

Part A Probability

Neil Laws

Hilary Term 2009

Acknowledgements

These are lecture notes, not a textbook. I strongly encourage students to look at books for more examples, more details and more background material. In particular I want to acknowledge my use of the excellent books by Grimmett & Welsh, Grimmett & Stirzaker, Stirzaker, and Norris in preparing these notes: the first half follows Grimmett & Welsh in several places; the second half follows Norris in some places, Grimmett & Stirzaker in others. I would also like to acknowledge Peter Donnelly's and Mary Lunn's 2nd year probability notes, which I also used when preparing these notes.

Introductory remarks

Mods Probability covered basic probability concepts (e.g. axioms, conditional probability, independence) and plenty of material on discrete and continuous random variables. We can think of this course being in two parts, about 8 lectures on each.

- More on random variables, especially continuous random variables:
 - we build up more techniques for handling mathematical models involving uncertainty
 - convergence results ('law of averages', Central Limit Theorem).
- Stochastic processes:
 - processes that evolve over time according to probabilistic rules (e.g. the gambler's ruin model)
 - in discrete time we study Markov chains
 - in continuous time we study the Poisson process (1–2 lectures)

- stochastic processes arise in many applications: e.g. stock market prices, population processes, queueing models, . . .
- this part of the subject is taken much further next year in Applied Probability.

Books

The books mentioned in the synopses are:

- Grimmett & Welsh – for lectures 1–8 and Poisson processes
- Grimmett & Stirzaker (more advanced than G&W) – for lectures 1–16 and much more, useful next year – the best buy?
- Stirzaker – for lectures 1–16
- Norris – for lectures 9–16 and more, useful next year.

1 Mods recap

We model an experiment with a ‘random’ outcome using a probability space (Ω, \mathcal{F}, P) where

- (i) the sample space Ω is the set of all possible outcomes
- (ii) \mathcal{F} is a collection of subsets of Ω [satisfying ...] – each element of \mathcal{F} is an event
- (iii) the probability measure P [satisfying ...] assigns a probability in $[0, 1]$ to each event in \mathcal{F} .

We often encode the outcome of our experiment as a number. A random variable (RV), which can be a *discrete* RV, or a *continuous* RV, is a function $X : \Omega \rightarrow \mathbb{R}$ such that

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \quad \text{for all } x \in \mathbb{R}.$$

We write this event as $\{X \leq x\}$.

X Discrete

The probability mass function (pmf) of X is

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\}) \quad \text{for } x \in \mathbb{R}$$

where $p_X(x) > 0$ for only a finite or countable number of values of x . We have

- (a) $p_X(x) \geq 0$ for all x
- (b) $\sum_x p_X(x) = 1$
- (c) $P(X \in A) = \sum_{x \in A} p_X(x)$ for $A \subseteq \mathbb{R}$.

Example. Let X be the number of flips of a coin, with probability p of heads, $q = 1 - p$ of tails, until we get a head. Assume flips independent. The pmf of X is

$$p_X(k) = q^{k-1}p \quad \text{for } k = 1, 2, \dots$$

The (cumulative) distribution function (cdf) of any RV X is given by

$$F_X(x) = P(X \leq x) \quad \text{for } x \in \mathbb{R}.$$

X Continuous

If

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

for some non-negative function f_X , then we say that X is a continuous RV with (probability) density function (pdf) f_X , and clearly $f_X(x) = F'_X(x)$. We have

- (a) $f_X(x) \geq 0$ for all x
- (b) $\int_{-\infty}^{\infty} f_X(x) dx = 1$
- (c) $P(X \in A) = \int_{x \in A} f_X(x) dx$
- (d) $P(X = x) = 0$ for each x .

Example. (Possible model for a lifetime.) X has an exponential distribution with parameter λ (> 0), which we will write as $X \sim \text{Exp}(\lambda)$, if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

We write $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (with $f_X(x) = 0$ for $x < 0$ understood).

Expectation

For any function g , the expectation or expected value of the RV $g(X)$ is

$$E[g(X)] = \begin{cases} \sum_x g(x)p_X(x) & X \text{ discrete} \\ \int_{-\infty}^{\infty} g(x)f_X(x) dx & X \text{ continuous.} \end{cases} \quad (1.1)$$

Remember that $E[g(X)]$ is only defined when $\sum_x |g(x)|p_X(x) < \infty$ or $\int_{-\infty}^{\infty} |g(x)|f_X(x) dx < \infty$.

We sometimes call $E(X)$ the mean of X . The variance of X is

$$\text{var}(X) = E[(X - \mu)^2] = E(X^2) - \mu^2$$

where $\mu = E(X)$.

Example. If $X \sim \text{Exp}(\lambda)$, then

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx = 1/\lambda.$$

So X is ‘exponential with parameter λ ’ or ‘exponential with mean $1/\lambda$ ’. Can check $\text{var}(X) = 1/\lambda^2$.

Simple transformations

If $Y = g(X)$, what can we say about Y ?

We can calculate $E(Y)$ using (1.1).

If X is continuous, then we can calculate the pdf of Y by first finding the cdf of Y , then differentiating.

Example. If $X \sim N(0, 1)$, find the pdf of X^2 .

We know

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

Let $Y = X^2$. Then, for $y \geq 0$,

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

Differentiating with respect to y ,

$$\begin{aligned} f_Y(y) &= f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \frac{1}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi y}} e^{-y/2} \quad \text{for } y \geq 0. \end{aligned}$$

Example. If $X \sim \text{Uniform}(1, 5)$, find the pdf of

$$Y = \frac{X}{5 - X}.$$

Since $X \in (1, 5)$, we have $Y \in (\frac{1}{4}, \infty)$. For $y > \frac{1}{4}$,

$$\begin{aligned} F_Y(y) &= P\left(\frac{X}{5 - X} \leq y\right) \\ &= P\left(X \leq \frac{5y}{1 + y}\right) \\ &= \int_1^{5y/(1+y)} \frac{1}{4} dx \\ &= \frac{1}{4} \left(\frac{5y}{1 + y} - 1\right). \end{aligned}$$

Now differentiate to find

$$f_Y(y) = \frac{5}{4(1 + y)^2} \quad \text{for } \frac{1}{4} < y < \infty.$$

1.1 Joint distributions and independence

(RVs can be discrete or continuous in Section 1.1.)

Given two RVs X and Y (on the probability space (Ω, \mathcal{F}, P)), we often think of (X, Y) as a random vector taking values in \mathbb{R}^2 .

Definition. The *joint (cumulative) distribution function* of X and Y is defined by

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) \quad \text{for } x, y \in \mathbb{R}.$$

The *marginal* distribution of X is

$$F_X(x) = P(X \leq x) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y)$$

and the *marginal* distribution of Y is

$$F_Y(y) = P(Y \leq y) = \lim_{x \rightarrow \infty} F_{X,Y}(x, y).$$

Definition. The RVs X and Y are called *independent* if

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

Otherwise X and Y are called *dependent*.

Given RVs X_1, \dots, X_n , their joint distribution function is defined in a similar way to above:

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, \dots, X_n \leq x_n) \quad \text{for } \mathbf{x} \in \mathbb{R}^n$$

where $\mathbf{x} = (x_1, \dots, x_n)$, and X_1, \dots, X_n are called *independent* if

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1}(x_1) \dots F_{X_n}(x_n) \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

and so on.

2 Jointly continuous random variables

Recall the joint cdf: $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$.

Definition. The pair of RVs X, Y is called (*jointly*) *continuous* if the joint cdf can be written as

$$F_{X,Y}(x, y) = \int_{u=-\infty}^x \int_{v=-\infty}^y f(u, v) du dv$$

for all $x, y \in \mathbb{R}$ and some $f : \mathbb{R}^2 \rightarrow [0, \infty)$.

We call f the joint pdf, we usually write it as $f_{X,Y}$, and we have

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

The joint density satisfies

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1 \quad (2.1)$$

and a consequence of the definition is

$$P((X, Y) \in A) = \iint_A f_{X,Y}(x, y) dx dy \quad \text{for } A \subseteq \mathbb{R}^2. \quad (2.2)$$

Example. Suppose we pick a point ‘at random’ in the square $S = \{(x, y) : -1 \leq x, y \leq 1\}$. ‘At random’ means we want a uniform distribution:

$$f_{X,Y}(x, y) = \begin{cases} c, \text{ constant} & (x, y) \in S \\ 0 & \text{otherwise.} \end{cases}$$

[DIAGRAM.] By (2.1),

$$1 = \int_{-1}^1 \int_{-1}^1 c dx dy = 4c$$

so $c = \frac{1}{4}$. By (2.2),

$$P((X, Y) \in D) = \iint_D \frac{1}{4} dx dy = \frac{1}{4} \text{area}(D) = \frac{\pi}{4}.$$

[Probability calculations reduce to area calculations when we have a uniform distribution.]

2.1 Marginal density functions and independence

The marginal pdf of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$$

and the marginal pdf of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx.$$

Independence

As in Section 1.1, X and Y are called *independent* if and only if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \text{for all } x, y \in \mathbb{R} \quad (2.3)$$

or, equivalently, if and only if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y \in \mathbb{R}. \quad (2.4)$$

We can prove (2.3) \iff (2.4) by differentiation/integration.

Example. As before

$$f_{X,Y}(x, y) = \frac{1}{4} \quad \text{for } (x, y) \in S.$$

[DIAGRAM.] Then, for $-1 \leq x \leq 1$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy = \int_{-1}^1 \frac{1}{4} dy = \frac{1}{2}.$$

Similarly

$$f_Y(y) = \frac{1}{2} \quad \text{for } -1 \leq y \leq 1$$

and X and Y are independent since $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.

Example. Pick a point at random in the disc D . [DIAGRAM.]

$$f_{X,Y}(x, y) = \frac{1}{\pi} \quad \text{for } (x, y) \in D.$$

So, for $-1 \leq x \leq 1$,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \\ &= \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{2}{\pi} \sqrt{1-x^2}. \end{aligned}$$

Similarly

$$f_Y(y) = \frac{2}{\pi} \sqrt{1-y^2} \quad \text{for } -1 \leq y \leq 1.$$

Clearly $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$ so X and Y are not independent.

Example. Suppose $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ are independent. Then their joint pdf is

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \lambda e^{-\lambda x} \mu e^{-\mu y} \quad \text{for } x, y \geq 0.$$

[DIAGRAM.]

$$\begin{aligned} P(Y > X) &= \int_{x=0}^{\infty} \int_{y=x}^{\infty} \lambda \mu e^{-\lambda x} e^{-\mu y} dy dx \quad \text{using (2.2)} \\ &= \frac{\lambda}{\lambda + \mu}. \end{aligned}$$

Let $U = \min(X, Y)$. For $u \geq 0$,

$$\begin{aligned} F_U(u) &= 1 - P(\min(X, Y) > u) \\ &= 1 - P(X > u, Y > u) \\ &= 1 - P(X > u)P(Y > u) \quad \text{by independence} \\ &= 1 - e^{-\lambda u} e^{-\mu u} = 1 - e^{-(\lambda+\mu)u} \end{aligned}$$

and so $U \sim \text{Exp}(\lambda + \mu)$.

We can extend this, by induction, to: if $X_i \sim \text{Exp}(\lambda_i)$, $i = 1, \dots, n$, are independent, then $\min_{1 \leq i \leq n} X_i \sim \text{Exp}(\lambda_1 + \dots + \lambda_n)$.

Exercise. If X and Y are independent and $N(0, 1)$, show that $P(X^2 + Y^2 \leq 1) = 1 - e^{-1/2}$.

Theorem 2.1. *The RVs X and Y are independent if and only if there exist functions g and h such that*

$$f_{X,Y}(x,y) = g(x)h(y) \quad \text{for all } x, y \in \mathbb{R}.$$

Proof. (i) If X and Y are independent, then take $g = f_X$ and $h = f_Y$.

(ii) Conversely, by (2.1),

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) dx dy = \left(\int_{-\infty}^{\infty} g(x) dx \right) \left(\int_{-\infty}^{\infty} h(y) dy \right)$$

and also

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} g(x)h(y) dy = g(x) \int_{-\infty}^{\infty} h(y) dy \\ f_Y(y) &= \int_{-\infty}^{\infty} g(x)h(y) dx = h(y) \int_{-\infty}^{\infty} g(x) dx. \end{aligned}$$

Therefore

$$\begin{aligned} f_X(x)f_Y(y) &= g(x) \left(\int_{-\infty}^{\infty} h(y) dy \right) h(y) \left(\int_{-\infty}^{\infty} h(y) dy \right) \\ &= g(x)h(y) \\ &= f_{X,Y}(x, y). \end{aligned} \quad \square$$

Note: In the ‘uniform in D ’ example, $f_{X,Y}(x, y)$ does *not* factorise as in Theorem 2.1 for *all* $x, y \in \mathbb{R}$ because D is *not* of the form $[a, b] \times [c, d]$.

2.2 Expectation

The expectation of $g(X, Y)$ is

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy$$

(provided $\iint |g(x, y)| f_{X,Y}(x, y) dx dy < \infty$).

The covariance of X and Y is

$$\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X\mu_Y$$

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

3 Conditional distributions and conditional expectation

Recall that if X and Y are discrete RVs, then the conditional distribution of Y given that $X = x$ is defined by

$$p_{Y|X}(y|x) = P(Y = y | X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

and the conditional expectation of Y given $X = x$ is

$$E(Y | X = x) = \sum_y y p_{Y|X}(y|x).$$

Now suppose X and Y are continuous RVs with joint pdf $f_{X,Y}(x,y)$. Instead of conditioning on $X = x$, which has probability zero, we condition on $x \leq X \leq x + \delta x$ and then let $\delta x \rightarrow 0$:

$$\begin{aligned} P(Y \leq y | x \leq X \leq x + \delta x) &= \frac{P(Y \leq y, x \leq X \leq x + \delta x)}{P(x \leq X \leq x + \delta x)} \\ &= \frac{\frac{1}{\delta x} \int_{u=x}^{x+\delta x} \int_{v=-\infty}^y f_{X,Y}(u,v) du dv}{\frac{1}{\delta x} \int_{u=x}^{x+\delta x} f_X(u) du} \\ &\rightarrow \int_{-\infty}^y \frac{f_{X,Y}(x,v)}{f_X(x)} dv \quad \text{as } \delta x \rightarrow 0. \end{aligned}$$

This is a conditional probability that $Y \leq y$, i.e. a cdf. To get the conditional density, differentiate with respect to y .

Definition. The *conditional density function* of Y given that $X = x$ is defined by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad \text{for } y \in \mathbb{R}$$

and the conditional expectation of Y given $X = x$ is

$$E(Y | X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy.$$

This definition is for any x such that $f_X(x) > 0$.

Example. Suppose (X, Y) is uniform within the disc D . [DIAGRAM.] Then

$$f_{X,Y}(x, y) = \frac{1}{\pi} \quad \text{for } (x, y) \in D.$$

From a previous example

$$f_Y(y) = \frac{2}{\pi} \sqrt{1 - y^2} \quad \text{for } -1 \leq y \leq 1.$$

So the conditional density of Y given $X = x$ is

$$f_{Y|X}(y | x) = \frac{1/\pi}{2\sqrt{1 - x^2}/\pi} \quad \text{for } |y| \leq \sqrt{1 - x^2}$$

i.e. a uniform density on $[-\sqrt{1 - x^2}, \sqrt{1 - x^2}]$. [DIAGRAM.]

Clearly $E(Y | X = x) = 0$.

Example. If X and Y are independent then of course $f_{Y|X}(y | x) = f_Y(y)$. (Check.)

Since $f_{X,Y}(x, y) = f_{Y|X}(y | x)f_X(x)$ we have, after integrating over x ,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y | x)f_X(x) dx. \quad (3.1)$$

Integrating (3.1) over $y \in A$ gives

$$P(Y \in A) = \int_{-\infty}^{\infty} P(Y \in A | X = x)f_X(x) dx. \quad (3.2)$$

Similarly, multiplying (3.1) by y and integrating over all y gives

$$E(Y) = \int_{-\infty}^{\infty} E(Y | X = x)f_X(x) dx. \quad (3.3)$$

(3.3) is often written $E(Y) = E[E(Y | X)]$.

(3.1)–(3.3) are all ‘partition results’: condition on x , multiply by $f_X(x)$, and integrate over x .

Example. Suppose $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ are independent. Then, conditioning on X ,

$$\begin{aligned} P(Y > X) &= \int_0^{\infty} P(Y > X | X = x)f_X(x) dx \\ &= \int_0^{\infty} P(Y > x)f_X(x) dx \\ &= \int_0^{\infty} e^{-\mu x} \lambda e^{-\lambda x} dx \\ &= \frac{\lambda}{\lambda + \mu}. \end{aligned}$$

Example. Suppose $-1 < \rho < 1$ and consider the bivariate normal density

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right) \quad \text{for } x, y \in \mathbb{R}.$$

What is the marginal distribution of Y ?

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x, y) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)}[(x - \rho y)^2 + y^2(1-\rho^2)]\right) dx \\ &\hspace{15em} \text{after completing the square in } x \\ &= \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \int_{-\infty}^{\infty} \text{pdf of a } N(\rho y, 1-\rho^2) dx \\ &= \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \quad \text{for } y \in \mathbb{R}. \end{aligned}$$

So $Y \sim N(0, 1)$ and by symmetry $X \sim N(0, 1)$ also.

The conditional density of Y given $X = x$ is

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f(x, y)}{f_X(x)} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(\frac{-1}{2(1-\rho^2)}(y - \rho x)^2\right) \quad \text{for } y \in \mathbb{R}. \end{aligned}$$

This is a $N(\rho x, 1-\rho^2)$ pdf, so given $X = x$ the conditional distribution of Y is $N(\rho x, 1-\rho^2)$.

So $E(Y|X = x) = \rho x$ and often this is written $E(Y|X) = \rho X$.

Since $E(X) = E(Y) = 0$, $\text{cov}(X, Y) = E(XY)$ and

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} E(XY|X = x) f_X(x) dx \quad \text{by (3.3)} \\ &= \int_{-\infty}^{\infty} x \underbrace{E(Y|X = x)}_{\rho x} f_X(x) dx \\ &= \int_{-\infty}^{\infty} \rho x^2 f_X(x) dx \\ &= \rho E(X^2) \\ &= \rho. \end{aligned}$$

So we have

$$\begin{aligned} X \text{ and } Y \text{ are independent} &\iff \rho = 0 \quad \text{by Theorem 2.1 (or defn of indep)} \\ &\iff \text{cov}(X, Y) = 0 \quad \text{from above.} \end{aligned}$$

Note: For general RVs, $\text{cov}(X, Y) = 0$ does *not* imply that X and Y are independent. But for normal RVs, $\text{cov}(X, Y) = 0$ does imply independence.

Example. Let X and Y have joint density function

$$f(x, y) = \begin{cases} 2e^{-x-y} & 0 < x < y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

[DIAGRAM.] We do not have $f(x, y) = g(x)h(y)$ for all $x, y \in \mathbb{R}$, so by Theorem 2.1 X and Y are not independent. E.g. If we are told $Y = y$, then we know $X \in (0, y)$.

The marginal density functions are

$$f_X(x) = \int_{y=x}^{\infty} 2e^{-x-y} dy = 2e^{-2x} \quad \text{for } x > 0$$

$$f_Y(y) = \int_{x=0}^y 2e^{-x-y} dx = 2e^{-y}(1 - e^{-y}) \quad \text{for } y > 0.$$

Given $X = x (> 0)$ the conditional density of Y is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = e^{x-y} \quad \text{for } y \in (x, \infty).$$

Given $Y = y (> 0)$ the conditional density of X is

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{e^{-x}}{1 - e^{-y}} \quad \text{for } x \in (0, y).$$

4 Change of variables

Our real interest may be in some function(s) of X and Y . How do we find the joint pdf of (U, V) where $U = u(X, Y)$ and $V = v(X, Y)$? E.g. $U = X + Y$, $V = X - Y$.

Suppose $f_{X,Y}(x, y)$ is non-zero on $D \subseteq \mathbb{R}^2$. Suppose T is a 1-1 transformation given by

$$T(x, y) = (u, v) \quad \text{where } u = u(x, y), v = v(x, y)$$

and that $S \subseteq \mathbb{R}^2$ is the image of D under the transformation.

T is 1-1, therefore invertible, with say $x = x(u, v)$, $y = y(u, v)$. E.g. $x = (u + v)/2$, $y = (u - v)/2$ above.

Theorem 4.1. *The joint density of (U, V) is given by*

$$f_{U,V}(u, v) = \begin{cases} f_{X,Y}(x(u, v), y(u, v)) |J(u, v)| & (u, v) \in S \\ 0 & \text{otherwise} \end{cases}$$

where $J(u, v)$ is the Jacobian

$$J(u, v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}.$$

Why? Suppose $A \subseteq D$ and that $T(A) = B$. Then

$$\begin{aligned} P((U, V) \in B) &= P((X, Y) \in A) \quad \text{since } T \text{ is 1-1} \\ &= \iint_A f_{X,Y}(x, y) dx dy \quad \text{by (2.2)} \end{aligned}$$

and changing variables in the integral

$$= \iint_B f_{X,Y}(x(u, v), y(u, v)) |J(u, v)| du dv.$$

Looking again at (2.2), the final integrand must be the pdf of (U, V) .

There is a similar result for mappings $\mathbb{R}^n \rightarrow \mathbb{R}^n$.

Example. Suppose X and Y are independent, both $\text{Exp}(\lambda)$. Find the joint density of $U = X + Y$ and $V = X/Y$.

$$f_{X,Y}(x, y) = \lambda e^{-\lambda x} \lambda e^{-\lambda y} = \lambda^2 e^{-\lambda(x+y)} \quad \text{for } x, y \geq 0.$$

Now $u = x + y$ and $v = x/y$ has inverse

$$x = \frac{uv}{1+v}, \quad y = \frac{u}{1+v}.$$

The Jacobian is

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{v}{1+v} & \frac{u}{(1+v)^2} \\ \frac{1}{1+v} & \frac{-u}{(1+v)^2} \end{vmatrix} = \frac{-u}{(1+v)^2}.$$

The image of the region $x, y \geq 0$ is $u, v \geq 0$. So, by Theorem 4.1,

$$\begin{aligned} f_{U,V}(u, v) &= \lambda^2 \exp\left(-\lambda \left[\frac{uv}{1+v} + \frac{u}{1+v}\right]\right) \left| \frac{-u}{(1+v)^2} \right| \\ &= \frac{\lambda^2 u e^{-\lambda u}}{(1+v)^2} \quad \text{for } u, v \geq 0. \end{aligned}$$

Since $f_{U,V}(u, v)$ factorises, U and V are independent by Theorem 2.1.

Example. Take X, Y as in the previous example, and let $U = X - Y$, $V = X + Y$. The inverse transformation is

$$x = \frac{u+v}{2}, \quad y = \frac{v-u}{2}$$

with Jacobian

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{vmatrix} = \frac{1}{2}.$$

Therefore, by Theorem 4.1,

$$\begin{aligned} f_{U,V}(u, v) &= \begin{cases} f_{X,Y}\left(\frac{u+v}{2}, \frac{v-u}{2}\right) |J| & (u, v) \in S \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{2} \lambda^2 e^{-\lambda v} & (u, v) \in S \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where S is the image of $x, y \geq 0$ under the transformation. Now

$$\begin{aligned} x \geq 0 &\iff \frac{u+v}{2} \geq 0 \iff v \geq -u \\ y \geq 0 &\iff \frac{v-u}{2} \geq 0 \iff v \geq u \end{aligned}$$

and so $S = \{(u, v) : v \geq |u|\}$. [DIAGRAM.]

Example (Sums of continuous RVs). For general X, Y consider $U = X + Y$, $V = X$, which has inverse $X = V$, $Y = U - V$ and $|\text{Jacobian}| = 1$. (Check.)

Then $f_{U,V}(u, v) = f_{X,Y}(v, u - v) \cdot 1$ by Theorem 4.1. So the marginal pdf of U is

$$f_U(u) = \int_{-\infty}^{\infty} f_{X,Y}(v, u - v) dv$$

and if X, Y are independent

$$f_U(u) = \int_{-\infty}^{\infty} f_X(v)f_Y(u - v) dv.$$

If X, Y are independent and $\text{Exp}(\lambda)$, then

$$\begin{aligned} f_U(u) &= \int_0^u \lambda e^{-\lambda v} \lambda e^{-\lambda(u-v)} dv \quad \text{note we can restrict to } \int_0^u \\ &= \lambda^2 e^{-\lambda u} \int_0^u dv \\ &= \lambda^2 u e^{-\lambda u} \quad \text{for } u \geq 0. \end{aligned}$$

This is a $\text{Gamma}(2, \lambda)$ density. The $\text{Gamma}(\alpha, \lambda)$ pdf is

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x} \quad \text{for } x \geq 0$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ (and $\Gamma(\alpha) = (\alpha - 1)!$ if α is a positive integer). When $\lambda = \frac{1}{2}$ and $\alpha = \frac{n}{2}$, f is known as the χ^2 density with n degrees of freedom.

Example. If X, Y have joint pdf $f(x, y)$ find the pdf of XY .

Consider $u = xy$, $v = x$, with inverse $x = v$, $y = u/v$:

$$\text{Jacobian} = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{vmatrix} = -\frac{1}{v}.$$

Hence

$$f_{U,V}(u, v) = f\left(v, \frac{u}{v}\right) \frac{1}{|v|}$$

by Theorem 4.1. So the pdf of U is

$$f_U(u) = \int_{-\infty}^{\infty} f\left(v, \frac{u}{v}\right) \frac{1}{|v|} dv.$$

Now suppose X, Y are independent and $\text{Uniform}(0, 1)$. Then

$$f(x, y) = \begin{cases} 1 & x, y \in (0, 1) \\ 0 & \text{otherwise.} \end{cases}$$

The image of the region $\{(x, y) : x, y \in (0, 1)\}$ is $\{(u, v) : 0 < u < v < 1\}$.
(Check.) [DIAGRAM.] So

$$f_U(u) = \int_{v=u}^1 1 \cdot \frac{1}{v} dv = -\log u \quad \text{for } 0 < u < 1.$$

5 Moment generating functions

If X is a discrete RV taking values in $\{0, 1, 2, \dots\}$, then its probability generating function (pgf) is $G_X(s) = E(s^X) = \sum_{k=0}^{\infty} s^k P(X = k)$. For more general RVs we consider a modification of the pgf.

The *moment generating function* (mgf) of a RV X is defined by $M_X(t) = E(e^{tX})$. So

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_k e^{tk} P(X = k) & X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx & X \text{ continuous.} \end{cases}$$

We will only be interested in values of t for which $M_X(t)$ is finite. Obviously $M_X(0) = 1$.

Theorem 5.1. *If $M_X(t) < \infty$ for all t such that $|t| < \delta$, for some $\delta > 0$, then*

(i) *there is a unique distribution with mgf $M_X(t)$*

(ii) $M_X(t) = \sum_{k=0}^{\infty} \frac{1}{k!} t^k E(X^k)$ for $|t| < \delta$

(iii) $E(X^k) = M_X^{(k)}(0)$ where $M_X^{(k)}$ is the k th derivative of M_X .

We call $E(X^k)$ the ' k th moment of X '.

Sketch proof. (i) not proved.

(ii)

$$\begin{aligned} M_X(t) &= E(e^{tX}) = E\left(1 + tX + \frac{1}{2!}(tX)^2 + \dots\right) \\ &= 1 + tE(X) + \frac{1}{2!}t^2E(X^2) + \dots \end{aligned}$$

assuming we can interchange E and \sum .

(iii)

$$\begin{aligned} M_X^{(k)}(t) &= \frac{d^k}{dt^k} \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\ &= \int_{-\infty}^{\infty} x^k e^{tx} f_X(x) dx \end{aligned}$$

assuming we can interchange $\frac{d^k}{dt^k}$ and \int . So

$$M_X^{(k)}(0) = \int_{-\infty}^{\infty} x^k f_X(x) dx = E(X^k). \quad \square$$

Using Theorem 5.1(iii) we have

$$\begin{aligned} E(X) &= M'(0) \\ \text{var}(X) &= M''(0) - [M'(0)]^2. \end{aligned}$$

Example. Suppose $X \sim \text{Exp}(\lambda)$. Then

$$\begin{aligned} M_X(t) &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx \\ &= \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda. \end{aligned}$$

($M_X(t)$ undefined for $t \geq \lambda$.)

$$\begin{aligned} E(X) &= M'_X(0) = \left. \frac{\lambda}{(\lambda - t)^2} \right|_{t=0} = \frac{1}{\lambda} \\ E(X^2) &= M''_X(0) = \left. \frac{2\lambda}{(\lambda - t)^3} \right|_{t=0} = \frac{2}{\lambda^2} \end{aligned}$$

Thus $\text{var}(X) = 2/\lambda^2 - (1/\lambda)^2 = 1/\lambda^2$.

Example. Let $X \sim N(\mu, \sigma^2)$. Then

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= e^{\mu t} \int_{-\infty}^{\infty} e^{y\sigma t} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \quad \text{by the substitution } y = \frac{x - \mu}{\sigma} \\ &= e^{\mu t} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(y-\sigma t)^2 - \sigma^2 t^2]} dy \quad \text{by completing the square} \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \int_{-\infty}^{\infty} \text{pdf of a } N(\sigma t, 1) dy \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2}. \end{aligned} \tag{5.1}$$

In particular, the mgf of a $N(0, 1)$ is $e^{t^2/2}$.

Exercise. If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ show, by calculating the mgf of Y , that $Y \sim N(a\mu + b, a^2\sigma^2)$.

Theorem 5.2. If X and Y are independent RVs with mgfs $M_X(t)$ and $M_Y(t)$, then $X + Y$ has mgf $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Proof.

$$\begin{aligned}
M_{X+Y}(t) &= E(e^{t(X+Y)}) \\
&= E(e^{tX} e^{tY}) \\
&= E(e^{tX})E(e^{tY}) \quad \text{since } X \text{ and } Y \text{ are independent} \\
&= M_X(t)M_Y(t). \quad \square
\end{aligned}$$

Suppose $Y = \sum_{i=1}^n X_i$ where X_1, \dots, X_n are independent. Then Theorem 5.2 extends to $M_Y(t) = \prod_{i=1}^n M_{X_i}(t)$.

Example. Suppose X and Y are independent with $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. If $Z = aX + bY$ then $Z \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.

$$\begin{aligned}
M_Z(t) &= E(e^{t(aX+bY)}) \\
&= E(e^{atX} e^{btY}) \\
&= E(e^{(at)X})E(e^{(bt)Y}) \quad \text{since } X \text{ and } Y \text{ are independent} \\
&= M_X(at)M_Y(bt) \\
&= e^{\mu_1(at) + \frac{1}{2}\sigma_1^2(at)^2} e^{\mu_2(bt) + \frac{1}{2}\sigma_2^2(bt)^2} \quad \text{using (5.1) twice} \\
&= e^{(a\mu_1 + b\mu_2)t + \frac{1}{2}(a^2\sigma_1^2 + b^2\sigma_2^2)t^2}
\end{aligned}$$

which is the mgf of a $N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$ by (5.1), and hence this is the distribution of Z by Theorem 5.1(i).

This result can be generalised to any linear combination of n independent normals.

Example. Let N be a non-negative integer valued RV with pgf $G_N(s) = \sum_{n=0}^{\infty} s^n P(N = n)$. Let X_1, X_2, \dots each have mgf $M_X(t)$. Assume N, X_1, X_2, \dots are independent.

Define $S = X_1 + \dots + X_N$ (a sum of a random number of RVs).

$$\begin{aligned}
M_S(t) &= E(e^{tS}) \\
&= \sum_{n=0}^{\infty} E(e^{tS} | N = n)P(N = n) \\
&= \sum_{n=0}^{\infty} E(e^{t(X_1 + \dots + X_n)})P(N = n) \\
&= \sum_{n=0}^{\infty} [M_X(t)]^n P(N = n) \\
&= G_N(M_X(t)).
\end{aligned}$$

Exercise. If $X_i \sim \text{Exp}(\lambda)$ and $P(N = k) = q^{k-1}p$ for $k \geq 1$, show that $S \sim \text{Exp}(\lambda p)$.

5.1 *Characteristic functions

Example. For a Cauchy RV, which has pdf

$$f_X(x) = \frac{1}{\pi(1+x^2)} \quad \text{for } -\infty < x < \infty,$$

$M_X(t)$ is undefined for $t \neq 0$.

The *characteristic function* (CF) of a RV X is defined by

$$\phi_X(t) = E(e^{itX}) = E(\cos tX) + iE(\sin tX) \quad \text{for } t \in \mathbb{R}$$

where $i = \sqrt{-1}$.

An advantage of the CF is that it is always defined for all $t \in \mathbb{R}$. In fact $|\phi_X(t)| \leq 1$ for all t .

When $M_X(t)$ is finite for $|t| < \delta$, for some $\delta > 0$,

$$\phi_X(t) = M_X(it) \quad \text{for } t \in \mathbb{R}.$$

Example. If X is Cauchy, then

$$\phi_X(t) = E(e^{itX}) = \int_{-\infty}^{\infty} \frac{e^{itx}}{\pi(1+x^2)} dx = e^{-|t|}.$$

6 Convergence

Suppose we have independent and identically distributed (i.i.d.) RVs X_1, X_2, \dots and let $\mu = E(X_i)$. We would like to prove that the average $\frac{1}{n}(X_1 + \dots + X_n)$ converges to μ as $n \rightarrow \infty$.

Theorem 6.1 (Markov's inequality). *If Y is any non-negative RV, then*

$$P(Y \geq a) \leq \frac{1}{a}E(Y) \quad \text{for all } a > 0.$$

Proof.

$$\begin{aligned} E(Y) &= E(Y | Y \geq a)P(Y \geq a) + \underbrace{E(Y | Y < a)P(Y < a)}_{\text{non-neg}} \quad \text{by partition thm} \\ &\geq E(Y | Y \geq a)P(Y \geq a) \\ &\geq aP(Y \geq a) \quad \text{since } E(Y | Y \geq a) \geq a. \quad \square \end{aligned}$$

If Y can take negative values the inequality becomes (same proof)

$$P(|Y| \geq a) \leq \frac{1}{a}E(|Y|).$$

For any event A , the *indicator function* of A is defined by

$$I(A) = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

So $I(A)$ is a discrete RV and

$$E[I(A)] = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A).$$

2nd proof of Markov's inequality. Let $A = \{|Y| \geq a\}$. Then

$$|Y| \geq aI(A)$$

and taking expectations gives

$$E(|Y|) \geq aP(|Y| \geq a). \quad \square$$

Theorem 6.2 (Chebyshev's inequality). *If Y is any RV, then*

$$P(|Y| \geq a) \leq \frac{1}{a^2}E(Y^2) \quad \text{for all } a > 0.$$

Proof.

$$\begin{aligned} P(|Y| \geq a) &= P(Y^2 \geq a^2) \\ &\leq \frac{1}{a^2} E(Y^2) \quad \text{by Markov's inequality.} \quad \square \end{aligned}$$

If X is any RV and $\mu = E(X)$, then applying Chebyshev to $Y = X - \mu$ gives

$$P(|X - \mu| \geq a) \leq \frac{1}{a^2} \text{var}(X). \quad (6.1)$$

Example. Let $Y \sim \text{Binomial}(100, \frac{1}{2})$, then $E(Y) = 50$ and $\text{var}(Y) = 100 \cdot \frac{1}{2} \cdot \frac{1}{2} = 25$. So

$$\begin{aligned} P(41 \leq Y \leq 59) &= P(|Y - 50| < 10) \\ &= 1 - P(|Y - 50| \geq 10) \\ &\geq 1 - \frac{\text{var}(Y)}{100} \quad \text{by (6.1)} \\ &= \frac{3}{4}. \end{aligned}$$

In fact, $P(41 \leq Y \leq 59) = 0.94\dots$

Definition. We say that the sequence of RVs X_1, X_2, \dots *converges in probability to X* as $n \rightarrow \infty$ if, for all $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

We write $X_n \xrightarrow{P} X$. (The limit X may be a RV, or a constant.)

Example. Suppose $P(X_n = 0) = 1 - \frac{1}{n}$, $P(X_n = 1) = \frac{1}{n}$.

Then for $\epsilon > 0$, $P(|X_n - 0| > \epsilon) \leq \frac{1}{n} \rightarrow 0$. Hence $X_n \xrightarrow{P} 0$.

Theorem 6.3 (Weak law of large numbers, WLLN). *Suppose X_1, X_2, \dots are i.i.d. RVs, each having mean μ and variance σ^2 . Then*

$$\frac{1}{n}(X_1 + \dots + X_n) \xrightarrow{P} \mu \quad \text{as } n \rightarrow \infty.$$

Proof. Let $S_n = \sum_{i=1}^n X_i$ and note $E(\frac{1}{n}S_n) = \mu$. So

$$\begin{aligned} P\left(\left|\frac{1}{n}S_n - \mu\right| > \epsilon\right) &\leq \frac{1}{\epsilon^2} \text{var}\left(\frac{1}{n}S_n\right) \quad \text{by (6.1)} \\ &= \frac{1}{n^2\epsilon^2} \text{var}(S_n) \\ &= \frac{1}{n^2\epsilon^2} n\sigma^2 \quad \text{since } \text{var}(S_n) = \sum_{i=1}^n \text{var}(X_i) = n\sigma^2 \\ &\rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad \square \end{aligned}$$

We also consider other types of convergence.

Definition. As $n \rightarrow \infty$ we say that

(a) X_n converges to X in mean square, written $X_n \xrightarrow{\text{m.s.}} X$, if

$$E[(X_n - X)^2] \rightarrow 0$$

(b) X_n converges to X in distribution, written $X_n \xrightarrow{D} X$, if

$$P(X_n \leq x) \rightarrow P(X \leq x)$$

for all points x at which the cdf $F_X(x) = P(X \leq x)$ is continuous.

Example. Suppose $Y_n \sim \text{Binomial}(n, p)$ and let $X_n = \frac{1}{n}Y_n$.

$E[(X_n - p)^2] = \text{var}(X_n) = \frac{1}{n^2}npq \rightarrow 0$. Therefore $X_n \xrightarrow{\text{m.s.}} p$.

Example. Suppose $P(X_n = 0) = 1 - \frac{1}{n}$, $P(X_n = n) = \frac{1}{n}$.

As earlier $X_n \xrightarrow{P} 0$. But $E[(X_n - 0)^2] = n^2\frac{1}{n} + 0^2(1 - \frac{1}{n}) \not\rightarrow 0$, so X_n does not converge in m.s. to 0.

Example. Suppose $P(U = 1) = P(U = -1) = \frac{1}{2}$ and let

$$X_n = \begin{cases} U & n \text{ odd} \\ -U & n \text{ even.} \end{cases}$$

Then $X_n \xrightarrow{D} U$ as all X_n 's have the same distribution.

For n even, $X_n - U = -2U$. Therefore

$$P(|X_{2m} - U| > 1) = P(|U| > \frac{1}{2}) = 1 \quad \text{for all } m$$

so X_n does not converge to U in probability.

Theorem 6.4. $(X_n \xrightarrow{\text{m.s.}} X) \implies (X_n \xrightarrow{P} X) \implies (X_n \xrightarrow{D} X)$.

The reverse implications do not hold in general (see the examples above).

Proof. (i) Suppose $X_n \xrightarrow{\text{m.s.}} X$ and let $\epsilon > 0$. Then

$$\begin{aligned} P(|X_n - X| > \epsilon) &\leq \frac{1}{\epsilon^2} E[(X_n - X)^2] \quad \text{by Chebyshev} \\ &\rightarrow 0 \quad \text{since } X_n \xrightarrow{\text{m.s.}} X. \end{aligned}$$

Hence $X_n \xrightarrow{P} X$.

(ii) Suppose $X_n \xrightarrow{P} X$ and let

$$F_n(x) = P(X_n \leq x), \quad F(x) = P(X \leq x).$$

Then, if $\epsilon > 0$,

$$\begin{aligned} F_n(x) &= P(X_n \leq x) \\ &= \underbrace{P(X_n \leq x, X \leq x + \epsilon)}_{\leq P(X \leq x + \epsilon)} + \underbrace{P(X_n \leq x, X > x + \epsilon)}_{\leq P(X - X_n > \epsilon)} \\ &\leq F(x + \epsilon) + P(|X_n - X| > \epsilon). \end{aligned}$$

Similarly

$$\begin{aligned} F(x - \epsilon) &= P(X \leq x - \epsilon) \\ &= P(X \leq x - \epsilon, X_n \leq x) + P(X \leq x - \epsilon, X_n > x) \\ &\leq F_n(x) + P(|X_n - X| > \epsilon). \end{aligned}$$

Therefore

$$F(x - \epsilon) - P(|X_n - X| > \epsilon) \leq F_n(x) \leq F(x + \epsilon) + P(|X_n - X| > \epsilon).$$

Letting $n \rightarrow \infty$, and then $\epsilon \rightarrow 0$, we have $F_n(x) \rightarrow F(x)$ at points of continuity of F . Hence $X_n \xrightarrow{D} X$. \square

7 Convergence and the Central Limit Theorem

Suppose X_1, X_2, \dots are i.i.d., each having mean μ and non-zero variance σ^2 . The WLLN says $\frac{1}{n}S_n \xrightarrow{P} \mu$ where $S_n = X_1 + \dots + X_n$. So, for large n , S_n is about as big as $n\mu$. What can we say about the difference $S_n - n\mu$? It turns out that this difference has order \sqrt{n} .

We work with the so-called *standardised version* of S_n defined by

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Note that $E(S_n) = n\mu$ and, since the X_i are independent, $\text{var}(S_n) = \sum_{i=1}^n \text{var}(X_i) = n\sigma^2$, so

$$\begin{aligned} E(Z_n) &= \frac{1}{\sigma\sqrt{n}}[E(S_n) - n\mu] = 0 \\ \text{var}(Z_n) &= \frac{1}{\sigma^2 n} \text{var}(S_n - n\mu) = \frac{1}{\sigma^2 n} \text{var}(S_n) = 1. \end{aligned}$$

From Section 5, any linear combination of independent normals is normal, so if $X_i \sim N(\mu, \sigma^2)$ then $Z_n \sim N(0, 1)$ for all n .

Theorem 7.1 (Central Limit Theorem, CLT). *Suppose X_1, X_2, \dots are i.i.d. RVs, each having mean μ and non-zero variance σ^2 . Let*

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \quad \text{where } S_n = X_1 + \dots + X_n.$$

Then, as $n \rightarrow \infty$,

$$P(Z_n \leq x) \rightarrow \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \quad \text{for } x \in \mathbb{R}.$$

So the CLT says $Z_n \xrightarrow{D} N(0, 1)$ as $n \rightarrow \infty$ *irrespective* of the original distribution of the X_i 's.

Theorem 7.2 (Continuity Theorem for mgfs). *Let Z and Z_1, Z_2, \dots be RVs such that, for some $\delta > 0$, $M_Z(t)$ and $M_{Z_1}(t), M_{Z_2}(t), \dots$ are all finite for $|t| < \delta$. Suppose that, as $n \rightarrow \infty$,*

$$M_{Z_n}(t) \rightarrow M_Z(t) \quad \text{for } |t| < \delta.$$

Then $Z_n \xrightarrow{D} Z$ as $n \rightarrow \infty$.

Proof of CLT. Let $Y_i = (X_i - \mu)/\sigma$. Then $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$. So

$$\begin{aligned} M_{Z_n}(t) &= E\left(e^{\frac{t}{\sqrt{n}} \sum_{i=1}^n Y_i}\right) \\ &= E(e^{\frac{t}{\sqrt{n}} Y_1}) \cdots E(e^{\frac{t}{\sqrt{n}} Y_n}) \quad \text{by independence of } Y_1, \dots, Y_n \\ &= \left[M_Y\left(\frac{t}{\sqrt{n}}\right) \right]^n \end{aligned} \tag{7.1}$$

where M_Y is mgf of each Y_i .

By Theorem 5.1(ii),

$$\begin{aligned} M_Y(s) &= 1 + sE(Y) + \frac{1}{2}s^2E(Y^2) + o(s^2) \\ &= 1 + \frac{1}{2}s^2 + o(s^2). \end{aligned}$$

Substituting this into (7.1) with $s = \frac{t}{\sqrt{n}}$ and t fixed,

$$\begin{aligned} M_{Z_n}(t) &= \left[1 + \frac{1}{2} \left(\frac{t}{\sqrt{n}} \right)^2 + o\left(\frac{1}{n}\right) \right]^n \\ &= \left[1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \right]^n \\ &\rightarrow e^{t^2/2} \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Since $e^{t^2/2}$ is the mgf of a $N(0, 1)$ distribution, by Theorem 7.2 we have $Z_n \xrightarrow{D} N(0, 1)$. \square

In the above proof we have assumed that mgfs exist: the CLT is true without the existence of mgfs.

Example. Suppose $S_n \sim \text{Binomial}(n, p)$. This is the case $P(X_i = 1) = p$, $P(X_i = 0) = q = 1 - p$. So $\mu = E(X_i) = p$ and $\sigma^2 = \text{var}(X_i) = pq$.

By the CLT

$$\frac{S_n - np}{\sqrt{npq}} \approx N(0, 1).$$

This is called the normal approximation to the binomial: it applies as $n \rightarrow \infty$ with p constant.

Example. Let Z have Poisson distribution with mean λ ,

$$P(Z = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \dots$$

Then the mgf of Z is

$$\begin{aligned}
 M_Z(t) &= E(e^{tZ}) = \sum_{k=0}^{\infty} e^{tk} \frac{e^{-\lambda} \lambda^k}{k!} \\
 &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} \\
 &= \exp(-\lambda + \lambda e^t) \\
 &= \exp(\lambda[e^t - 1]).
 \end{aligned} \tag{7.2}$$

If $Z_n \sim \text{Binomial}(n, \frac{\lambda}{n})$ then

$$\begin{aligned}
 E(e^{tZ_n}) &= \sum_{k=0}^n e^{tk} \binom{n}{k} p^k q^{n-k} \quad \text{where } p = 1 - q = \frac{\lambda}{n} \\
 &= \sum_{k=0}^n \binom{n}{k} (pe^t)^k q^{n-k} \\
 &= (pe^t + q)^n \\
 &= \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n \\
 &\rightarrow \exp(\lambda[e^t - 1]) \quad \text{as } n \rightarrow \infty.
 \end{aligned} \tag{7.3}$$

Comparing (7.2) and (7.3) we see $Z_n \xrightarrow{D} \text{Poisson}(\lambda)$ by the Continuity Theorem, i.e. $Z_n \approx \text{Poisson}(\lambda)$.

This is the Poisson approximation to the binomial: it applies as $n \rightarrow \infty$ with $np = \text{constant}$.

Example. Let Y_n be a geometric RV with success probability $p = \frac{\lambda}{n}$ so that

$$P(Y_n = k) = q^{k-1} p \quad \text{for } k = 1, 2, \dots$$

and

$$M_{Y_n}(t) = \sum_{k=1}^{\infty} e^{tk} q^{k-1} p = \frac{pe^t}{1 - qe^t}.$$

Let $X_n = \frac{1}{n}Y_n$, which has mgf

$$\begin{aligned}M_{X_n}(t) &= E(e^{\frac{t}{n}Y_n}) \\&= M_{Y_n}\left(\frac{t}{n}\right) \\&= \frac{\frac{\lambda}{n}e^{\frac{t}{n}}}{1 - \left(1 - \frac{\lambda}{n}\right)e^{\frac{t}{n}}} \quad \text{using } M_{Y_n} \text{ with } p = \frac{\lambda}{n} \\&= \frac{\lambda e^{\frac{t}{n}}}{\lambda e^{\frac{t}{n}} - n(e^{\frac{t}{n}} - 1)} \\&\rightarrow \frac{\lambda}{\lambda - t} \quad \text{as } n \rightarrow \infty.\end{aligned}$$

Therefore, by the Continuity Theorem, $X_n \xrightarrow{D} \text{Exp}(\lambda)$.

8 Examples

Example. Let $P(X = 1) = P(X = -1) = \frac{1}{2}$. Suppose Y_1, Y_2, \dots are independent, Y_k having the same distribution as $2^{-k}X$, and let $Z_n = \sum_{k=1}^n Y_k$. What is the limiting distribution of Z_n as $n \rightarrow \infty$?

$Z_n \in [-1, 1]$ and Z_n is equally likely to be any of

$$\pm \frac{(2m-1)}{2^n} \quad \text{for } m \in \{1, 2, \dots, 2^{n-1}\}.$$

(Check. $Z_1 = \pm \frac{1}{2}$, $Z_2 = \pm \frac{1}{2} \pm \frac{1}{4}$, ...) So we might guess that $Z_n \xrightarrow{D} U$ as $n \rightarrow \infty$ where $U \sim \text{Uniform}[-1, 1]$.

The pdf of U is

$$f_U(u) = \frac{1}{2} \quad \text{for } -1 \leq u \leq 1$$

and so the mgf is

$$M_U(t) = \int_{-1}^1 e^{tu} \frac{1}{2} du = \frac{1}{2} \left[\frac{e^{tu}}{t} \right]_{-1}^1 = \frac{e^t - e^{-t}}{2t} = \frac{\sinh t}{t}. \quad (8.1)$$

The mgf of Y_k is

$$M_{Y_k}(t) = E\left(e^{\frac{t}{2^k}X}\right) = \frac{1}{2}e^{\frac{t}{2^k}} + \frac{1}{2}e^{-\frac{t}{2^k}} = \cosh\left(\frac{t}{2^k}\right).$$

So the mgf of Z_n is

$$\begin{aligned} M_{Z_n}(t) &= E(e^{t(Y_1 + \dots + Y_n)}) \\ &= \prod_{k=1}^n M_{Y_k}(t) \quad \text{since the } Y_k \text{ are independent} \\ &= \prod_{k=1}^n \cosh\left(\frac{t}{2^k}\right). \end{aligned}$$

We'd like to consider $n \rightarrow \infty$.

$$\begin{aligned} \sinh t &= 2 \sinh\left(\frac{t}{2}\right) \cosh\left(\frac{t}{2}\right) \\ &= 2 \cosh\left(\frac{t}{2}\right) 2 \sinh\left(\frac{t}{4}\right) \cosh\left(\frac{t}{4}\right) \\ &= 2^n \cosh\left(\frac{t}{2}\right) \cosh\left(\frac{t}{4}\right) \dots \cosh\left(\frac{t}{2^n}\right) \sinh\left(\frac{t}{2^n}\right) \end{aligned}$$

Hence

$$M_{Z_n}(t) = \frac{\sinh t}{2^n \sinh\left(\frac{t}{2^n}\right)}$$

and $\sinh x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \dots$ so

$$M_{Z_n}(t) = \frac{\sinh t}{2^n \left(\frac{t}{2^n} + \frac{t^3}{3!2^{3n}} + \dots\right)} \rightarrow \frac{\sinh t}{t} \quad \text{as } n \rightarrow \infty. \quad (8.2)$$

From (8.1) and (8.2), by the Continuity Theorem, $Z_n \xrightarrow{D} U$ as $n \rightarrow \infty$.

*Characteristic functions

Recall that the CF of X is $\phi_X(t) = E(e^{itX})$ for $t \in \mathbb{R}$.

(i) X and Y have the same distribution if and only if $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$.

(ii) If $E(|X|^N) < \infty$, then

$$\phi_X(t) = \sum_{k=0}^N \frac{1}{k!} (it)^k E(X^k) + o(t^N)$$

and $\phi_X^{(k)}(0) = i^k E(X^k)$ for $k = 1, \dots, N$.

(iii) If X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

(iv) Let Z, Z_1, Z_2, \dots have CFs $\phi, \phi_1, \phi_2, \dots$. Then $Z_n \xrightarrow{D} Z$ as $n \rightarrow \infty$ if and only if

$$\phi_n(t) \rightarrow \phi(t) \quad \text{for all } t \in \mathbb{R}.$$

Our proof of the CLT assumed mgfs existed. The CLT holds whether or not mgfs exist: use CFs and (iv) in place of mgfs.

Example. Suppose X_1, X_2, \dots are i.i.d. with Cauchy pdf $f(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$. Then

$$\phi_{X_1}(t) = e^{-|t|}.$$

Consider $Z_n = \frac{1}{n}(X_1 + \dots + X_n)$.

$$\phi_{Z_n}(t) = E(e^{i\frac{t}{n}(X_1 + \dots + X_n)}) = \left[\phi_{X_1}\left(\frac{t}{n}\right) \right]^n = \left[e^{-\frac{|t|}{n}} \right]^n = e^{-|t|}.$$

Therefore Z_n has a Cauchy distribution for all n .

Note the WLLN does not apply here since the mean and variance of a Cauchy RV are undefined.*

Example. Suppose X and Y are independent Poisson RVs each with parameter n . Show that

$$P(X - Y \leq x\sqrt{2n}) \rightarrow \Phi(x) \quad \text{as } n \rightarrow \infty$$

where Φ is the $N(0, 1)$ cdf.

Let $U_1, U_2, \dots, V_1, V_2, \dots$ be i.i.d. Poisson with parameter 1. Then $\sum_{i=1}^n U_i$ is Poisson with parameter n (use pgfs, mgfs, or $U_1 + U_2 \sim$ Poisson and induction). Therefore

$$\begin{aligned} P(X - Y \leq x\sqrt{2n}) &= P\left(\sum_1^n U_i - \sum_1^n V_i \leq x\sqrt{2n}\right) \\ &= P\left(\frac{\sum_1^n W_i}{\sqrt{2n}} \leq x\right) \end{aligned}$$

where $W_i = X_i - Y_i$.

Now $E(W_i) = E(U_i) - E(V_i) = 0$, $\text{var}(W_i) = \text{var}(U_i) + \text{var}(V_i) = 2$. Therefore, by the CLT,

$$\begin{aligned} P\left(\frac{\sum_1^n W_i}{\sqrt{2n}} \leq x\right) &\rightarrow P((N(0, 1) \leq x) \\ &= \Phi(x). \end{aligned}$$

Example. Suppose X_1, \dots, X_n are independent with mgfs $M_1(t), \dots, M_n(t)$. Find RVs with the following mgfs.

(a) $M_1(t) \times \dots \times M_n(t)$

answer = $X_1 + \dots + X_n$

(b) $M_1(t)^2$

answer = $X_1 + X'_1$, where X_1, X'_1 are i.i.d.

(c) $e^{at}M_1(bt)$

answer = $a + bX$

(d) $\sum_{j=1}^n p_j M_j(t)$ where $p_j \geq 0$, $\sum_1^n p_j = 1$

answer = X_N where N is a RV independent of X_1, \dots, X_n with $P(N = j) = p_j$.

$$\begin{aligned} E(e^{tX_N}) &= \sum_j E(e^{tX_N} | N = j)P(N = j) \\ &= \sum_j E(e^{tX_j})P(N = j) \end{aligned}$$

$$(e) \int_0^\infty M_1(yt)e^{-y} dy$$

answer = $Y X_1$ where $Y \sim \text{Exp}(1)$ is independent of X_1 .

$$\begin{aligned} E(e^{tYX_1}) &= \int_0^\infty E(e^{tYX_1} | Y = y) f_Y(y) dy \\ &= \int_0^\infty E(e^{tyX_1}) e^{-y} dy \\ &= \int_0^\infty M_1(ty) e^{-y} dy \end{aligned}$$

9 Markov chains

Mods reminder: if we fix B and define $Q(A) = P(A|B)$ for all A , then Q is itself a probability measure.

Introduction

Examples, motivation, diagrams, etc.

9.1 Definition and basic properties

Let S be a countable set. Each $i \in S$ is called a *state* and S is called the *state space*. (S will usually be a subset of \mathbb{Z} .)

A *probability distribution* on S is a row vector $\lambda = (\lambda_i : i \in S)$ such that $\lambda_i \geq 0$ for all i and $\sum_{i \in S} \lambda_i = 1$.

A RV Y taking values in S is said to have *distribution* λ if

$$P(Y = i) = \lambda_i \quad \text{for } i \in S.$$

Let $X = (X_n)_{n \geq 0} = (X_0, X_1, X_2, \dots)$ be a sequence of RVs taking values in S . We say that X is a *discrete-time stochastic process*.

Definition. The process X is called a *Markov chain* if

$$P(X_{n+1} = i_{n+1} | X_n = i_n, \dots, X_0 = i_0) = P(X_{n+1} = i_{n+1} | X_n = i_n).$$

for all $n \geq 0$ and all $i_0, i_1, \dots, i_{n+1} \in S$.

We think of X as modelling a random system, with possible states S , where X_n denotes the state of the system at time n . E.g. X_n = the price of a commodity, the size of a population, the length of a queue, \dots [DIAGRAM.]

A Markov chain X is *homogeneous* if

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i) = p_{ij}.$$

for all $n \geq 0$ and all $i, j \in S$. From now on, we restrict to this case.

In words: for $n \geq 0$, conditional on $X_n = i$, X_{n+1} has distribution $(p_{ij} : j \in S)$ and is independent of X_0, \dots, X_{n-1} .

The *transition matrix* $P = (p_{ij})$ is the $|S| \times |S|$ matrix of *transition probabilities* p_{ij} . (Sometimes we write $p_{i,j}$.) We have

$$p_{ij} \geq 0 \quad \text{for all } i, j$$
$$\sum_{j \in S} p_{ij} = \sum_{j \in S} P(X_1 = j | X_0 = i) = P(X_1 \in S | X_0 = i) = 1 \quad \text{for all } i.$$

Any P with these properties is called a *stochastic matrix* (non-negative entries and row sums equal to one.)

Example (Ehrenfest model of diffusion). [DIAGRAM.] Suppose m gas molecules are distributed between containers A and B . At each time $n = 1, 2, \dots$, one molecule is picked at random from the m available:

- if it is in A , it moves to B with probability α
- if it is in B , it moves to A with probability β
- otherwise the molecules stay where they are.

Let X_n be the number of molecules in A after n units of time. Then X is a Markov chain on $S = \{0, 1, \dots, m\}$ with transition probabilities

$$\begin{aligned} p_{i,i+1} &= \beta \frac{(m-i)}{m} \\ p_{i,i-1} &= \alpha \frac{i}{m} \\ p_{ii} &= 1 - \alpha \frac{i}{m} - \beta \frac{(m-i)}{m} \\ p_{ij} &= 0 \quad \text{otherwise.} \end{aligned}$$

Example (Simple symmetric random walk on \mathbb{Z}^d). [DIAGRAM.] A particle jumps to each of its $2d$ neighbours with equal probability. So $S = \mathbb{Z}^d$ and

$$p_{ij} = \begin{cases} \frac{1}{2d} & \text{if } i \text{ and } j \text{ are neighbours} \\ 0 & \text{otherwise.} \end{cases}$$

Example (Simple random walk on \mathbb{Z}). [DIAGRAM.] $S = \mathbb{Z}$ and

$$\begin{aligned} p_{i,i+1} &= p \\ p_{i,i-1} &= q = 1 - p \\ p_{ij} &= 0 \quad \text{otherwise.} \end{aligned}$$

Lemma 9.1. *If $A, B, \{C_k\}$ are events with $\{C_k\}$ a partition of the sample space Ω and if $P(A | B \cap C_k) = \alpha$ for all k , then $P(A | B) = \alpha$.*

Proof. We have $P(A \cap B \cap C_k) = \alpha P(B \cap C_k)$. Now sum over k to get $P(A \cap B) = \alpha P(B)$. □

If X is a Markov chain and

$$P(X_0 = i) = \lambda_i \quad \text{for } i \in S$$

then λ is called the *initial distribution* of X . We say that X is Markov(λ, P) if X is a Markov chain with initial distribution λ and transition matrix P .

Theorem 9.2. X is Markov(λ, P) if and only if

$$P(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$

for all $n \geq 0$ and all $i_0, i_1, \dots, i_n \in S$.

Proof. \implies : Recall that

$$P(A_0 \cap A_1 \cap \cdots \cap A_n) = P(A_0)P(A_1 | A_0) \cdots P(A_n | A_0 \cap \cdots \cap A_{n-1}).$$

Let $A_r = \{X_r = i_r\}$. Then

$$\begin{aligned} P(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) &= P(X_0 = i_0)P(X_1 = i_1 | X_0 = i_0) \\ &\quad \cdots P(X_n = i_n | X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\ &= \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}. \end{aligned}$$

\impliedby : Take $n = 0$ to see the initial distribution of X is λ . For $n \geq 0$,

$$\begin{aligned} P(\underbrace{X_{n+1} = i_{n+1}}_A | \underbrace{X_n = i_n}_B, \underbrace{\dots, X_0 = i_0}_{C_k}) &= \frac{\lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_n i_{n+1}}}{\lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}} \\ &= p_{i_n i_{n+1}}. \end{aligned}$$

So $P(A | B) = p_{i_n i_{n+1}}$ by Lemma 9.1 and X is Markov(λ, P). \square

Let

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

and let $\delta_i = (\delta_{ij} : j \in S) = (0, \dots, 0, 1, 0, \dots, 0)$ where the 1 is in the i th place. So if X has initial distribution δ_i , then $X_0 = i$ with probability 1.

Theorem 9.3 (Markov property). *Let $(X_n)_{n \geq 0}$ be Markov(λ, P). Then, conditional on $X_m = i$, the process $(X_{m+n})_{n \geq 0} = (X_m, X_{m+1}, X_{m+2}, \dots)$ is Markov(δ_i, P) and is independent of X_0, \dots, X_m .*

Proof. Let

$$A = \{X_0 = i_0, \dots, X_m = i_m\}, \quad B = \{X_m = i_m, \dots, X_{m+n} = i_{m+n}\}.$$

We need to show

$$P(A \cap B | X_m = i) = P(A | X_m = i)P(B | X_m = i) \quad (9.1)$$

and

$$P(B | X_m = i) = \delta_{ii_m} p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}} \quad (9.2)$$

since then, conditional on $X_m = i$, (9.1) shows the required independence and (9.2) shows that $(X_{m+n})_{n \geq 0}$ is Markov(δ_i, P) by Theorem 9.2.

If $i \neq i_m$ then (9.1) and (9.2) hold since both sides of both equations are zero. So suppose $i = i_m$. Then

$$P(A \cap B | X_m = i) \quad (9.3)$$

$$= \frac{P(X_0 = i_0, \dots, X_{m+n} = i_{m+n})}{P(X_m = i)}$$

$$= \frac{\lambda_{i_0} p_{i_0 i_1} \cdots p_{i_{m+n-1} i_{m+n}}}{P(X_m = i)} \quad \text{by Theorem 9.2}$$

$$= P(A | X_m = i) p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}} \quad \text{using Theorem 9.2 again.} \quad (9.4)$$

Summing (9.4) over $i_0, \dots, i_{m-1} \in S$, we get

$$P(B | X_m = i) = p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}}. \quad (9.5)$$

So (9.4) and (9.5) show that (9.1) and (9.2) hold in the case $i = i_m$ also. \square

9.2 n -step transition probabilities

What is the probability that after n steps our Markov chain is in a given state?

Define the n -step transition probabilities by

$$\begin{aligned} p_{ij}(n) &= P(X_n = j \mid X_0 = i) \quad \text{for } n \geq 1 \\ p_{ij}(0) &= \delta_{ij}. \end{aligned}$$

Theorem 9.4. (i) Let P^n be the n th power of P . Then

$$p_{ij}(n) = (P^n)_{ij}.$$

(ii) $p_{ij}(m+n) = \sum_{k \in S} p_{ik}(m)p_{kj}(n)$ (Chapman–Kolmogorov equations).

Proof. (i) By induction. The result holds for $n = 0, 1$ by definition of $p_{ij}(n)$. Assuming it holds for $n - 1$, then

$$\begin{aligned} p_{ij}(n) &= P(X_n = j \mid X_0 = i) \\ &= \sum_{k \in S} P(X_n = j, X_{n-1} = k \mid X_0 = i) \\ &= \sum_{k \in S} P(X_n = j \mid X_{n-1} = k, X_0 = i)P(X_{n-1} = k \mid X_0 = i) \\ &\quad \text{using } P(A \cap B \mid C) = P(A \mid B \cap C)P(B \mid C) \\ &= \sum_{k \in S} P(X_n = j \mid X_{n-1} = k)P(X_{n-1} = k \mid X_0 = i) \quad \text{by Markov prop} \\ &= \sum_{k \in S} p_{kj}(1)p_{ik}(n-1) \\ &= \sum_{k \in S} (P^{n-1})_{ik}(P^1)_{kj} \quad \text{by inductive hypothesis} \\ &= (P^n)_{ij} \quad \text{by definition of matrix multiplication.} \end{aligned}$$

(ii)

$$(P^{m+n})_{ij} = (P^m P^n)_{ij} = \sum_{k \in S} (P^m)_{ik}(P^n)_{kj}$$

and using (i) completes the proof. \square

[Here we have proved (i) then deduced (ii). The HT2008 lectures, and a question on Paper AO2/AS2 in 2008, proved (ii) then deduced (i).]

By Theorem 9.4(i), in principle $p_{ij}(n)$ can be calculated by first calculating P^n and then taking the (i, j) component of this matrix. (Of course $p_{ij}(n) \neq (p_{ij})^n$.)

Corollary 9.5. *Let X be Markov(λ, P). Then*

$$P(X_n = i) = (\lambda P^n)_i = \sum_{j \in S} \lambda_j p_{ji}(n).$$

Remember that $\lambda = (\lambda_i : i \in S)$ is a *row* vector.

Proof. Conditioning on the value of X_0 ,

$$\begin{aligned} P(X_n = i) &= \sum_{j \in S} P(X_n = i | X_0 = j) P(X_0 = j) \\ &= \sum_{j \in S} p_{ji}(n) \lambda_j. \end{aligned} \quad \square$$

Example. The most general two-state chain has transition matrix of the form

$$P = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}.$$

[DIAGRAM.] . Using $P^{n+1} = P^n P$, and taking the $(1, 1)$ -element, we have

$$p_{11}(n+1) = p_{11}(n)(1 - \alpha) + p_{12}(n)\beta$$

and $p_{11}(n) + p_{12}(n) = P(X_n = 1 \text{ or } 2 | X_0 = 1) = 1$, so

$$p_{11}(n+1) = (1 - \alpha)p_{11}(n) + \beta(1 - p_{11}(n)).$$

Hence

$$p_{11}(n+1) = (1 - \alpha - \beta)p_{11}(n) + \beta, \quad p_{11}(0) = 1.$$

Solving this recurrence relation,

$$p_{11}(n) = \begin{cases} \frac{\beta}{\alpha + \beta} + \frac{\alpha}{\alpha + \beta}(1 - \alpha - \beta)^n & \text{if } \alpha + \beta > 0 \\ 1 & \text{if } \alpha + \beta = 0. \end{cases}$$

Example (Alternative method). First find the eigenvalues of P : these are 1 and $1 - \alpha - \beta$. (Check.)

Assume $\alpha + \beta > 0$ (as $\alpha = \beta = 0$ is trivial since $P = \text{identity matrix}$). Then the eigenvalues are distinct and P is diagonalizable

$$P = U \begin{pmatrix} 1 & 0 \\ 0 & 1 - \alpha - \beta \end{pmatrix} U^{-1}$$

for some U . Hence

$$P^n = U \begin{pmatrix} 1^n & 0 \\ 0 & (1 - \alpha - \beta)^n \end{pmatrix} U^{-1}.$$

Considering the $(1, 1)$ element,

$$p_{11}(n) = a1^n + b(1 - \alpha - \beta)^n$$

for some a and b . But $p_{11}(0) = 1$, so $a + b = 1$. Also $p_{11}(1) = p_{11} = 1 - \alpha$, so $a + b(1 - \alpha - \beta) = 1 - \alpha$, and so

$$a = \frac{\beta}{\alpha + \beta}, \quad b = \frac{\alpha}{\alpha + \beta}$$

giving $p_{11}(n)$ as before.

(In fact, $p_{ij}(n) = a1^n + b(1 - \alpha - \beta)^n$ for any i, j , but the constants a and b will depend on i and j .)

10 Class structure

Sometimes we can break a Markov chain into pieces, with each piece relatively simple to understand.

Definition. We say that i leads to j and write $i \rightarrow j$ if $p_{ij}(n) > 0$ for some $n \geq 0$. That is, $i \rightarrow j$ if the chain may ever visit state j with positive probability, starting from i .

We say that i communicates with j and write $i \leftrightarrow j$ if both $i \rightarrow j$ and $j \rightarrow i$.

Note that \leftrightarrow is an equivalence relation. (Clearly $i \leftrightarrow i$ since $p_{ii}(0) = 1$; if $i \leftrightarrow j$ then $j \leftrightarrow i$; and if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.) So we can partition S into the equivalence classes of \leftrightarrow , called *communicating classes*.

Definition. (i) A communicating class C is called *closed* if $p_{ij} = 0$ for all $i \in C, j \notin C$.

(ii) A state i is called *absorbing* if $\{i\}$ is a closed class.

(iii) A chain, or transition matrix P , is called *irreducible* if S consists of one (closed) class.

Thus a closed class is one from which we cannot escape. And i is an absorbing state iff $p_{ii} = 1$.

Example. Find the communicating classes associated to the transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Draw diagram: vertices $i \in S$, an arrow from i to j if $p_{ij} > 0$. [DIAGRAM.]

From the diagram the classes are:

$\{3\}$ absorbing

$\{4\}$ absorbing

$\{5, 6\}$ closed

$\{1, 2\}$ not closed.

Exercise. When S is finite show that there must be at least one closed class.

11 Hitting times and absorption probabilities

$\{X_n = i\}$ is the event ‘ X is in state i at time n ’.

The *hitting time* of a subset A of S is the RV

$$H_A = \begin{cases} \min\{n \geq 0 : X_n \in A\} & \text{if this set of times is non-empty} \\ +\infty & \text{otherwise.} \end{cases}$$

So H_A is the first time $n \geq 0$ that X hits the set A , and $H_A = +\infty$ if X never hits A .

For any event B we write $P_i(B) = P(B | X_0 = i)$. That is, P_i denotes probabilities conditional on starting in state i . Similarly for any RV Y write $E_i(Y) = E(Y | X_0 = i)$.

The probability starting from i that X ever hits set A is then

$$h_i^A = P_i(H_A < \infty) = P_i(X_n \in A \text{ for some } n \geq 0).$$

When A is a closed class, h_i^A is called the *absorption probability*. The mean (i.e. expected) time for X to reach A is given by

$$k_i^A = E_i(H_A) = \sum_{n=0}^{\infty} n P_i(H_A = n) + \infty P_i(H_A = \infty).$$

We sometimes write

$$h_i^A = P_i(\text{hit } A), \quad k_i^A = E_i(\text{time to hit } A).$$

These quantities can be calculated from systems of linear equations.

Example. Consider the following chain. [DIAGRAM.]

Let

$$h_i = P_i(\text{hit } 4), \quad k_i = E_i(\text{time to hit } \{1, 4\}).$$

Clearly $h_1 = 0$, $h_4 = 1$ and $k_1 = k_4 = 0$. Suppose we start at 2: then conditioning on the first step

$$h_2 = \frac{1}{3}h_1 + \frac{2}{3}h_3, \quad k_2 = \frac{1}{3}(1 + k_1) + \frac{2}{3}(1 + k_3).$$

Similarly,

$$h_3 = \frac{1}{4}h_2 + \frac{3}{4}h_4, \quad k_3 = \frac{1}{4}(1 + k_2) + \frac{3}{4}(1 + k_4).$$

Hence

$$\begin{aligned} h_2 &= \frac{2}{3}h_3 = \frac{2}{3}\left(\frac{1}{4}h_2 + \frac{3}{4}\right) \\ k_2 &= 1 + \frac{2}{3}k_3 = 1 + \frac{2}{3}\left(1 + \frac{1}{4}k_2\right) \end{aligned}$$

so $h_2 = \frac{3}{5}$, $h_3 = \frac{9}{10}$ and $k_2 = 2$, $k_3 = \frac{3}{2}$.

Theorem 11.1. *The vector of hitting probabilities $h^A = (h_i^A : i \in S)$ is the minimal non-negative solution to the equations*

$$h_i^A = \begin{cases} 1 & \text{if } i \in A \\ \sum_{j \in S} p_{ij} h_j^A & \text{if } i \notin A. \end{cases} \quad (11.1)$$

(Minimality means that if $x = (x_i : i \in S)$ is another solution with $x_i \geq 0$ for all i , then $x_i \geq h_i$ for all i .)

Proof. If $i \in A$, then $h_i^A = P_i(\text{hit } A) = 1$. If $i \notin A$, then

$$\begin{aligned} h_i^A &= P_i(\text{hit } A) \\ &= \sum_{j \in S} \underbrace{P_i(\text{hit } A \mid X_1 = j)}_{h_j^A \text{ by Markov prop}} \underbrace{P_i(X_1 = j)}_{p_{ij}} \quad \text{by conditioning on } X_1 \\ &= \sum_{j \in S} p_{ij} h_j^A \end{aligned}$$

So h^A satisfies (11.1).

To prove minimality suppose $x = (x_i : i \in S)$ is any non-negative solution to (11.1). Then for $i \in A$, $x_i = h_i^A = 1$. For $i \notin A$,

$$\begin{aligned} x_i &= \sum_{j \in S} p_{ij} x_j = \sum_{j \in A} p_{ij} + \sum_{j \notin A} p_{ij} x_j \\ &= \sum_{j \in A} p_{ij} + \sum_{j \notin A} p_{ij} \left(\sum_{k \in A} p_{jk} + \sum_{k \notin A} p_{jk} x_k \right) \\ &= P_i(X_1 \in A) + P_i(X_1 \notin A, X_2 \in A) + \sum_{j \notin A} \sum_{k \notin A} p_{ij} p_{jk} x_k. \end{aligned}$$

By repeated substitution for x in the last term we get

$$\begin{aligned} x_i &= P_i(X_1 \in A) + \cdots + P_i(X_1 \notin A, \dots, X_{n-1} \notin A, X_n \in A) \\ &\quad + \sum_{j_1 \notin A} \cdots \sum_{j_n \notin A} p_{ij_1} p_{j_1 j_2} \cdots p_{j_{n-1} j_n} x_{j_n}. \end{aligned}$$

But x is non-negative so

$$\begin{aligned} x_i &\geq P_i(X_1 \in A) + \cdots + P_i(X_1 \notin A, \dots, X_{n-1} \notin A, X_n \in A) \\ &= P_i(\text{hit } A \text{ before or at time } n). \end{aligned}$$

Hence

$$x_i \geq \lim_{n \rightarrow \infty} P_i(\text{hit } A \text{ before or at time } n) = P_i(\text{hit } A) = h_i$$

as required. □

Example (Gambler's ruin). [DIAGRAM.] Let $0 < q = 1 - p < 1$. The transition probabilities are

$$\begin{aligned} p_{00} &= 1 \\ p_{i,i+1} &= p, \quad p_{i,i-1} = q \quad \text{for } i = 1, 2, \dots \end{aligned}$$

Let $h_i = P_i(\text{hit } 0)$ be the probability of going broke starting from i (playing against an infinite bank). By Theorem 11.1, h is the minimal non-negative solution to

$$h_0 = 1 \tag{11.2}$$

$$h_i = ph_{i+1} + qh_{i-1} \quad \text{for } i = 1, 2, \dots \tag{11.3}$$

If $p \neq q$ the general solution of (11.3) is

$$h_i = A + B \left(\frac{q}{p} \right)^i.$$

If $p < q$ then the requirement $0 \leq h_i \leq 1$ forces $B = 0$, and then (11.2) gives $h_i = 1$ for all i .

If $p > q$ then (11.2) gives $A + B = 1$ so

$$h_i = \left(\frac{q}{p} \right)^i + A \left(1 - \left(\frac{q}{p} \right)^i \right)$$

and for a non-negative solution we must have $A \geq 0$, so the minimal non-negative solution is when $A = 0$ and $h_i = (q/p)^i$.

If $p = q$ the general solution of (11.3) is

$$h_i = A + Bi$$

and $0 \leq h_i \leq 1$ forces $B = 0$, so $h_i = 1$ for all i .

Example. [DIAGRAM.] Let $0 < p_i = 1 - q_i < 1$.

$$\begin{aligned} p_{00} &= 1 \\ p_{i,i+1} &= p_i, \quad p_{i,i-1} = q_i \quad \text{for } i = 1, 2, \dots \end{aligned}$$

This chain could model the size of a population, recorded each time it changes. Then $h_i = P_i(\text{hit } 0)$ is the extinction probability starting from i . As usual

$$\begin{aligned} h_0 &= 1 \\ h_i &= p_i h_{i+1} + q_i h_{i-1} \quad \text{for } i = 1, 2, \dots \end{aligned}$$

To solve note

$$(p_i + q_i)h_i = p_i h_{i+1} + q_i h_{i-1} \quad \text{since } p_i + q_i = 1$$

$$p_i u_{i+1} = q_i u_i$$

where $u_i = h_{i-1} - h_i$. So

$$u_{i+1} = \frac{q_i}{p_i} u_i = \left(\frac{q_i}{p_i} \frac{q_{i-1}}{p_{i-1}} \cdots \frac{q_1}{p_1} \right) u_1 = \gamma_i u_1.$$

Then $u_1 + \cdots + u_i = h_0 - h_i$ so

$$h_i = h_0 - (u_1 + \cdots + u_i)$$

$$= 1 - u_1(\gamma_0 + \cdots + \gamma_{i-1})$$

where u_1 remains to be determined.

If $\sum_{i=0}^{\infty} \gamma_i = \infty$, then $0 \leq h_i \leq 1$ forces $u_1 = 0$ and so $h_i = 1$ for all i .

If $\sum_{i=0}^{\infty} \gamma_i < \infty$, then we can have $u_1 > 0$ provided

$$1 - u_1(\gamma_0 + \cdots + \gamma_{i-1}) \geq 0 \quad \text{for all } i.$$

The minimal non-negative solution is $u_1 = (\sum_{i=0}^{\infty} \gamma_i)^{-1}$ and then

$$h_i = \frac{\sum_{j=i}^{\infty} \gamma_j}{\sum_{j=0}^{\infty} \gamma_j}.$$

In this case $h_i < 1$ for $i = 1, 2, \dots$, so the population survives with positive probability.

Theorem 11.2. *The vector of mean hitting times $k^A = (k_i^A : i \in S)$ is the minimal non-negative solution to the equations*

$$k_i^A = \begin{cases} 0 & \text{if } i \in A \\ 1 + \sum_{j \notin A} p_{ij} k_j^A & \text{if } i \notin A. \end{cases}$$

Proof. If $X_0 = i \in A$, then $H^A = 0$ with probability 1, so $k_i^A = 0$. If $i \notin A$, then

$$k_i^A = E_i(H^A) = \sum_{j \in S} E_i(H^A | X_1 = j) P_i(X_1 = j) \quad \text{by conditioning on } X_1$$

$$= \sum_{j \in S} p_{ij} (1 + k_j^A) \quad \text{by the Markov property}$$

$$= 1 + \sum_{j \notin A} p_{ij} k_j^A.$$

For proof of minimality see Norris p17. □

12 Recurrence and transience

Starting from state i , is it certain that X will return to i ?

Definition. State i is called *recurrent* (or *persistent*) if

$$P_i(X_n = i \text{ for some } n \geq 1) = 1$$

and *transient* if

$$P_i(X_n = i \text{ for some } n \geq 1) < 1.$$

The *first passage time* to state i is the RV

$$T_i = \begin{cases} \min\{n \geq 1 : X_n = i\} & \text{if this set of times is non-empty} \\ +\infty & \text{otherwise.} \end{cases}$$

So T_i is the first time $n \geq 1$ that X hits state i , and $T_i = +\infty$ if X never visits i over times $n \geq 1$.

Define

$$f_{ij} = P_i(T_j < \infty) = P_i(X_n = j \text{ for some } n \geq 1).$$

So i is recurrent if $f_{ii} = 1$ and transient if $f_{ii} < 1$. We would like a criterion for a state to be recurrent in terms of the n -step transition probabilities.

Let

$$f_{ij}(n) = P_i(T_j = n) = P_i(X_1 \neq j, \dots, X_{n-1} \neq j, X_n = j) \quad \text{for } n \geq 1$$

and set $f_{ij}(0) = 0$. Then

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}(n).$$

Define the generating functions

$$P_{ij}(s) = \sum_{n=0}^{\infty} p_{ij}(n)s^n, \quad F_{ij}(s) = \sum_{n=0}^{\infty} f_{ij}(n)s^n.$$

Clearly $f_{ij} = F_{ij}(1)$.

Theorem 12.1. $P_{ij}(s) = \delta_{ij} + F_{ij}(s)P_{jj}(s)$.

Proof. For $n \geq 1$,

$$\begin{aligned}
p_{ij}(n) &= P_i(X_n = j) \\
&= \sum_{r=1}^n P_i(X_n = j \mid T_j = r) P_i(T_j = r) \quad \text{by conditioning on } T_j \\
&= \sum_{r=1}^n P_i(X_n = j \mid X_r = j, X_{r-1} \neq j, \dots, X_1 \neq j) f_{ij}(r) \\
&= \sum_{r=1}^n P_i(X_n = j \mid X_r = j) f_{ij}(r) \quad \text{by the Markov property} \\
&= \sum_{r=1}^n p_{jj}(n-r) f_{ij}(r). \tag{12.1}
\end{aligned}$$

Consider $F_{ij}(s)P_{jj}(s)$: there is no s^0 term since $f_{ij}(0) = 0$, and for $n \geq 1$ the coefficient of s^n is the RHS of (12.1). So multiply (12.1) by s^n and sum over $n \geq 1$ to obtain

$$P_{ij}(s) - \delta_{ij} = F_{ij}(s)P_{jj}(s)$$

as required. □

Theorem 12.2. *State i is recurrent if and only if $\sum_{n=0}^{\infty} p_{ii}(n) = \infty$.
State i is transient if and only if $\sum_{n=0}^{\infty} p_{ii}(n) < \infty$.*

Proof. From Theorem 12.1

$$P_{ii}(s) = \frac{1}{1 - F_{ii}(s)}.$$

So

$$\begin{aligned}
i \text{ transient} &\iff f_{ii} < 1 \\
&\iff \sum_{n=0}^{\infty} f_{ii}(n) < 1 \\
&\iff \lim_{s \uparrow 1} F_{ii}(s) < 1 \\
&\iff \lim_{s \uparrow 1} P_{ii}(s) < \infty \\
&\iff \sum_{n=0}^{\infty} p_{ii}(n) < \infty. \tag{□}
\end{aligned}$$

Example (Simple symmetric random walk on \mathbb{Z}^d). [DIAGRAM.]

Then

$$p_{00}(2n+1) = 0$$

$$p_{00}(2n) \sim \frac{c_d}{n^{d/2}} \quad \text{by Stirling's formula (for some constant } c_d).$$

$d = 1, 2$: $\sum_n p_{00}(n) = \infty$ and state 0 is recurrent.

$d \geq 3$: $\sum_n p_{00}(n) < \infty$ and state 0 is transient.

Example (Simple random walk on \mathbb{Z}). [DIAGRAM.]

Then (Stirling's formula gives)

$$p_{00}(2n) \sim \frac{(4pq)^n}{\sqrt{\pi n}}.$$

If $p \neq q$, state 0 is transient. If $p = q$, state 0 is recurrent as in the previous example.

In this example it is clear that either (i) all states are recurrent, or (ii) all are transient. (Similarly in the previous example.) Case (i) occurs when $p = q$, (ii) when $p \neq q$. The following result shows, in general, that recurrence and transience are *class properties*.

Theorem 12.3. *Let C be a communicating class. Then either all states in C are transient, or all states in C are recurrent.*

Proof. We must show that C does not contain both a transient state and a recurrent state.

Take any pair of states $i, j \in C$ and suppose i recurrent. There exist $n, m \geq 0$ such that $p_{ij}(n) > 0$ and $p_{ji}(m) > 0$. Then for all $r \geq 0$

$$\begin{aligned} p_{jj}(m+r+n) &= P_j(X_{m+r+n} = j) \\ &\geq P_j(X_{m+r+n} = j, X_{m+r} = i, X_m = i) \\ &= p_{ji}(m)p_{ii}(r)p_{ij}(n) \end{aligned}$$

so

$$\sum_r p_{jj}(m+r+n) \geq p_{ji}(m)p_{ij}(n) \sum_r p_{ii}(r) = \infty$$

so j is recurrent by Theorem 12.2. □

Two facts that we won't prove are:

- every recurrent class is closed

- every finite closed class is recurrent.

(See Norris p27.) So a finite class is recurrent if and only if it is closed.

Example. The indicator function of the event $\{X_n = i\}$ is defined by

$$I(X_n = i) = \begin{cases} 1 & \text{if } X_n = i \\ 0 & \text{otherwise.} \end{cases}$$

Then the *number of visits* to state i can be written as

$$V_i = \sum_{n=0}^{\infty} I(X_n = i)$$

Let $X_0 = i$. [DIAGRAM.] Then

$$\begin{aligned} P_i(V_i \geq r) &= \text{probability of at least } r - 1 \text{ returns to } i \\ &= (f_{ii})^{r-1} \end{aligned}$$

because, probabilistically, the process ‘starts again’ when it returns to i : that is $(X_{T_i+n})_{n \geq 0}$ is $\text{Markov}(\delta_i, P)$ and independent of the past. (The Strong Markov property justifies this, not part of the course.)

$$\text{So } P_i(V_i = r) = P_i(V_i \geq r) - P_i(V_i \geq r + 1) = (f_{ii})^{r-1}(1 - f_{ii}).$$

(a) i transient: V_i is geometric, parameter $1 - f_{ii}$.

$$P_i(V_i < \infty) = \sum_{r=1}^{\infty} P_i(V_i = r) = 1.$$

$$E_i(V_i) = \frac{1}{1 - f_{ii}}.$$

(b) i recurrent, $P_i(V_i \geq r) = 1$ for all r .

$$\text{So } P_i(V_i = \infty) = \lim_{r \rightarrow \infty} P_i(V_i \geq r) = 1.$$

$$\text{We say } E_i(V_i) = +\infty.$$

Note that

$$E_i(V_i) = E_i\left[\sum_{n=0}^{\infty} I(X_n = i)\right] = \sum_{n=0}^{\infty} P_i(X_n = i) = \sum_{n=1}^{\infty} p_{ii}(n)$$

So this gives a 2nd proof of Theorem 12.2.

There are two possibilities:

- i transient, $f_{ii} < 1$, $E_i(V_i) = \frac{1}{1-f_{ii}}$, $P_i(V_i < \infty) = 1$, $\sum_n p_{ii}(n) < \infty$
- i recurrent, $f_{ii} = 1$, $E_i(V_i) = +\infty$, $P_i(V_i < \infty) = 0$, $\sum_n p_{ii}(n) = \infty$.

13 Stationary distributions and convergence to equilibrium

How does X behave after a long time has elapsed? The existence of a limiting distribution for X_n as $n \rightarrow \infty$ is closely linked to the existence of a ‘stationary distribution’.

Definition. The row vector $\pi = (\pi_i : i \in S)$ is called a *stationary distribution* if

- (i) $\pi_i \geq 0$ for all i and $\sum_{i \in S} \pi_i = 1$
- (ii) $\pi = \pi P$, that is $\pi_j = \sum_{i \in S} \pi_i p_{ij}$ for all j .

The first result explains the name ‘stationary’.

Theorem 13.1. *If π is a stationary distribution and X_0 has distribution π , then X_n has distribution π for all n .*

Proof. If π is stationary then, for all n ,

$$\pi P^n = (\pi P)P^{n-1} = \pi P^{n-1} = \dots = \pi \quad (13.1)$$

so

$$\begin{aligned} P(X_n = i) &= (\pi P^n)_i \quad \text{by Corollary 9.5} \\ &= \pi_i \quad \text{by (13.1)} \end{aligned}$$

as required. □

Example.

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

[DIAGRAM.] To find a stationary distribution: $\pi = \pi P$ gives

$$\begin{aligned} \pi_1 &= \frac{1}{2}\pi_3 \\ \pi_2 &= \pi_1 + \frac{1}{2}\pi_2 \\ \pi_3 &= \frac{1}{2}\pi_2 + \frac{1}{2}\pi_3. \end{aligned}$$

One of these equations is redundant (if π is a solution, so is $c\pi$ for any constant c). To fix π uniquely we need the equation

$$\pi_1 + \pi_2 + \pi_3 = 1$$

and we find that $\pi = (1/5, 2/5, 2/5)$.

Does a stationary distribution always exist?

13.1 Positive recurrence

Recall that a state i is recurrent if

$$P_i(T_i < \infty) = P_i(X_n = i \text{ for some } n \geq 1) = 1.$$

Define the expected return time to state i by

$$m_i = E_i(T_i).$$

Definition. Let i be recurrent. We say that

- (i) i is *positive recurrent* if $m_i < \infty$
- (ii) i is *null recurrent* if $m_i = \infty$.

Recall that P is called irreducible if S is a single communicating class.

Theorem 13.2. *Let P be irreducible. Then the following are equivalent:*

- (i) *every state is positive recurrent*
- (ii) *some state i is positive recurrent*
- (iii) *P has a stationary distribution, π say.*

When (iii) holds, π is unique and $m_i = 1/\pi_i$ for all i .

No proof (see Norris p37).

Example (Simple symmetric random walk on \mathbb{Z}). [DIAGRAM.] P is irreducible. Look for a stationary distribution: $\pi_j = \sum_i \pi_i p_{ij}$, that is

$$\pi_j = \pi_{j-1} \frac{1}{2} + \pi_{j+1} \frac{1}{2}.$$

Hence

$$\pi_{j+1} - \pi_j = \pi_j - \pi_{j-1} = \cdots = \pi_1 - \pi_0 = a$$

so $\pi_j = \pi_0 + aj$ for all $j \in \mathbb{Z}$. But $\pi_j \geq 0$ for all j , so $a = 0$. But we'd need $\sum_{j \in \mathbb{Z}} \pi_j = 1$, so no such π exists.

Hence, by Theorem 13.2, no state is positive recurrent. We know all states are recurrent (see earlier), so all are null recurrent.

13.2 Convergence to equilibrium

Example.

$$\begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}$$

[DIAGRAM.] Ignore the trivial case $\alpha = \beta = 0$.

Look for a stationary distribution: $\pi = \pi P$ gives

$$\begin{aligned} \pi_1 &= (1 - \alpha)\pi_1 + \beta\pi_2 \\ \pi_2 &= \alpha\pi_1 + (1 - \beta)\pi_2. \end{aligned}$$

So $\alpha\pi_1 = \beta\pi_2$, hence

$$\pi = (\pi_1, \pi_2) = \left(\frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta} \right).$$

Previously we calculated $p_{11}(n)$, and $p_{12}(n) = 1 - p_{11}(n)$, and similarly we can obtain $p_{21}(n)$, $p_{22}(n)$. We find that

$$P^n \rightarrow \begin{pmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{pmatrix} \quad \text{as } n \rightarrow \infty$$

that is $p_{ij}(n) \rightarrow \pi_j$, provided we omit the case $\alpha = \beta = 1$.

What happens when $\alpha = \beta = 1$? $P^{2n} = I$, $P^{2n+1} = P$, so $p_{ij}(n)$ does *not* converge as $n \rightarrow \infty$, but $\pi = (\frac{1}{2}, \frac{1}{2})$ is stationary. The problem is that we have periodic behaviour.

Definition. State i is called *aperiodic* if $p_{ii}(n) > 0$ for all sufficiently large n .

Exercise. Suppose P is irreducible and has an aperiodic state i . Show that all states are aperiodic. (Hint: adapt the proof of Theorem 12.3.)

An equivalent definition is that i is aperiodic if and only if the greatest common divisor of $\{n : p_{ii}(n) > 0\}$ is 1. The *period* $d(i)$ of state i is defined to be the greatest common divisor of $\{n : p_{ii}(n) > 0\}$.

Theorem 13.3 (Convergence to equilibrium). *Let P be irreducible and aperiodic, and suppose P has a stationary distribution π . For any initial distribution λ , if X is Markov(λ, P) then*

$$P(X_n = j) \rightarrow \pi_j \quad \text{as } n \rightarrow \infty \text{ for all } j.$$

In particular,

$$p_{ij}(n) \rightarrow \pi_j \quad \text{as } n \rightarrow \infty \text{ for all } i, j.$$

When the theorem holds

- X ‘forgets where it started’: the limit π_j does not depend on the initial distribution λ
- π is sometimes called an *equilibrium distribution*.

Sketch proof. Let X be Markov(λ, P). Let Y be Markov(π, P) and independent of X . So Y is in equilibrium already,

$$P(Y_n = j) = (\pi P^n)_j = \pi_j \quad \text{for all } n.$$

Let $W_n = (X_n, Y_n) \in S \times S$. W is a Markov chain with transition probabilities

$$\tilde{p}_{(i,k)(j,l)} = p_{ij}p_{kl}.$$

Fix a state b and let

$$T = \min\{n \geq 1 : X_n = Y_n = b\} = \min\{n \geq 1 : W_n = (b, b)\}.$$

Construct a new process Z by

$$Z_n = \begin{cases} X_n & n < T \\ Y_n & n \geq T. \end{cases}$$

[DIAGRAM.]

The proof relies on the following facts (see Norris p41)

- $P(T < \infty) = 1$
- Z is Markov(λ, P).

From (ii), $P(Z_n = j) = P(X_n = j)$ for all n, j .

Then

$$\begin{aligned} P(X_n = j) - \pi_j &= P(Z_n = j) - P(Y_n = j) \\ &= P(Z_n = j, T > n) + P(Z_n = j, T \leq n) - P(Y_n = j) \\ &= P(X_n = j, T > n) + P(Y_n = j, T \leq n) \\ &\quad - P(Y_n = j, T > n) - P(Y_n = j, T \leq n) \\ &= P(X_n = j, T > n) - P(Y_n = j, T > n). \end{aligned}$$

So

$$|P(X_n = j) - \pi_j| \leq P(T > n) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

So $P(X_n = j) \rightarrow \pi_j$ as required, and $p_{ij}(n) \rightarrow \pi_j$ follows immediately by taking $\lambda = \delta_i$. \square

14 Ergodic theorem

The *number of visits to state i before time n* can be written in terms of indicator functions as

$$V_i(n) = \sum_{k=0}^{n-1} I(X_k = i).$$

So $V_i(n)/n$ is the proportion of time before n spent in state i .

Theorem 14.1 (Ergodic theorem). *Let P be irreducible and let λ be any distribution. If X is Markov(λ, P) then, as $n \rightarrow \infty$,*

$$\frac{V_i(n)}{n} \rightarrow \frac{1}{m_i} \quad \text{with probability 1}$$

where $m_i = E_i(T_i)$ is the expected return time to state i .

In the positive recurrent case, for any bounded function $f : S \rightarrow \mathbb{R}$, as $n \rightarrow \infty$,

$$\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \rightarrow \sum_{i \in S} \pi_i f(i) \quad \text{with probability 1}$$

where π is the unique stationary distribution.

No proof.

Informally, the theorem says that the time average of $f(X_k)$ converges to the expectation of f under the stationary distribution. Note that, unlike Theorem 13.3, aperiodicity is not required: here any periodic effects average out over time.

Note that in the irreducible and positive recurrent case,

$$\frac{V_i(n)}{n} \rightarrow \pi_i \quad \text{with probability 1}$$

since $m_i = 1/\pi_i$ by Theorem 13.2.

Example. Consider a queue with a single server. In each unit of time, independently of past events:

- either the service of the person at the front of the queue is completed with probability α
- or a new customer joins the queue with probability β
- or nothing happens

where $0 < \beta < \alpha$ and $\alpha + \beta < 1$.

Let X_n be the number of people present at time n . [DIAGRAM.] The transition probabilities are

$$\begin{aligned} p_{i,i+1} &= \beta & i \geq 0 \\ p_{i,i-1} &= \alpha & i \geq 1 \\ p_{ii} &= 1 - \alpha - \beta & i \geq 1 \\ p_{00} &= 1 - \beta \\ p_{ij} &= 0 & \text{otherwise.} \end{aligned}$$

The chain is irreducible. It is also aperiodic: consider state 0 and note $p_{00} = 1 - \beta > 0$.

Look for a solution of $\pi = \pi P$: then $\pi_j = \sum_{i=0}^{\infty} \pi_i p_{ij}$ gives

$$\pi_0 = \pi_0(1 - \beta) + \pi_1\alpha \quad (14.1)$$

$$\pi_j = \pi_{j-1}\beta + \pi_j(1 - \alpha - \beta) + \pi_{j+1}\alpha \quad \text{for } j \geq 1. \quad (14.2)$$

From (14.1)

$$\pi_1 = \frac{\beta}{\alpha}\pi_0. \quad (14.3)$$

The general solution of the recurrence relation (14.2) is

$$\pi_j = A \left(\frac{\beta}{\alpha}\right)^j + B.$$

Then substituting in (14.3) gives

$$A\frac{\beta}{\alpha} + B = \frac{\beta}{\alpha}(A + B)$$

and so $B = 0$. Finally

$$1 = \sum_{j=0}^{\infty} \pi_j = A \sum_{j=0}^{\infty} \left(\frac{\beta}{\alpha}\right)^j = \frac{A}{1 - (\beta/\alpha)}.$$

So

$$\pi_j = \left(1 - \frac{\beta}{\alpha}\right) \left(\frac{\beta}{\alpha}\right)^j \quad j \geq 0.$$

Note there is no stationary distribution in the case $\beta \geq \alpha$.

By Theorem 13.3, $p_{ij}(n) \rightarrow \pi_j$ as $n \rightarrow \infty$.

By Theorem 13.2, starting from $X_0 = 0$, the mean time until the queue is next empty is $m_0 = 1/\pi_0 = \alpha/(\alpha - \beta)$.

By Theorem 14.1, the long-run (i.e. as $n \rightarrow \infty$) proportion of time for which the server is busy is $1 - \pi_0 = \beta/\alpha$.

15 The Poisson process

The Poisson process is used to model *arrivals*: e.g.

- of radioactive particles at a Geiger counter
- of telephone calls at an exchange
- of web page requests at a server.

[DIAGRAM.]

Let the RV N_t denote the number of arrivals by time t . Here time is indexed by $t \in [0, \infty)$ so $N = (N_t)_{t \geq 0}$ is a *continuous-time* process. (For Markov chains time was discrete: we had $n \in \{0, 1, 2, \dots\}$.)

Sometimes N_t is written $N(t)$.

We will call N a *counting process* if N_t takes values in $\{0, 1, 2, \dots\}$ and $N_s \leq N_t$ for $s < t$.

Let T_n be the time of the n th arrival. If $N_0 = 0$ (as it usually will be)

$$\begin{aligned} N_t &= 0 & \text{for } 0 \leq t < T_1 \\ N_t &= 1 & \text{for } T_1 \leq t < T_2 \\ N_t &= 2 & \text{for } T_2 \leq t < T_3 \\ & & \text{and so on.} \end{aligned}$$

[DIAGRAM.]

Let $N(A)$ denote the number of arrivals in a time interval A , and write $N(s, t]$ for $N(A)$ when $A = (s, t]$. Note $N(s, t] = N_t - N_s$.

We give two definitions, which we will show are equivalent.

Definition (Description D1). The counting process N is said to be a *Poisson process of rate (or intensity) λ* (> 0) if $N_0 = 0$ and

- (i) the numbers of arrivals $N(A_1), \dots, N(A_n)$ in any finite collection A_1, \dots, A_n of *disjoint* time intervals are *independent* RVs
- (ii) the number of arrivals in any interval of length t has a Poisson distribution with mean λt .

We write $PP(\lambda)$ for a Poisson process of rate λ . A $PP(\lambda)$ ‘on $[0, \infty)$ ’ simply means that $t \in [0, \infty)$.

Definition (Description D2 – infinitesimal definition). The counting process N is said to be a $PP(\lambda)$ if $N_0 = 0$ and

- (i) the numbers of arrivals $N(A_1), \dots, N(A_n)$ in any finite collection A_1, \dots, A_n of *disjoint* time intervals are *independent* RVs [i.e. property D1(i)]
- (ii) the distribution of the number of arrivals in an interval depends only on the length of the interval
- (iii) for small positive h ,

$$\begin{aligned} P(N(t, t+h] = 0) &= 1 - \lambda h + o(h) \\ P(N(t, t+h] = 1) &= \lambda h + o(h). \end{aligned}$$

[$f(h) = o(h)$ means $f(h)/h \rightarrow 0$ as $h \rightarrow 0$.]

From D2(iii) we clearly have $P(N(t, t+h] \geq 2) = o(h)$.

The change, or *increment*, of N over interval $(s, t]$ is $N_t - N_s$. So D1(i) says: N has *independent* increments (over any finite collection of disjoint time intervals). N is said to have *stationary* increments if the distribution of $N_{s+t} - N_s$ depends only on t , which is D2(ii).

So, for a counting process N with $N_0 = 0$,

- D1 is: N has independent increments, plus D1(ii)
- D2 is: N has stationary independent increments, plus D2(iii).

Theorem 15.1. *Descriptions D1 and D2 are equivalent.*

Proof. D2 \implies D1. It is enough to show that $N(0, t] \sim \text{Poisson}(\lambda t)$.

Let $p_n(t) = P(N(0, t] = n)$ for $n \geq 0$. Then

$$\begin{aligned} p_0(t+h) &= P(N(0, t+h] = 0) \\ &= P(N(0, t] = 0, N(t, t+h] = 0) \\ &= P(N(0, t] = 0)P(N(t, t+h] = 0) \quad \text{by D2(i)} \\ &= p_0(t)(1 - \lambda h + o(h)) \quad \text{by D2(iii)}. \end{aligned}$$

So

$$\frac{p_0(t+h) - p_0(t)}{h} = -\lambda p_0(t) + \frac{o(h)}{h}$$

and letting $h \rightarrow 0$ gives

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t).$$

Since $N_0 = 0$ the initial conditions are

$$p_0(0) = 1, \quad p_n(0) = 0 \quad \text{for } n \geq 1.$$

Solving the differential equation for $p_0(t)$, and using $p_0(0) = 1$, we get $p_0(t) = e^{-\lambda t}$.

For $n \geq 1$,

$$\begin{aligned} p_n(t+h) &= \sum_{k=0}^n P(N(0,t] = k, N(t,t+h] = n-k) \\ &= \sum_{k=0}^n p_k(t) P(N(t,t+h] = n-k) \quad \text{by D2(i)} \\ &= p_n(t)(1 - \lambda h + o(h)) + p_{n-1}(t)(\lambda h + o(h)) + o(h) \quad \text{by D2(iii)}. \end{aligned}$$

So

$$\frac{p_n(t+h) - p_n(t)}{h} = -\lambda p_n(t) + \lambda p_{n-1}(t) + \frac{o(h)}{h}$$

and letting $h \rightarrow 0$

$$\frac{dp_n(t)}{dt} = -\lambda p_n(t) + \lambda p_{n-1}(t). \quad (15.1)$$

We already know $p_0(t) = e^{-\lambda t}$, so we can solve (15.1) inductively. First let $n = 1$ to get

$$\frac{dp_1(t)}{dt} = -\lambda p_1(t) + \lambda e^{-\lambda t}$$

and remember $p_1(0) = 0$, so we can solve for $p_1(t)$. Then let $n = 2, \dots$, etc. Induction shows

$$p_n(t) = \frac{e^{-\lambda t} (\lambda t)^n}{n!} \quad \text{for } n \geq 0$$

as required.

D1 \implies D2 is almost immediate – note that D2(ii) follows immediately from D1(ii). For D2(iii) note

$$\begin{aligned} P(N(t,t+h] = 0) &= e^{-\lambda h} \quad \text{by D1(ii)} \\ &= 1 - \lambda h + o(h) \end{aligned}$$

and

$$\begin{aligned} P(N(t,t+h] = 1) &= e^{-\lambda h} \lambda h \quad \text{using D1(ii)} \\ &= \lambda h(1 - \lambda h + o(h)) \\ &= \lambda h + o(h). \end{aligned} \quad \square$$

15.1 Interarrival times

Let N be a $PP(\lambda)$. The *arrival times* T_0, T_1, T_2, \dots of N are defined as before by $T_0 = 0$ and

$$T_n = \inf\{t \geq 0 : N_t = n\} \quad \text{for } n \geq 1. \quad (15.2)$$

[DIAGRAM.]

The *interarrival times* are the RVs X_1, X_2, \dots given by

$$X_n = T_n - T_{n-1}. \quad (15.3)$$

From knowledge of N we can find the values of X_1, X_2, \dots by (15.2) and (15.3). Conversely, we can reconstruct N from knowledge of the X_i by

$$T_n = \sum_{i=1}^n X_i, \quad N_t = \max\{n \geq 0 : T_n \leq t\}.$$

Theorem 15.2. X_1, X_2, \dots are independent $\text{Exp}(\lambda)$ RVs.

Proof. First note

$$P(X_1 > t) = P(T_1 > t) = P(N(0, t] = 0) = e^{-\lambda t} \quad (15.4)$$

and so $X_1 \sim \text{exponential}(\lambda)$ ((15.4) implies $f_{X_1}(t) = \lambda e^{-\lambda t}$).

Next

$$\begin{aligned} P(X_{k+1} > t \mid X_1 = s_1, \dots, X_k = s_k) \\ &= P(N(s_1 + \dots + s_k, s_1 + \dots + s_k + t] = 0 \mid X_1 = s_1, \dots, X_k = s_k) \\ &= P(N(s_1 + \dots + s_k, s_1 + \dots + s_k + t] = 0) \quad \text{from D1(i)} \\ &= e^{-\lambda t} \quad \text{from D1(ii)}. \end{aligned}$$

Hence $X_{k+1} \sim \text{Exp}(\lambda)$, is independent of X_1, \dots, X_k , and the result follows. \square

Since $T_n = \sum_{i=1}^n X_i$ and the X_i are i.i.d. $\text{Exp}(\lambda)$, we have $T_n \sim \text{Gamma}(n, \lambda)$.

15.2 Examples

Example. Suppose $N(0, t] = 1$, i.e. exactly one arrival in $(0, t]$. What can we say about the time T_1 at which this arrival occurred?

$$\begin{aligned} P(T_1 < s \mid N(0, t] = 1) &= \frac{P(T_1 < s, N(0, t] = 1)}{P(N(0, t] = 1)} \\ &= \frac{P(N(0, s] = 1, N(s, t] = 0)}{P(N(0, t] = 1)} \\ &= \frac{P(N(0, s] = 1)P(N(s, t] = 0)}{P(N(0, t] = 1)} \end{aligned}$$

since $(0, s]$ and $(s, t]$ are disjoint time intervals (use D1(i))

$$\begin{aligned} &= \frac{\lambda s e^{-\lambda s} e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \\ &= \frac{s}{t}. \end{aligned}$$

So the conditional density of T_1 is

$$f(s) = \frac{d}{ds} \left(\frac{s}{t} \right) = \frac{1}{t} \quad \text{for } 0 < s \leq t$$

which is a uniform density. So, conditional on one arrival in $(0, t]$, the time at which the arrival occurred is $\text{Uniform}(0, t]$.

Example (Superposition). Suppose N and M are independent Poisson processes of rates λ and μ . Then $(Z_t)_{t \geq 0}$ defined by $Z_t = N_t + M_t$ is a $PP(\lambda + \mu)$.

Why?

D1(i) follows for Z since it is true for N and M separately.

D1(ii) requires: ‘if U and V are independent Poisson RVs with parameters λt and μt , then $U + V \sim \text{Poisson}((\lambda + \mu)t)$ ’ – this bit is Mods.

Example. Above, what is the probability that the first arrival of process Z comes from process N ?

[DIAGRAM.] If the interarrival times are X_1, \dots for N , and Y_1, \dots for M , then

$$P(\text{first is from } N) = P(X_1 < Y_1)$$

where $X_1 \sim \text{Exp}(\lambda)$ and $Y_1 \sim \text{Exp}(\mu)$ are independent. We calculated this earlier in the course.

Example (Thinning). Emails arrive as a Poisson process $(N_t)_{t \geq 0}$ of rate λ . Messages are independently useful, with probability p , or junk, with probability $q = 1 - p$.

Then the arrival processes $(U_t)_{t \geq 0}$ of useful emails and $(J_t)_{t \geq 0}$ of junk emails are independent Poisson processes of rates $p\lambda$ and $q\lambda$. (Problem Sheet.)

Let Y be the number of emails up to and including the first useful one. [DIAGRAM.] Then

$$P(Y = k) = q^{k-1} p \quad \text{for } k \geq 1.$$

The first useful email arrives at time

$$T = \sum_{k=1}^Y X_k \tag{15.5}$$

where the X_i are i.i.d. $\text{Exp}(\lambda)$.

- (i) Either: T is also the first interarrival time for $N \sim PP(\lambda p)$, so $T \sim \text{Exp}(\lambda p)$ by Theorem 15.2.
- (ii) or: T as in (15.5) has an $\text{Exp}(\lambda p)$ by finding the mgf of T (Problem Sheet 2).

Either way: ‘the sum of a geometric number of independent exponential RVs is itself exponential’.