \quad i \notin A. \end{cases}$

But because g_{iA} is non-negative and satisfies $(\star)$, $\begin{cases} g_{iA} = 1 & \text{if } i \in A, \\ g_{iA} \geq 0 & \text{for all } i. \end{cases}$

So $g_{iA} \geq h_{iA}^{(0)}$ for all i.

The inductive hypothesis is true for time $t = 0$.

<u>Time t:</u> Suppose the inductive hypothesis holds for time t, i.e.

$$h_{jA}^{(t)} \leq g_{jA} \quad \text{for all } j.$$

Consider

$$
\begin{aligned}
h_{iA}^{(t+1)} &= \mathbb{P}(\text{hit } A \text{ by time } t+1 \mid X_0 = i) \\
&= \sum_{j \in S} \mathbb{P}(\text{hit } A \text{ by time } t+1 \mid X_1 = j)\mathbb{P}(X_1 = j \mid X_0 = i) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Partition Rule)} \\
&= \sum_{j \in S} h_{jA}^{(t)}\, p_{ij} \qquad \text{by definitions} \\
&\leq \sum_{j \in S} g_{jA}\, p_{ij} \qquad \text{by inductive hypothesis} \\
&= g_{iA} \qquad \text{because } \{g_{iA}\} \text{ satisfies } (\star).
\end{aligned}
$$

Thus $h_{iA}^{(t+1)} \leq g_{iA}$ for all i, so the inductive hypothesis is proved.

By the Continuity Theorem (Chapter 2), $h_{iA} = \lim_{t \to \infty} h_{iA}^{(t)}$.

So $h_{iA} \leq g_{iA}$ as required. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ $\square$

8.12 Expected hitting times

In the previous section we found the **probability** of hitting set A, starting at state i. Now we study **how long** it takes to get from i to A. As before, it is best to solve problems using first-step analysis and common sense. However, a general formula is also available.

Definition: Let A be a subset of the state space S. The **hitting time** of A is the random variable T_A, where

$$T_A = \min\{t \geq 0 : X_t \in A\}.$$

T_A is the time taken before hitting set A *for the first time.*

The hitting time T_A can take values $0, 1, 2, \ldots,$ *and* ∞.
If the chain *never* hits set A, then $T_A = \infty$.

Note: The hitting time is also called the **reaching time**. If A is a closed class, it is also called the *absorption time.*

Definition: The **mean hitting time** for A, starting from state i, is

$$m_{iA} = \mathbb{E}(T_A \,|\, X_0 = i).$$

Note: If there is any possibility that the chain *never* reaches A, starting from i, *i.e. if the hitting probability $h_{iA} < 1$, then* $\mathbb{E}(T_A \,|\, X_0 = i) = \infty$.

Calculating the mean hitting times

Theorem 8.12: The vector of expected hitting times $\boldsymbol{m_A} = (m_{iA} : i \in S)$ is *the minimal non-negative solution to the following equations:*

$$m_{iA} = \begin{cases} 0 & \text{for} \quad i \in A, \\ 1 + \displaystyle\sum_{j \notin A} p_{ij} m_{jA} & \text{for} \quad i \notin A. \end{cases}$$

Proof (sketch):

Consider the equations $m_{iA} = \begin{cases} 0 & \text{for} \quad i \in A, \\ 1 + \sum_{j \notin A} p_{ij} m_{jA} & \text{for} \quad i \notin A. \end{cases}$ $(\star)$.

We need to show that:

(i) the mean hitting times $\{m_{iA}\}$ collectively satisfy the equations $(\star)$;

(ii) if $\{u_{iA}\}$ is any other non-negative solution to $(\star)$, then the mean hitting times $\{m_{iA}\}$ satisfy $m_{iA} \le u_{iA}$ for all i (minimal solution).

We will prove point (i) only. A proof of (ii) can be found online at:
`http://www.statslab.cam.ac.uk/~james/Markov/` , Section 1.3.

Proof of (i): Clearly, $m_{iA} = 0$ if $i \in A$ (as the chain hits A immediately).

Suppose that $i \notin A$. Then

$$
\begin{aligned}
m_{iA} &= \mathbb{E}(T_A \mid X_0 = i) \\[2mm]
&= 1 + \sum_{j \in S} \mathbb{E}(T_A \mid X_1 = j)\mathbb{P}(X_1 = j \mid X_0 = i) \\
&\quad \text{(conditional expectation: take 1 step to get to state } j \\
&\quad \text{ at time 1, then find } \mathbb{E}(T_A) \text{ from there)} \\[2mm]
&= 1 + \sum_{j \in S} m_{jA}\, p_{ij} \qquad \text{(by definitions)} \\[2mm]
&= 1 + \sum_{j \notin A} p_{ij}\, m_{jA}, \qquad \text{because } m_{jA} = 0 \text{ for } j \in A.
\end{aligned}
$$

Thus the mean hitting times $\{m_{iA}\}$ must satisfy the equations $(\star)$.

$\square$

Example: Let $\{X_t : t \ge 0\}$ have the same transition diagram as before:

Starting from state 2, find the
expected time to absorption.
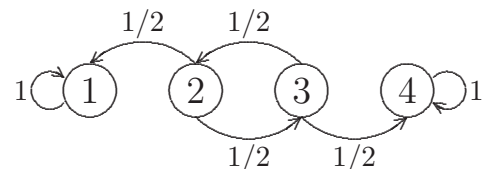
Solution:

Starting from state $i = 2$, we wish to find the expected time to reach the set $A = \{1, 4\}$ (the set of absorbing states).

Thus we are looking for $m_{iA} = m_{2A}$.

$$\text{Now} \quad m_{iA} = \begin{cases} 0 & \text{if} \quad i \in \{1, 4\}, \\ 1 + \displaystyle\sum_{j \notin A} p_{ij} m_{jA} & \text{if} \quad i \notin \{1, 4\}. \end{cases}$$

Thus,

$$m_{1A} = 0 \quad (\text{because } 1 \in A)$$

$$m_{4A} = 0 \quad (\text{because } 4 \in A)$$

$$m_{2A} = 1 + \tfrac{1}{2} m_{1A} + \tfrac{1}{2} m_{3A}$$

$$\Rightarrow \quad m_{2A} = 1 + \tfrac{1}{2} m_{3A}$$

$$m_{3A} = 1 + \tfrac{1}{2} m_{2A} + \tfrac{1}{2} m_{4A}$$

$$= 1 + \tfrac{1}{2} m_{2A}$$

$$= 1 + \tfrac{1}{2} \left(1 + \tfrac{1}{2} m_{3A} \right)$$

$$\Rightarrow \quad \tfrac{3}{4} m_{3A} = \tfrac{3}{2}$$

$$\Rightarrow \quad m_{3A} = 2.$$

Thus,

$$m_{2A} = 1 + \frac{1}{2} m_{3A} = 2.$$

The expected time to absorption is therefore $\mathbb{E}(T_A) = 2$ steps.
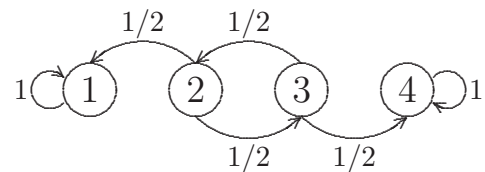
Example: Glee-flea hops around on a triangle. At each step he moves to one of the other two vertices at random. What is the expected time taken for Glee-flea to get from vertex 1 to vertex 2?

Solution:
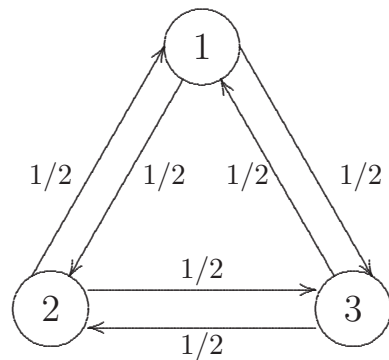
transition matrix, $P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$.

We wish to find m_{12}.

Now $\quad m_{i2} = \begin{cases} 0 & \text{if } i = 2, \\ 1 + \sum_{j \neq 2} p_{ij} m_{j2} & \text{if } i \neq 2. \end{cases}$

Thus

$$m_{22} = 0$$

$$m_{12} = 1 + \tfrac{1}{2}m_{22} + \tfrac{1}{2}m_{32} = 1 + \tfrac{1}{2}m_{32}.$$

$$m_{32} = 1 + \tfrac{1}{2}m_{22} + \tfrac{1}{2}m_{12}$$

$$= 1 + \tfrac{1}{2}m_{12}$$

$$= 1 + \tfrac{1}{2}\left(1 + \tfrac{1}{2}m_{32}\right)$$

$$\Rightarrow \quad m_{32} = 2.$$

Thus $m_{12} = 1 + \tfrac{1}{2}m_{32} = 2$ *steps.*

Chapter 9: Equilibrium
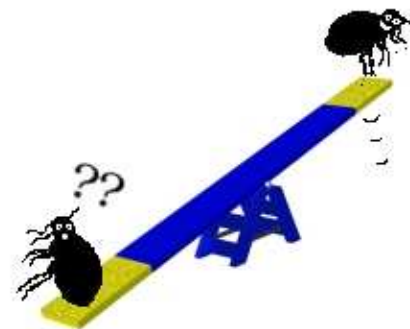
In Chapter 8, we saw that if $\{X_0, X_1, X_2, \ldots\}$ is a Markov chain with transition matrix P, then
$$X_t \sim \boldsymbol{\pi}^T \quad \Rightarrow \quad X_{t+1} \sim \boldsymbol{\pi}^T P.$$

This raises the question: is there any distribution $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$?

If $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$, then

$$
\begin{aligned}
X_t \sim \boldsymbol{\pi}^T \quad &\Rightarrow \quad X_{t+1} \sim \boldsymbol{\pi}^T P = \boldsymbol{\pi}^T \\
&\Rightarrow \quad X_{t+2} \sim \boldsymbol{\pi}^T P = \boldsymbol{\pi}^T \\
&\Rightarrow \quad X_{t+3} \sim \boldsymbol{\pi}^T P = \boldsymbol{\pi}^T \\
&\Rightarrow \quad \ldots
\end{aligned}
$$

In other words, if $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$, and $X_t \sim \boldsymbol{\pi}^T$, then

$$X_t \sim X_{t+1} \sim X_{t+2} \sim X_{t+3} \sim \ldots$$

Thus, once a Markov chain has reached a distribution $\boldsymbol{\pi}^T$ such that $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$, *it will stay there.*

If $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$, we say that the distribution $\boldsymbol{\pi}^T$ is an *equilibrium distribution.*

Equilibrium means a ***level position:*** there is *no more change* in the distribution of X_t as we wander through the Markov chain.

Note: Equilibrium does not mean that the <u>**value**</u> of X_{t+1} equals the <u>**value**</u> of X_t. It means that the <u>**distribution**</u> of X_{t+1} is the same as the <u>**distribution**</u> of X_t:

e.g. $\mathbb{P}(X_{t+1} = 1) = \mathbb{P}(X_t = 1) = \pi_1$;

$\mathbb{P}(X_{t+1} = 2) = \mathbb{P}(X_t = 2) = \pi_2$,   *etc.*

In this chapter, we will first see how to *calculate* the equilibrium distribution $\boldsymbol{\pi}^T$. We will then see the remarkable result that many Markov chains automatically *find their own way* to an equilibrium distribution as the chain wanders through time. This happens for many Markov chains, but not all. We will see the conditions required for the chain to find its way to an equilibrium distribution.
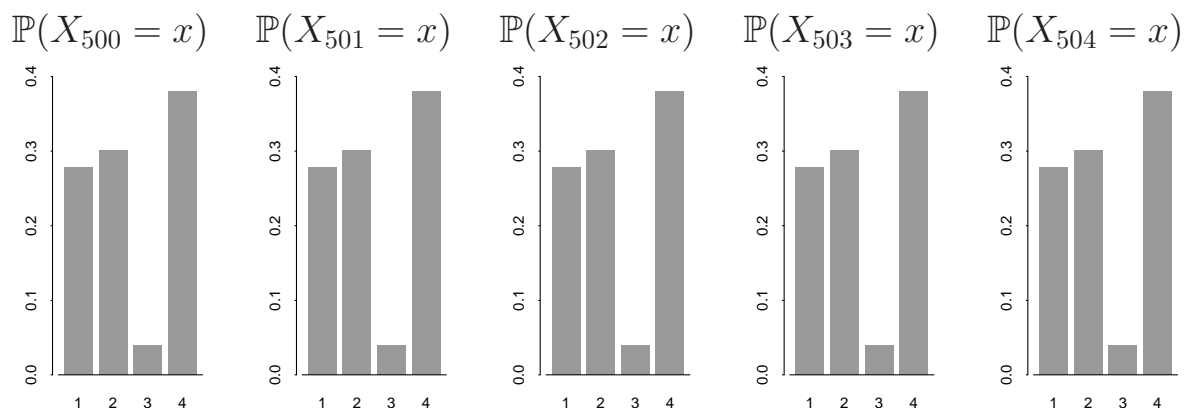
9.1 Equilibrium distribution in pictures

Consider the following 4-state Markov chain:

$$P = \begin{pmatrix} 0.0 & 0.9 & 0.1 & 0.0 \\ 0.8 & 0.1 & 0.0 & 0.1 \\ 0.0 & 0.5 & 0.3 & 0.2 \\ 0.1 & 0.0 & 0.0 & 0.9 \end{pmatrix}$$

Suppose we start at time 0 with $X_0 \sim \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$: so the chain is equally likely to start from any of the four states. Here are pictures of the distributions of $X_0, X_1, \ldots, X_4$:

$$\mathbb{P}(X_0 = x) \qquad \mathbb{P}(X_1 = x) \qquad \mathbb{P}(X_2 = x) \qquad \mathbb{P}(X_3 = x) \qquad \mathbb{P}(X_4 = x)$$
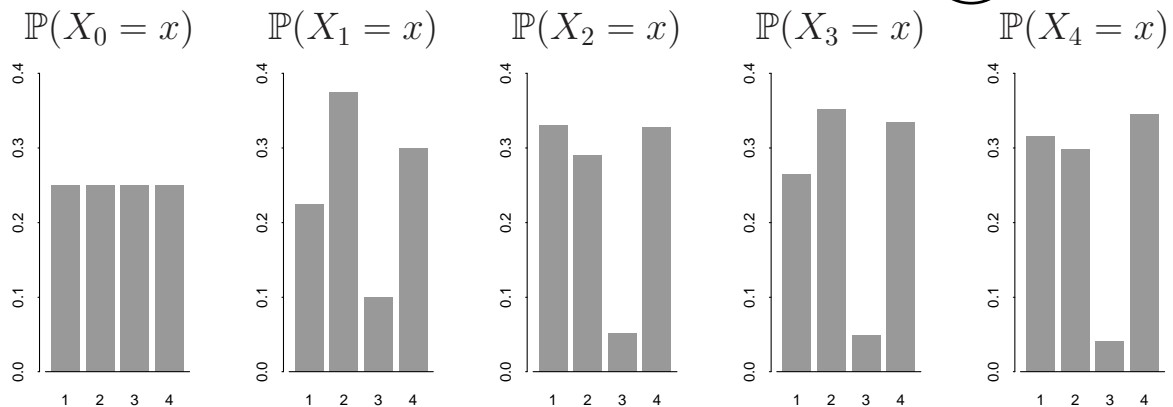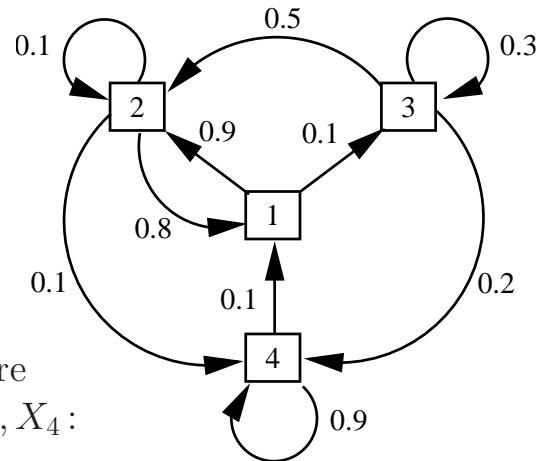
The distribution starts off level, but quickly changes: for example the chain is least likely to be found in state 3. The distribution of X_t changes between each $t = 0, 1, 2, 3, 4$. Now look at the distribution of X_t 500 steps into the future:

$$\mathbb{P}(X_{500} = x) \qquad \mathbb{P}(X_{501} = x) \qquad \mathbb{P}(X_{502} = x) \qquad \mathbb{P}(X_{503} = x) \qquad \mathbb{P}(X_{504} = x)$$

The distribution has reached a steady state: it **does not change** between $t = 500, 501, \ldots, 504$. *The chain has reached equilibrium of its own accord.*

9.2 Calculating equilibrium distributions

Definition: Let $\{X_0, X_1, \ldots\}$ be a Markov chain with transition matrix P and state space S, where $|S| = N$ (possibly infinite). Let $\boldsymbol{\pi}^T$ be a row vector denoting a probability distribution on S: so each element π_i denotes the probability of being in state i, and $\sum_{i=1}^{N} \pi_i = 1$, where $\pi_i \geq 0$ for all $i = 1, \ldots, N$. The probability distribution $\boldsymbol{\pi}^T$ is an **equilibrium** distribution for the Markov chain if $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$.

That is, $\boldsymbol{\pi}^T$ is an equilibrium distribution if

$$\left(\boldsymbol{\pi}^T P\right)_j = \sum_{i=1}^{N} \pi_i p_{ij} = \pi_j \quad \text{for all } j = 1, \ldots, N.$$

By the argument given on page 164, we have the following Theorem:

Theorem 9.2: Let $\{X_0, X_1, \ldots\}$ be a Markov chain with transition matrix P. Suppose that $\boldsymbol{\pi}^T$ is an equilibrium distribution for the chain. If $X_t \sim \boldsymbol{\pi}^T$ for any t, then $X_{t+r} \sim \boldsymbol{\pi}^T$ for all $r \geq 0$. $\qquad\square$

Once a chain has hit an equilibrium distribution, *it stays there for ever.*

Note: There are several other names for an equilibrium distribution. If $\boldsymbol{\pi}^T$ is an equilibrium distribution, it is also called:

- **invariant:** *it doesn't change:* $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$;
- **stationary:** *the chain 'stops' here.*

Stationarity: the Chain Station

a B U S station is where a B U S stops

a t r a i n station is where a t r a i n stops

a **workstation** is where . . . ? ? ?

a stationary distribution is where a Markov chain stops

9.3 Finding an equilibrium distribution

Vector $\boldsymbol{\pi}^T$ is an equilibrium distribution for P if:

1. $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$;

2. $\sum_{i=1}^{N} \pi_i = 1$;

3. $\pi_i \geq 0$ *for all* i.

Conditions 2 and 3 ensure that $\boldsymbol{\pi}^T$ *is a genuine probability distribution.*

Condition 1 means that $\boldsymbol{\pi}$ is a <u>row eigenvector</u> of P.

Solving $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$ by itself will just specify $\boldsymbol{\pi}$ up to *a scalar multiple.*
We need to include Condition 2 to scale $\boldsymbol{\pi}$ to a genuine probability distribution,
and then check with Condition 3 that the scaled distribution is valid.

Example: Find an equilibrium distribution for the Markov chain below.

$$
P = \begin{pmatrix} 0.0 & 0.9 & 0.1 & 0.0 \\ 0.8 & 0.1 & 0.0 & 0.1 \\ 0.0 & 0.5 & 0.3 & 0.2 \\ 0.1 & 0.0 & 0.0 & 0.9 \end{pmatrix}
$$
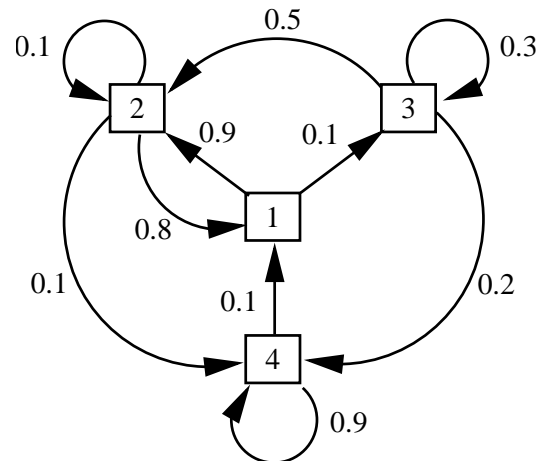
Solution:

Let $\boldsymbol{\pi}^T = (\pi_1, \ \pi_2, \ \pi_3, \ \pi_4)$.
The equations are $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$ *and* $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$.

$$
\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T \quad \Rightarrow \quad (\pi_1 \ \pi_2 \ \pi_3 \ \pi_4) \begin{pmatrix} 0.0 & 0.9 & 0.1 & 0.0 \\ 0.8 & 0.1 & 0.0 & 0.1 \\ 0.0 & 0.5 & 0.3 & 0.2 \\ 0.1 & 0.0 & 0.0 & 0.9 \end{pmatrix} = (\pi_1 \ \pi_2 \ \pi_3 \ \pi_4)
$$

$$.8\pi_2 + .1\pi_4 = \pi_1 \qquad (1)$$

$$.9\pi_1 + .1\pi_2 + .5\pi_3 = \pi_2 \qquad (2)$$

$$.1\pi_1 + .3\pi_3 = \pi_3 \qquad (3)$$

$$.1\pi_2 + .2\pi_3 + .9\pi_4 = \pi_4 \qquad (4)$$

$$\textit{Also} \qquad \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1. \qquad (5)$$

$$(3) \quad \Rightarrow \quad \pi_1 = 7\pi_3$$

$$\textit{Substitute in (2)} \quad \Rightarrow \quad .9(7\pi_3) + .5\pi_3 = .9\pi_2$$

$$\Rightarrow \quad \pi_2 = \frac{68}{9}\pi_3$$

$$\textit{Substitute in (1)} \quad \Rightarrow \quad .8\left(\frac{68}{9}\pi_3\right) + .1\pi_4 = 7\pi_3$$

$$\Rightarrow \quad \pi_4 = \frac{86}{9}\pi_3$$

$$\textit{Substitute all in (5)} \quad \Rightarrow \quad \pi_3\left(7 + \frac{68}{9} + 1 + \frac{86}{9}\right) = 1$$

$$\Rightarrow \quad \pi_3 = \frac{9}{226}$$

Overall:

$$\boldsymbol{\pi}^T = \left(\frac{63}{226}, \ \frac{68}{226}, \ \frac{9}{226}, \ \frac{86}{226}\right)$$

$$= (0.28, \ 0.30, \ 0.04, \ 0.38).$$

This is the distribution the chain converged to in Section 9.1.

9.4 Long-term behaviour

In Section 9.1, we saw an example where the Markov chain wandered of its own accord into its equilibrium distribution:

$$\mathbb{P}(X_{500} = x) \quad \mathbb{P}(X_{501} = x) \quad \mathbb{P}(X_{502} = x) \quad \mathbb{P}(X_{503} = x)$$
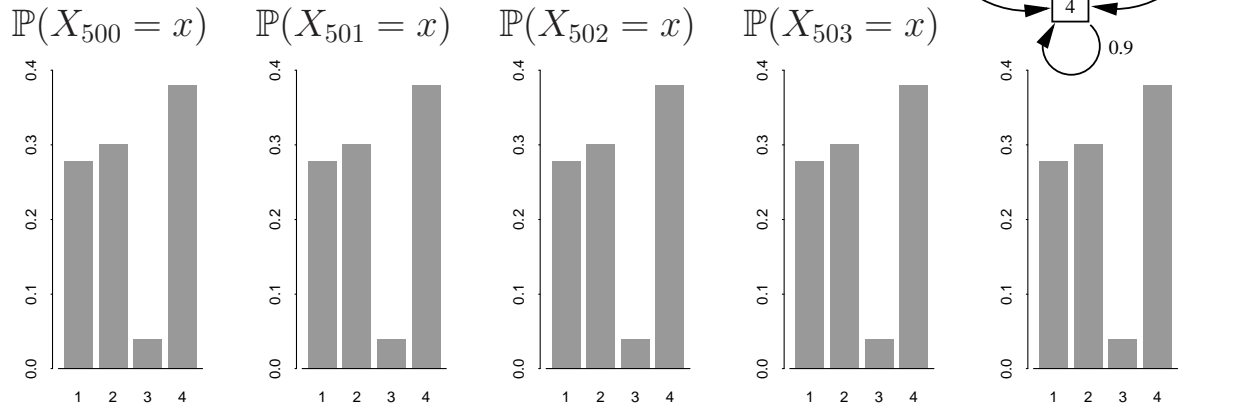
This will always happen for this Markov chain. In fact, the distribution it converges to (found above) does not depend upon the starting conditions: *for ANY value of X_0, we will always have $X_t \sim (0.28,\ 0.30,\ 0.04,\ 0.38)$ as $t \to \infty$.*

What is happening here is that *each row of the transition matrix P^t converges to the equilibrium distribution* $(0.28,\ 0.30,\ 0.04,\ 0.38)$ *as $t \to \infty$:*

$$P = \begin{pmatrix} 0.0 & 0.9 & 0.1 & 0.0 \\ 0.8 & 0.1 & 0.0 & 0.1 \\ 0.0 & 0.5 & 0.3 & 0.2 \\ 0.1 & 0.0 & 0.0 & 0.9 \end{pmatrix} \quad \Rightarrow \quad P^t \to \begin{pmatrix} 0.28 & 0.30 & 0.04 & 0.38 \\ 0.28 & 0.30 & 0.04 & 0.38 \\ 0.28 & 0.30 & 0.04 & 0.38 \\ 0.28 & 0.30 & 0.04 & 0.38 \end{pmatrix} \quad \text{as } t \to \infty.$$

(If you have a calculator that can handle matrices, try finding P^t for $t = 20$ and $t = 30$: you will find the matrix is already converging as above.)

This convergence of P^t means that *for large t, no matter WHICH state we start in, we always have probability*

- about *0.28* of being in State *1* after t steps;
- about *0.30* of being in State *2* after t steps;
- about *0.04* of being in State *3* after t steps;
- about *0.38* of being in State *4* after t steps.
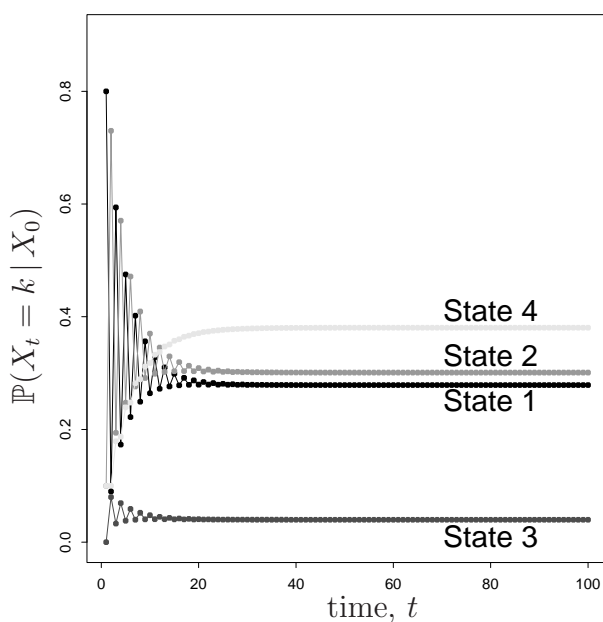
Start at $X_0 = 2$ Start at $X_0 = 4$
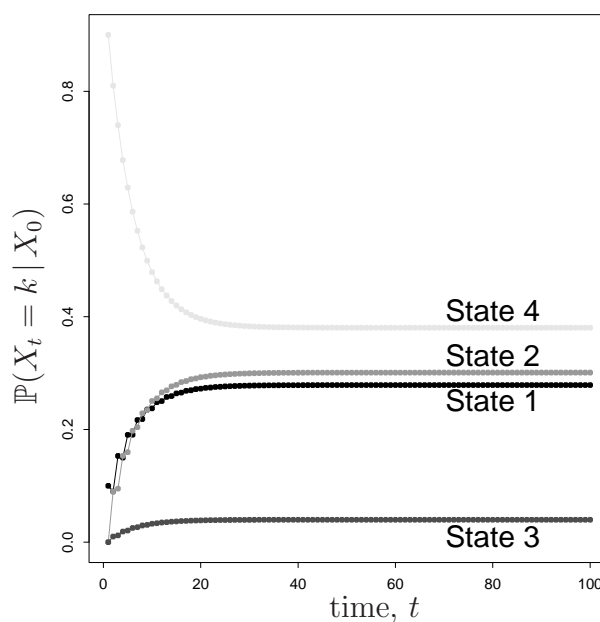
The **left graph** shows the probability of getting from state 2 to state k in t steps, as t changes: $(P^t)_{2,k}$ for $k = 1, 2, 3, 4$.

The **right graph** shows the probability of getting from state 4 to state k in t steps, as t changes: $(P^t)_{4,k}$ for $k = 1, 2, 3, 4$.

The ***initial behaviour*** differs greatly for the different start states.

The ***long-term behaviour*** (large t) is the same for both start states.

However, this does not always happen. Consider the two-state chain below:

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
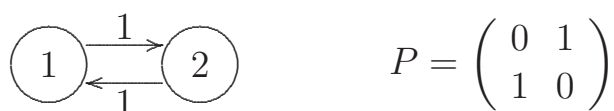
As t gets large, P^t ***does not converge:***

$$P^{500} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad P^{501} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad P^{502} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad P^{503} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdots$$

For this Markov chain, *we never 'forget' the initial start state.*

General formula for P^t

We have seen that we are interested in whether P^t converges to *a fixed matrix with all rows equal as* $t \to \infty$.

If it does, then the Markov chain will *reach an equilibrium distribution that does not depend upon the starting conditions.*

The equilibrium distribution is then given by *any row of the converged P^t.*

It can be shown that a general formula is available for P^t for any t, based on the eigenvalues of P. Producing this formula is beyond the scope of this course, but if you are given the formula, you should be able to recognise whether P^t is going to converge to a fixed matrix with all rows the same.

Example 1:

$$P = \begin{pmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{pmatrix}$$
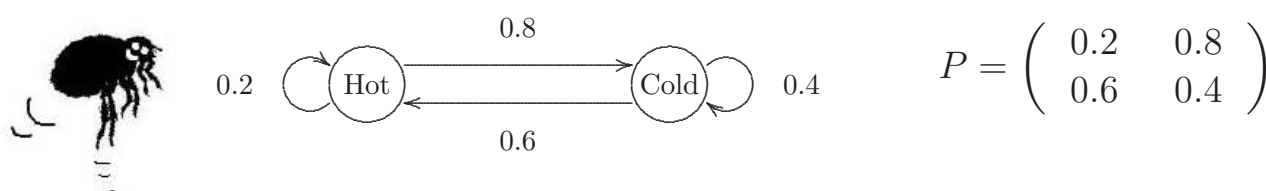
We can show that the general solution for P^t is:

$$P^t = \frac{1}{7}\left\{ \begin{pmatrix} 3 & 4 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 4 & -4 \\ -3 & 3 \end{pmatrix}(-0.4)^t \right\}$$

As $t \to \infty$, $(-0.4)^t \to 0$, *so*

$$P^t \to \frac{1}{7}\begin{pmatrix} 3 & 4 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} \frac{3}{7} & \frac{4}{7} \\ \frac{3}{7} & \frac{4}{7} \end{pmatrix}$$

This Markov chain will therefore converge to the equilibrium distribution $\boldsymbol{\pi}^T = \left(\frac{3}{7}, \frac{4}{7}\right)$ *as* $t \to \infty$, *regardless of whether the flea starts in state 1 or state 2.*

Exercise: Verify that $\boldsymbol{\pi}^T = \left(\frac{3}{7}, \frac{4}{7}\right)$ is the same as the result you obtain from solving the equilibrium equations: $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$ and $\pi_1 + \pi_2 = 1$.

Example 2: Purposeflea knows exactly what he is doing, so his probabilities are all 1:
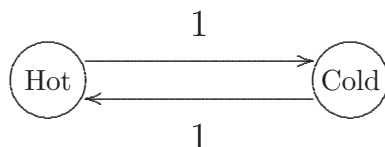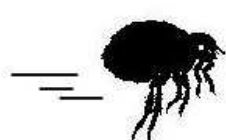
$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

We can show that the general solution for P^t is:

$$P^t = \frac{1}{2} \left\{ \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} (-1)^t \right\}$$

As $t \to \infty$, $(-1)^t$ does not converge to 0, so

$$P^t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{if } t \text{ is even,}$$

$$P^t = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{if } t \text{ is odd,}$$

for all t.

In this example, P^t never converges to a matrix with both rows identical as t gets large. The chain never 'forgets' its starting conditions as $t \to \infty$.

Exercise: Verify that this Markov chain *does* have an equilibrium distribution, $\boldsymbol{\pi}^T = \left(\frac{1}{2}, \frac{1}{2} \right)$. However, the chain does not *converge* to this distribution as $t \to \infty$.

These examples show that some Markov chains forget their starting conditions in the long term, and ensure that X_t will have the same distribution as $t \to \infty$ regardless of where we started at X_0. However, for other Markov chains, the initial conditions are never forgotten. In the next sections we look for general criteria that will ensure the chain converges.

Target Result:

- If a Markov chain is ***irreducible*** and ***aperiodic***, and if an equilibrium distribution $\boldsymbol{\pi}^T$ ***exists***, then the chain converges to this distribution as $t \to \infty$, regardless of the initial starting states.

To make sense of this, we need to revise the concept of ***irreducibility***, and introduce the idea of ***aperiodicity.***

9.5 Irreducibility

Recall from Chapter 8:

Definition: A Markov chain or transition matrix P is said to be **irreducible** if $i \leftrightarrow j$ *for all* $i, j \in S$. *That is, the chain is irreducible if the state space S is a single communicating class.*

An irreducible Markov chain consists of a single class.
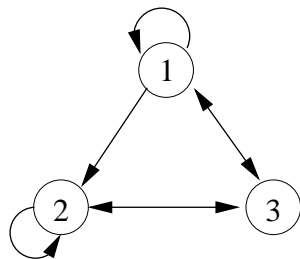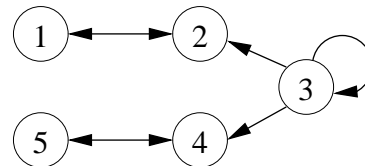
Irreducible   *Not irreducible*

Irreducibility of a Markov chain is important for convergence to equilibrium as $t \to \infty$, because *we want the convergence to be independent of start state.*

This can happen if the chain is irreducible. When the chain is not irreducible, different start states might cause the chain to get stuck in different closed classes. In the example above, a start state of $X_0 = 1$ means that the chain is restricted to states 1 and 2 as $t \to \infty$, whereas a start state of $X_0 = 4$ means that the chain is restricted to states 4 and 5 as $t \to \infty$. A single convergence that 'forgets' the initial state is therefore not possible.

9.6 Periodicity

Consider the Markov chain with transition matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.
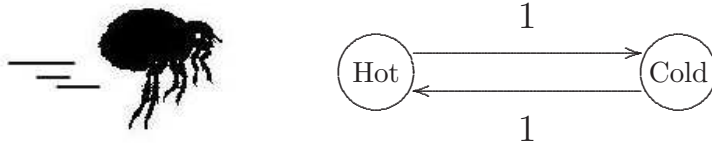
Suppose that $X_0 = 1$.

Then $X_t = 1$ *for all even values of* t, *and* $X_t = 2$ *for all odd values of* t.

This sort of behaviour is called **periodicity:** *the Markov chain can only return to a state at particular values of* t.

Clearly, periodicity of the chain will interfere with convergence to an equilibrium distribution as $t \to \infty$. For example,

$$\mathbb{P}(X_t = 1 \mid X_0 = 1) = \begin{cases} 1 \text{ for even values of } t, \\ 0 \text{ for odd values of } t. \end{cases}$$

Therefore, the probability can not converge to any <u>single</u> value as $t \to \infty$.

Period of state i

To formalize the notion of periodicity, we define the **period** of a state i. Intuitively, *the period is defined so that the time taken to get from state* i *back to state* i *again is always a multiple of the period.*

In the example above, the chain can return to state 1 after *2 steps, 4 steps, 6 steps, 8 steps, ...*

The period of state 1 is therefore *2.*

In general, the chain can return from state i back to state i again in t steps if $(P^t)_{ii} > 0$. This prompts the following definition.

Definition: The **period** $d(i)$ of a state i is

$$d(i) = \gcd\{t : (P^t)_{ii} > 0\},$$

the greatest common divisor of the times at which return is possible.

Definition: The state i is said to be **periodic** if $d(i) > 1$.

For a periodic state i, $(P^t)_{ii} = 0$ if t is not a multiple of $d(i)$.

*Definition:* The state i is said to be **aperiodic** if $d(i) = 1$.

If state i is aperiodic, it means that *return to state i is not limited only to regularly repeating times.*

For convergence to equilibrium as $t \to \infty$, we will be interested only in *aperiodic states.*

The following examples show how to calculate the period for both aperiodic and periodic states.

Examples: Find the periods of the given states in the following Markov chains, and state whether or not the chain is irreducible.

1. The simple random walk.
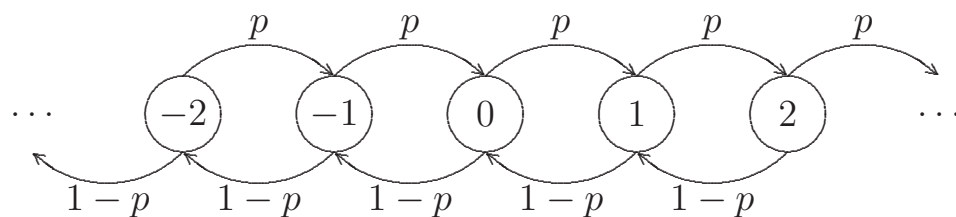
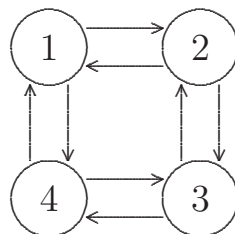$$d(0) = \gcd\{2, 4, 6, \ldots\} = 2.$$

Chain is irreducible.

2.
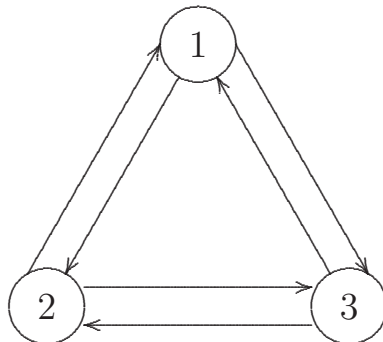
$$d(1) = \gcd\{2, 3, 4, \ldots\} = 1.$$
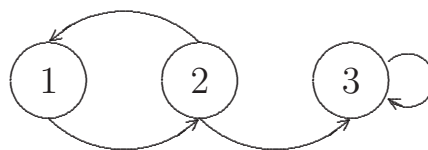
Chain is irreducible.

3.

$$d(1) = \gcd\{2, 4, 6, \ldots\} = 2.$$

Chain is irreducible.

4.

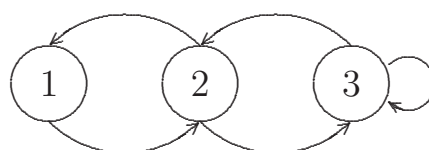$$d(1) = \gcd\{2, 4, 6, \ldots\} = 2.$$

Chain is NOT irreducible (i.e. Reducible).

5.

$$d(1) = \gcd\{2, 4, 5, 6, \ldots\} = 1.$$

Chain is irreducible.

9.7 Convergence to Equilibrium

We now draw together the threads of the previous sections with the following results.

Fact: If $i \leftrightarrow j$, then i and j have the same period. (Proof omitted.)

This leads immediately to the following result:

> If a Markov chain is ***irreducible*** and has ***one*** aperiodic state,
> then ***all*** states are aperiodic.

We can therefore talk about an **irreducible, aperiodic chain,** *meaning that all states are aperiodic.*

Theorem 9.7: Let $\{X_0, X_1, \ldots\}$ be an **irreducible and aperiodic** Markov chain with transition matrix P. Suppose that there ***exists*** an equilibrium distribution π^T. Then, from ***any*** starting state i, and for any end state j,

$$\mathbb{P}(X_t = j \mid X_0 = i) \to \pi_j \quad \text{as } t \to \infty.$$

In particular,
$$\left(P^t\right)_{ij} \to \pi_j \quad \text{as } t \to \infty, \text{ for } \underline{\text{all}} \ i \text{ and } j,$$

so P^t converges to a matrix with all rows identical and equal to π^T. $\qquad \square$

> *For an irreducible, aperiodic Markov chain,*
>
> *with finite or infinite state space,*
>
> *the $\underline{\text{existence}}$ of an equilibrium distribution π^T ensures*
>
> *that the Markov chain will $\underline{\text{converge}}$ to π^T as $t \to \infty$.*

Note: If the state space is infinite, it is not guaranteed that an equilibrium distribution $\boldsymbol{\pi}^T$ exists. See Example 3 below.
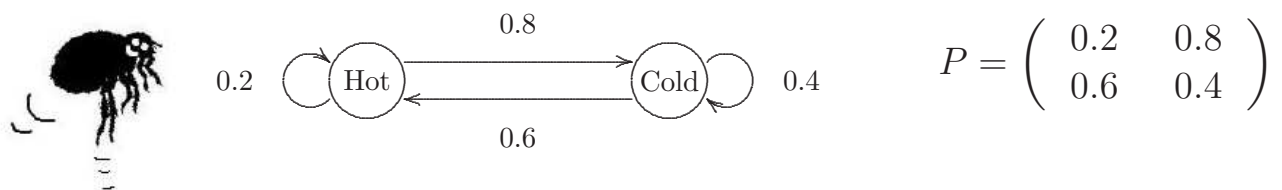
Note: If the chain converges to an equilibrium distribution $\boldsymbol{\pi}^T$ as $t \to \infty$, then *the long-run proportion of time spent in state k is π_k.*

9.8 Examples

A typical exam question gives you a Markov chain on a finite state space and asks if it converges to an equilibrium distribution as $t \to \infty$. An equilibrium distribution will always exist for a finite state space. You need to check whether the chain is irreducible and aperiodic. If so, it will converge to equilibrium. If the chain is *periodic*, it cannot converge to an equilibrium distribution that is independent of start state. If the chain is *reducible*, it may or may not converge.

The first two examples are the same as the ones given in Section 9.4.

Example 1: State whether the Markov chain below converges to an equilibrium distribution as $t \to \infty$.

$$P = \left(\begin{array}{cc} 0.2 & 0.8 \\ 0.6 & 0.4 \end{array} \right)$$

The chain is irreducible and aperiodic, and an equilibrium distribution will exist for a finite state space. So the chain does converge.

(From Section 9.4, the chain converges to $\boldsymbol{\pi}^T = \left(\frac{3}{7}, \frac{4}{7} \right)$ as $t \to \infty$.)

Example 2: State whether the Markov chain below converges to an equilibrium distribution as $t \to \infty$.
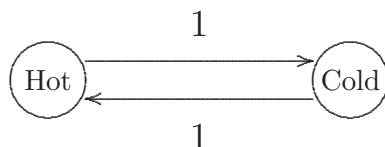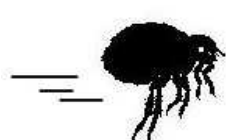

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

The chain is irreducible, but it is NOT aperiodic: period = 2.

Thus the chain does NOT converge to an equilibrium distribution as $t \to \infty$.

It is important to check for aperiodicity, because the existence of an equilibrium distribution does NOT ensure convergence to this distribution if the matrix is not aperiodic.

Example 3: Random walk with retaining barrier at 0.
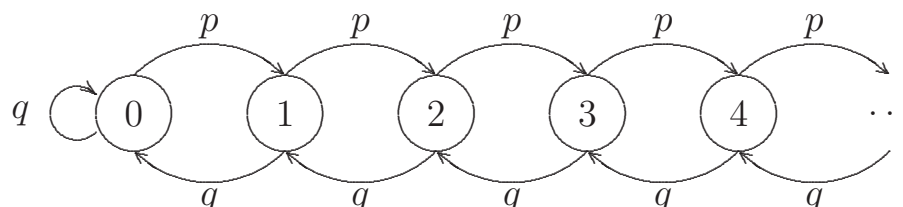
Find whether the chain converges to equilibrium as $t \to \infty$, and if so, find the equilibrium distribution.

The chain is irreducible and aperiodic, so <u>if</u> an equilibrium distribution exists, then the chain will converge to this distribution as $t \to \infty$.

However, the chain has an infinite state space, so we cannot guarantee that an equilibrium distribution exists.

Try to solve the equilibrium equations:

$\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$ and $\sum_{i=0}^{\infty} \pi_i = 1$.

$$P = \begin{pmatrix} q & p & 0 & 0 & \dots \\ q & 0 & p & 0 & \dots \\ 0 & q & 0 & p & \dots \\ \vdots & & & & \end{pmatrix}$$

$$\begin{aligned} q\pi_0 + q\pi_1 &= \pi_0 \qquad (\star) \\ p\pi_0 + q\pi_2 &= \pi_1 \\ p\pi_1 + q\pi_3 &= \pi_2 \\ &\vdots \\ p\pi_{k-1} + q\pi_{k+1} &= \pi_k \quad \text{for} \quad k = 1, 2, \dots \end{aligned}$$

From $(\star)$, we have $p\pi_0 = q\pi_1$,

$$\text{so} \quad \pi_1 = \frac{p}{q}\pi_0$$

$$\Rightarrow \pi_2 = \frac{1}{q}(\pi_1 - p\pi_0) = \frac{1}{q}\left(\frac{p}{q}\pi_0 - p\pi_0\right) = \frac{p}{q}\left(\frac{1-q}{q}\right)\pi_0 = \left(\frac{p}{q}\right)^2 \pi_0.$$

We suspect that $\pi_k = \left(\dfrac{p}{q}\right)^k \pi_0$. Prove by induction.

The hypothesis is true for $k = 0, 1, 2$. Suppose that $\pi_k = \left(\dfrac{p}{q}\right)^k \pi_0$. Then

$$\begin{aligned} \pi_{k+1} &= \frac{1}{q}\left(\pi_k - p\pi_{k-1}\right) \\ &= \frac{1}{q}\left\{ \left(\frac{p}{q}\right)^k \pi_0 - p\left(\frac{p}{q}\right)^{k-1} \pi_0 \right\} \\ &= \frac{p^k}{q^k}\left(\frac{1}{q} - 1\right)\pi_0 \\ &= \left(\frac{p}{q}\right)^{k+1} \pi_0. \end{aligned}$$

The inductive hypothesis holds, so $\pi_k = \left(\dfrac{p}{q}\right)^k \pi_0$ for all $k \geq 0$.

We now need $\displaystyle\sum_{i=0}^{\infty}\pi_i = 1,$ *i.e.* $\displaystyle\pi_0\sum_{k=0}^{\infty}\left(\frac{p}{q}\right)^k = 1.$

The sum is a Geometric series, and converges only for $\left|\dfrac{p}{q}\right| < 1$*. Thus when* $p < q$*, we have*

$$\pi_0\left(\frac{1}{1-\frac{p}{q}}\right) = 1 \quad\Rightarrow\quad \pi_0 = 1 - \frac{p}{q}.$$

If $p \geq q$*, there is no equilibrium distribution.*

Solution:

If $p < q$*, the chain converges to an equilibrium distribution* $\boldsymbol{\pi}$*, where* $\pi_k = \left(1-\frac{p}{q}\right)\left(\frac{p}{q}\right)^k$ *for* $k = 0, 1, \ldots$.

If $p \geq q$*, the chain does not converge to an equilibrium distribution as* $t \to \infty$*.*

Note: Equilibrium results also exist for chains that are ***not*** aperiodic. Also, states can be classified as **transient** (return to the state is not certain), **null recurrent** (return to the state is certain, but the expected return time is infinite), and **positive recurrent** (return to the state is certain, and the expected return time is finite). For each type of state, the long-term behaviour is known:

- If the state k is **transient** or **null-recurrent**,

$$\mathbb{P}(X_t = k \,|\, X_0 = k) = \left(P^t\right)_{kk} \to 0 \text{ as } t \to \infty.$$

- If the state is **positive recurrent**, then

$$\mathbb{P}(X_t = k \,|\, X_0 = k) = \left(P^t\right)_{kk} \to \pi_k \text{ as } t \to \infty, \text{ where } \pi_k > 0.$$

  The expected return time for the state is $1/\pi_k$.

A detailed treatment is available at
`http://www.statslab.cam.ac.uk/~james/Markov/`.