Mathematical Methods for Computer Science

Probability methods

Dr R.J. Gibbens

Computer Laboratory
University of Cambridge

Computer Science Tripos, Part IB
Michaelmas Term 2016/17

Last revised: 2016-09-28 (09025b0)

Outline

- Probability methods (10 lectures, Dr R.J. Gibbens)
  - Probability generating functions (2 lectures)
  - Inequalities and limit theorems (3 lectures)
  - Stochastic processes (5 lectures)
- Fourier and related methods (6 lectures, Professor J. Daugman)

Reference books (Probability methods)

- (*) Ross, Sheldon M.
  Probability Models for Computer Science.
  Harcourt/Academic Press, 2002
- Mitzenmacher, Michael & Upfal, Eli.
  Probability and Computing: Randomized Algorithms and Probabilistic Analysis.
  Cambridge University Press, 2005

Some notation

RV	random variable
IID	independent, identically distributed
PGF	probability generating function $G_X(z)$
MGF	moment generating function $M_X(t)$
$X \sim U(0,1)$	RV X has the distribution $U(0,1)$, etc
$\mathbb{I}(A)$	indicator function of the event A
$\mathbb{P}(A)$	probability that event A occurs
$\mathbb{E}(X)$	expected value of RV X
$\mathbb{E}(X^n)$	n^{th} moment of RV X, for $n = 1, 2, \ldots$
$F_X(x)$	distribution function, $F_X(x) = \mathbb{P}(X \leq x)$
$f_X(x)$	density of RV X given, when it exists, by $F'_X(x)$

Case studies

Case studies

We will consider three short cases studies where probability plays a pivitol role:

1. Birthday problem (birthday attack)
   - cryptographic attacks
2. Probabilistic classification (naive Bayes classifier)
   - email spam filtering
3. Gambler's ruin problem (cryptocurrency)
   - Bitcoin

We will use the first two to help us recall elementary probability theory (and establish our notation) and defer the third until we have studied the theory of random walks.

The birthday problem

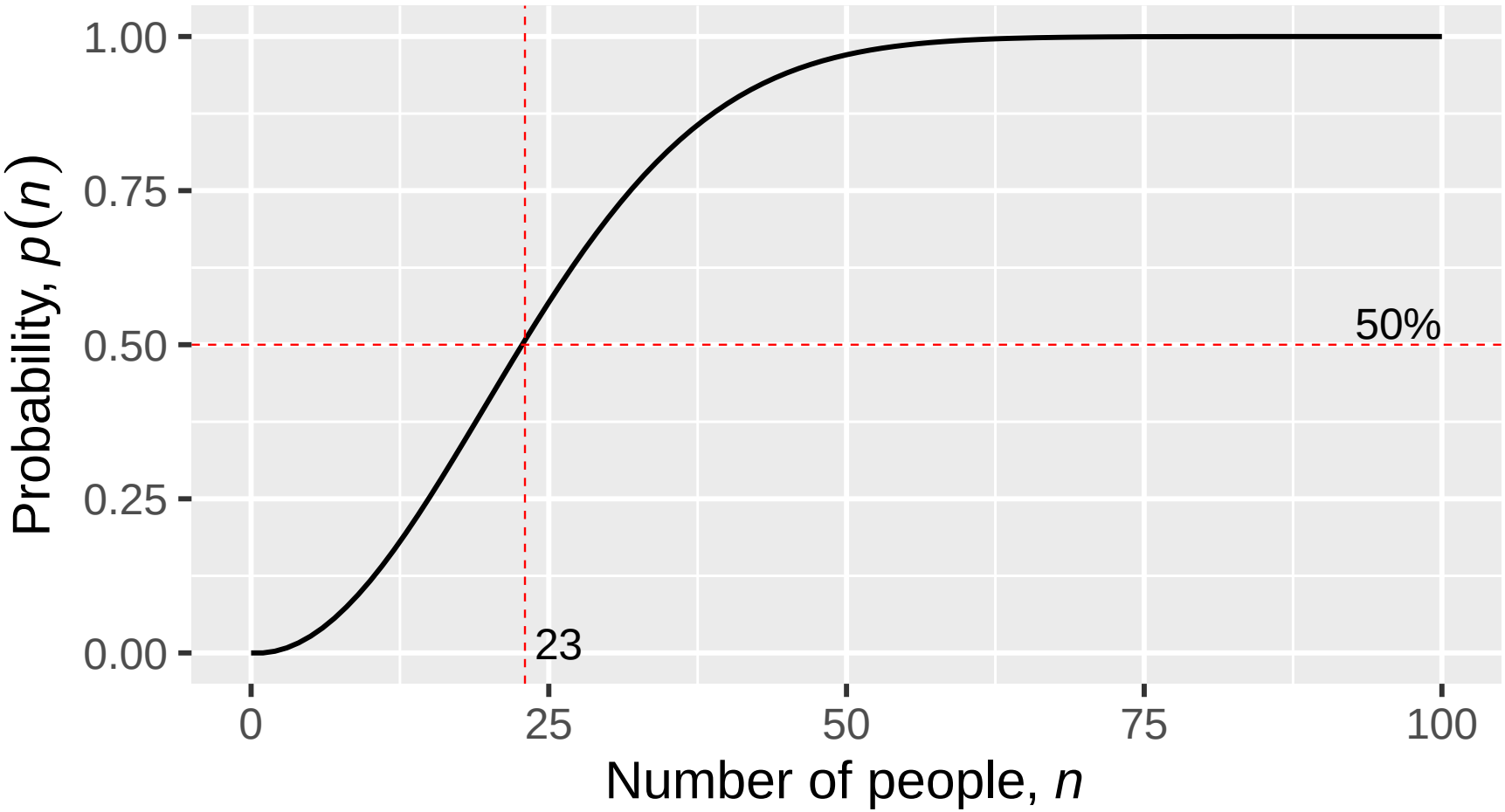
Consider the problem of computing the probability, $p(n)$, that in a party of n people at least two people share a birthday (that is, the same day and month but not necessarily same year).

It is easiest to first work out $1 - p(n) = q(n)$, say,
where $q(n) = \mathbb{P}(\text{none of the } n \text{ people share a birthday})$ then

$$\begin{aligned} q(n) &= \left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\cdots\left(\frac{365-n+1}{365}\right) \\ &= \left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\cdots\left(1-\frac{n-1}{365}\right) \\ &= \prod_{k=1}^{n-1}\left(1-\frac{k}{365}\right). \end{aligned}$$

Surprisingly, $n = 23$ people suffice to make $p(n)$ greater than 50%.

Graph of $p(n) = 1 - q(n)$ vs n



Assumptions

We should record some of our assumptions behind the calculation of $p(n)$.

1. Ignore leap days (29 Feb)
2. Each birthday is equally likely
3. People are selected independently and without regard to their birthday to attend the party (ignore twins, etc)

Examples: coincidences on the football field

Ian Stewart writing in Scientific American illustrates the birthday problem with an interesting example.

In a football match there are 23 people (two teams of 11 plus the referee) and on 19 April 1997 out of 10 UK Premier Division games there were 6 games with birthday coincidences and 4 games without.

Ian Stewart
What a coincidence!
Mathematical Recreations, Scientific American, Jun 1998, 95–96.

Examples: cryptographic hash functions

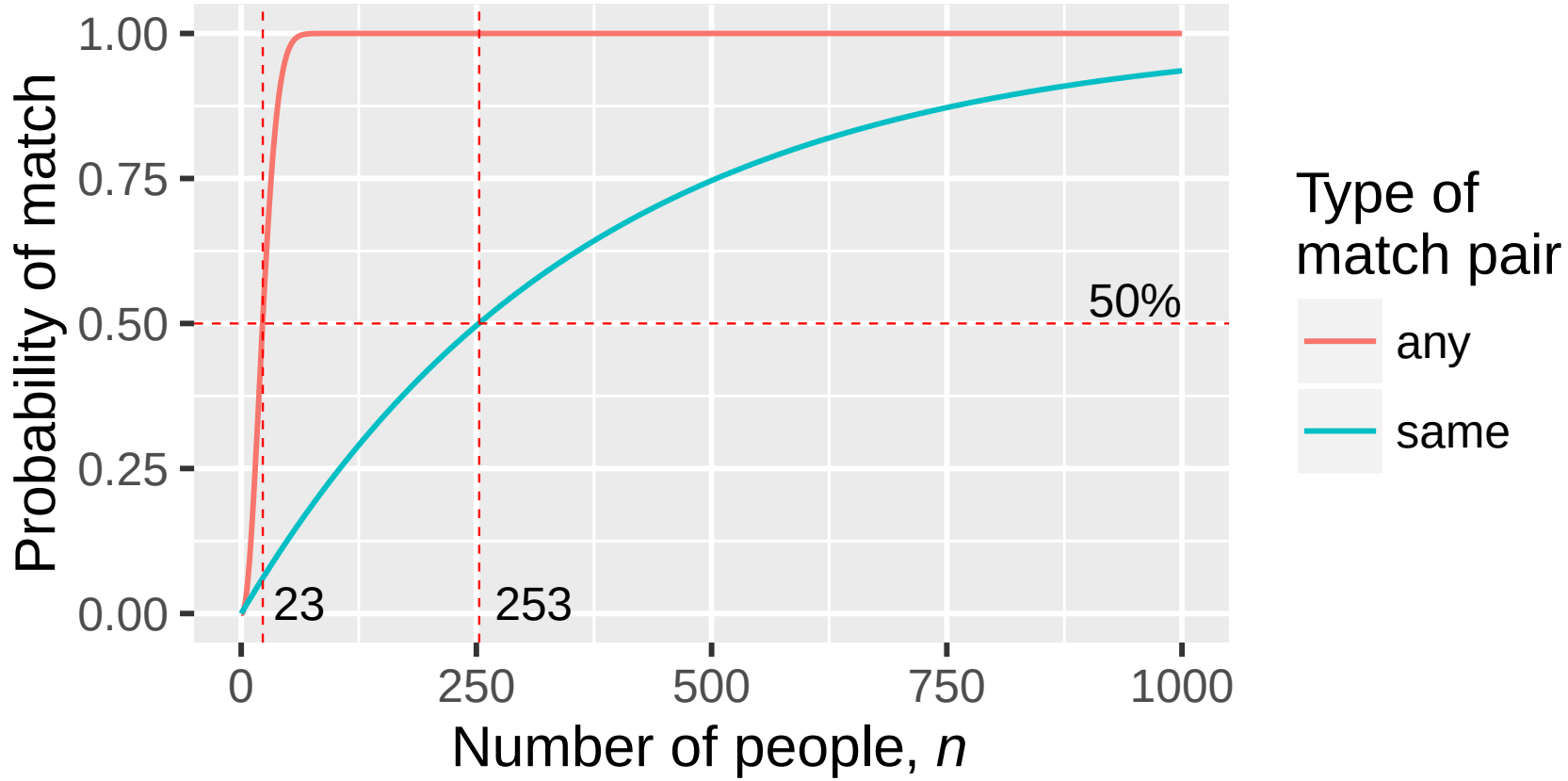
A hash function $y = f(x)$ used in cryptographic applications is usually required to have the following two properties (amongst others):

1. one-way function: computationally intractable to find an x given y.
2. collision-resistant: computationally intractable to find distinct x_1 and x_2 such that $f(x_1) = f(x_2)$.

Probability of same birthday as you

Note that in calculating $p(n)$ we are not specifying which birthday (for example, your own) matches. For the case of finding a match to your own birthday amongst a party of n other people we would calculate

$$1 - \left(\frac{364}{365}\right)^n .$$



General birthday problem

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ of size n where X_i are IID with $X_i \sim U(1, d)$ and let $p(n, d)$ be the probability that there are at least two outcomes that coincide.

Then

$$p(n,d) = \begin{cases} 1 - \prod_{k=1}^{n-1} \left(1 - \frac{k}{d}\right) & n \leq d \\ 1 & n > d\,. \end{cases}$$

The usual birthday problem is the special case when $d = 365$.

Approximations

One useful approximation is to note that for small $x > 0$ then $1 - x \approx e^{-x}$. Hence for $n \leq d$

$$\begin{aligned} p(n,d) &= 1 - \prod_{k=1}^{n-1} \left(1 - \frac{k}{d}\right) \\ &\approx 1 - \prod_{k=1}^{n-1} e^{-\frac{k}{d}} \\ &= 1 - e^{-(\Sigma_{k=1}^{n-1} k)/d} \\ &= 1 - e^{-n(n-1)/(2d)} . \end{aligned}$$

We can further approximate the last expression as

$$p(n,d) \approx 1 - e^{-n^2/(2d)} .$$

Inverse birthday problem

Using the last approximation

$$p(n,d) \approx 1 - e^{-n^2/(2d)}$$

we can invert the birthday problem to find $n = n(p,d)$, say, such that $p(n,d) \approx p$ so then

$$\begin{aligned} e^{-n(p,d)^2/(2d)} &\approx 1-p \\ -\frac{n(p,d)^2}{2d} &\approx \log(1-p) \\ n(p,d)^2 &\approx 2d \log\left(\frac{1}{1-p}\right) \\ n(p,d) &\approx \sqrt{2d \log\left(\frac{1}{1-p}\right)}. \end{aligned}$$

In the special case of $d = 365$ and $p = 1/2$ this gives the approximation $n(0.5, 365) \approx \sqrt{2 \times 365 \times \log(2)} \approx 22.49$.

Expected waiting times for a collision/match

Let W_d be the random variable specifiying the number of iterations when you choose one of d values independently and uniformly at random (with replacement) and stop when any value is selected a second time (that is, a "collision" or "match" occurs).

It is possible to show that

$$\mathbb{E}(W_d) \approx \sqrt{\frac{\pi d}{2}}.$$

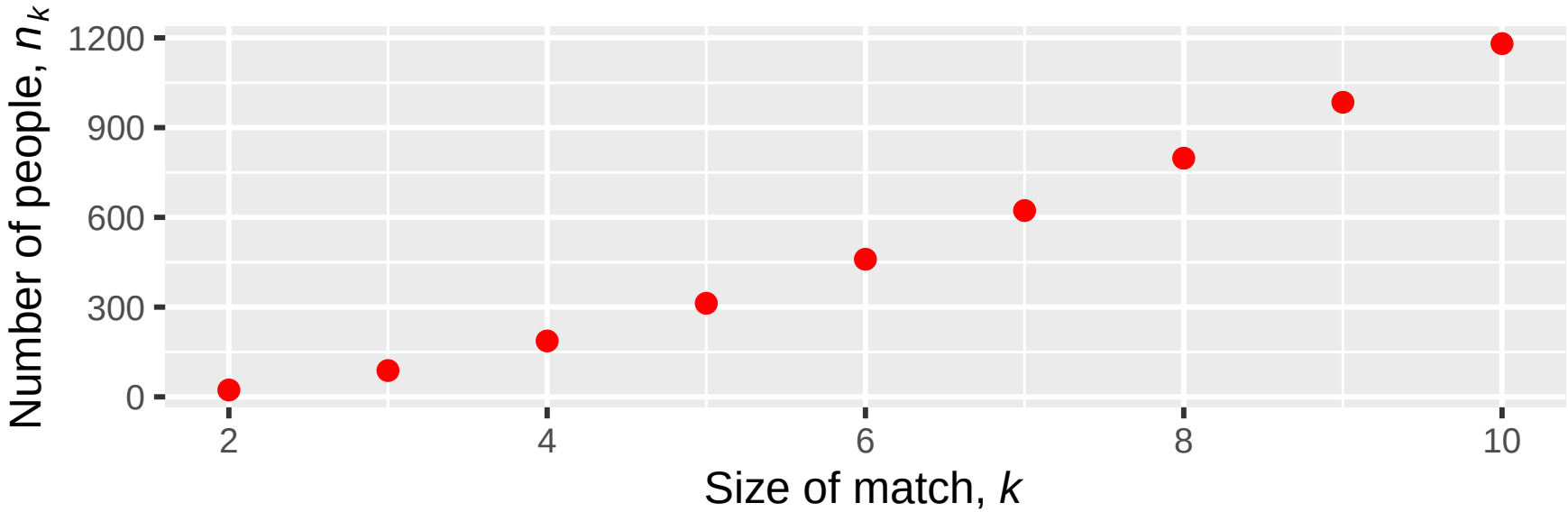
Thus in the special case of the birthday problem where $d = 365$ we have that $\mathbb{E}(W_{365}) \approx \sqrt{\frac{\pi \times 365}{2}} \approx 23.94$.

In the case that we have a cryptographic hash function with 160-bit outputs ($d = 2^{160}$) then $\mathbb{E}(W_{2^{160}}) \approx 1.25 \times 2^{80}$. This level of reduction leads to so-called "birthday attacks". (See the IB course Security I for further background discussion.)

Further results

This table (see ref below) gives the minimum number of people, n_k, such that the probability is $> 1/2$ of k or more matches with $d = 365$.

k	2	3	4	5	6	7	8	9	10
n_k	23	88	187	313	460	623	798	985	1181



Persi Diaconis and Frederick Mosteller
Methods for studying coincidences.
Journal of American Statistical Association, Vol 84, No 408, Dec 1989, 853–861.

Email spam filtering

Suppose that an email falls into exactly one of two classes (spam or ham) and that various features $F_1, F_2, \ldots, F_n$ of an email message can be measured. Such features could be the presence or absence of particular words or groups of words, etc, etc.

We would like to determine $\mathbb{P}(C \mid F_1, F_2, \ldots, F_n)$ the probability that an email message falls into a class C given the measured features $F_1, F_2, \ldots, F_n$.

We can use Bayes' theorem to help us.

Bayes' theorem for emails

We have that

$$\mathbb{P}(C \,|\, F_1, F_2, \ldots, F_n) = \frac{\mathbb{P}(C)\mathbb{P}(F_1, F_2, \ldots, F_n \,|\, C)}{\mathbb{P}(F_1, F_2, \ldots, F_n)}$$

which can be expressed in words as

$$\text{posterior probability} = \frac{\text{prior probability} \times \text{likelihood}}{\text{evidence}}\,.$$

Naive Bayes classifier

In the naive Bayes classifier we make the assumption of conditional independence across features. So that

$$\mathbb{P}(F_1, F_2, \ldots, F_n \,|\, C) = \prod_{i=1}^{n} \mathbb{P}(F_i \,|\, C)$$

and then

$$\begin{aligned}\mathbb{P}(C \,|\, F_1, F_2, \ldots, F_n) &= \frac{\mathbb{P}(C)\prod_{i=1}^{n}\mathbb{P}(F_i \,|\, C)}{\mathbb{P}(F_1, F_2, \ldots, F_n)} \\ &\propto \mathbb{P}(C)\prod_{i=1}^{n}\mathbb{P}(F_i \,|\, C)\,.\end{aligned}$$

Note that here we are referring to conditional independence in the sense of independence relative to the conditional joint probability distribution.

Decision rule for naive Bayes classifier

We then use the following decision rule to classify an email with observed features $F_1, F_2, \ldots, F_n$ as spam if

$$\mathbb{P}(C = \text{spam}) \prod_{i=1}^{n} \mathbb{P}(F_i \,|\, C = \text{spam}) > \mathbb{P}(C = \text{ham}) \prod_{i=1}^{n} \mathbb{P}(F_i \,|\, C = \text{ham}).$$

This decision rule is known as the maximum a posteriori (MAP) rule.

A training set of manually classified emails is needed to estimate the values of $\mathbb{P}(C)$ and $\mathbb{P}(F_i \,|\, C)$.

Probability generating functions

Probability generating functions (PGF)

A very common situation is when a RV, X, can take only non-negative integer values. For example, X may count the number of random events to occur in a fixed period of time. The probability mass function, $\mathbb{P}(X = k)$, is given by a sequence of values $p_0, p_1, p_2, \ldots$ where

$$p_k = \mathbb{P}(X = k) \geq 0 \quad \forall k \in \{0, 1, 2, \ldots\} \quad \text{and} \quad \sum_{k=0}^{\infty} p_k = 1 .$$

This sequence of terms can be "wrapped together" to define a function called the probability generating function (PGF) as follows.

Definition (Probability generating function)

The probability generating function, $G_X(z)$, of a (non-negative integer-valued) RV X is defined as

$$G_X(z) = \sum_{k=0}^{\infty} p_k z^k$$

for all real values of z for which the sum converges.

Elementary properties of the PGF

1. $G_X(z) = \sum_{k=0}^{\infty} p_k z^k$ so

$$G_X(0) = p_0 \qquad \text{and} \qquad G_X(1) = \sum_{k=0}^{\infty} p_k 1^k = \sum_{k=0}^{\infty} p_k = 1\,.$$

2. If $g(t) = z^t$ then

$$G_X(z) = \sum_{k=0}^{\infty} p_k z^k = \sum_{k=0}^{\infty} g(k)\mathbb{P}(X=k) = \mathbb{E}(g(X)) = \mathbb{E}(z^X)\,.$$

3. The PGF is defined for all $|z| \leq 1$ since

$$\sum_{k=0}^{\infty} |p_k z^k| \leq \sum_{k=0}^{\infty} p_k = 1\,.$$

4. Importantly, the PGF characterizes the distribution of a RV in the sense that

$$G_X(z) = G_Y(z) \qquad \forall z$$

if and only if

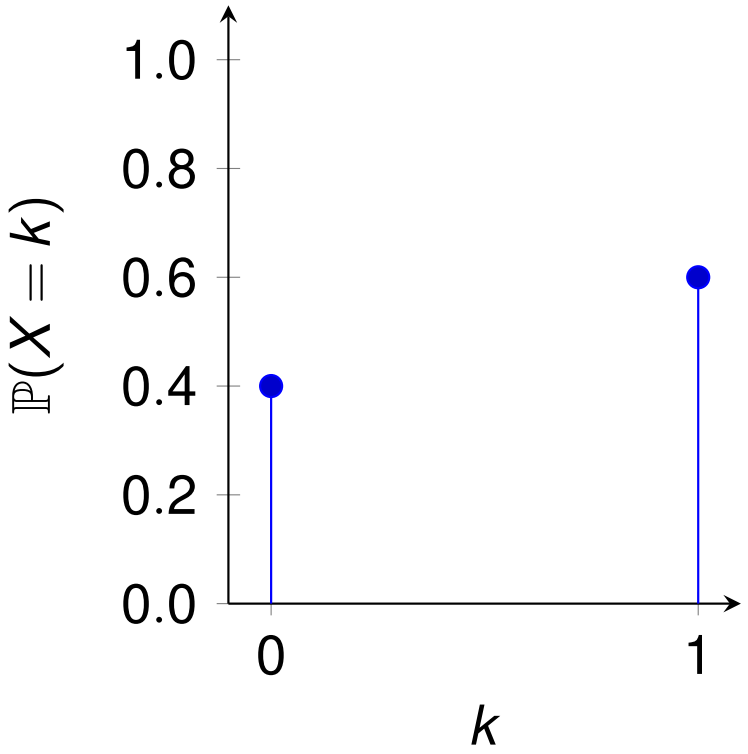
$$\mathbb{P}(X=k) = \mathbb{P}(Y=k) \qquad \forall k \in \{0,1,2,\ldots\}\,.$$

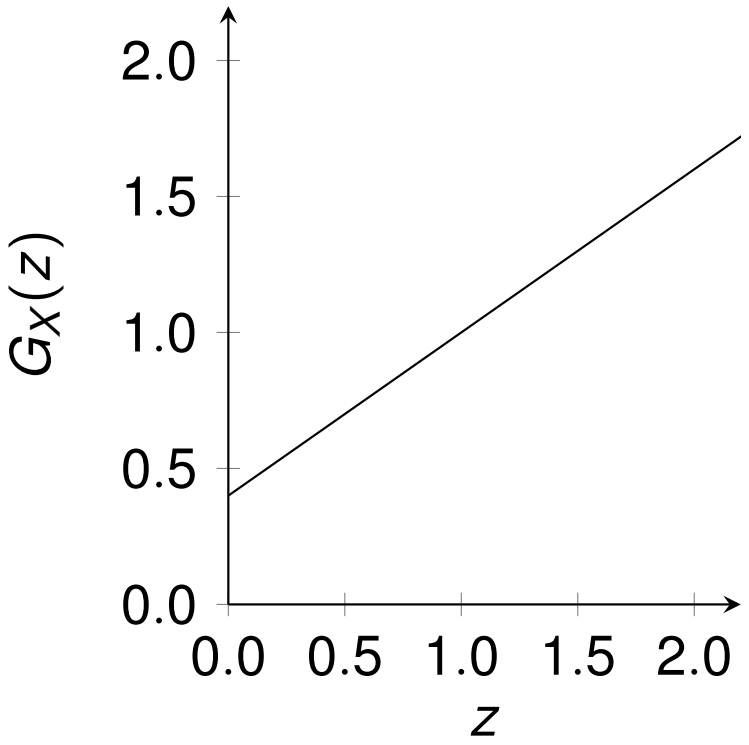
Examples of PGFs

Example (Bernoulli distribution)

$$G_X(z) = q + pz \qquad \text{where } q = 1 - p.$$

E.g. $p = 0.6, q = 1 - p = 0.4$

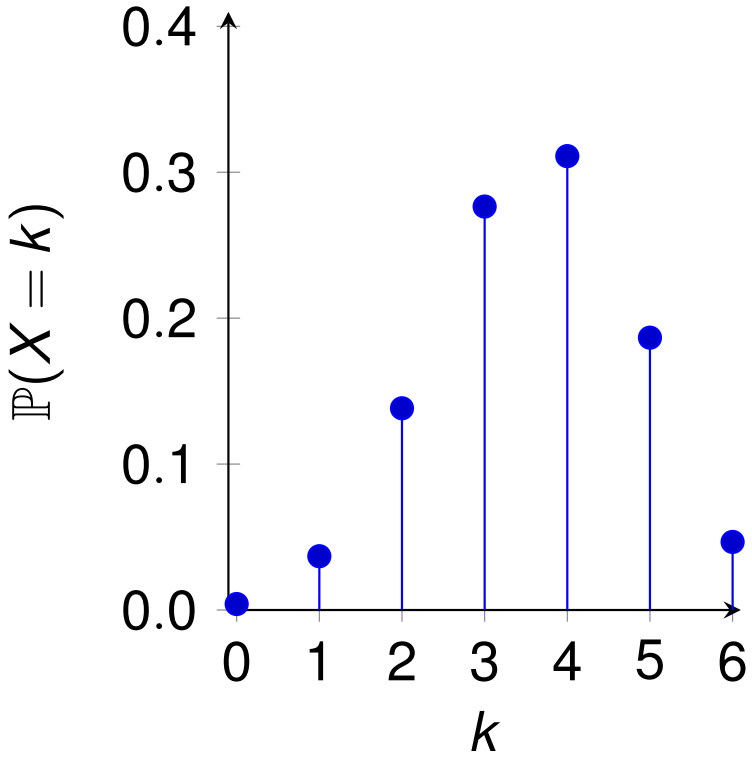


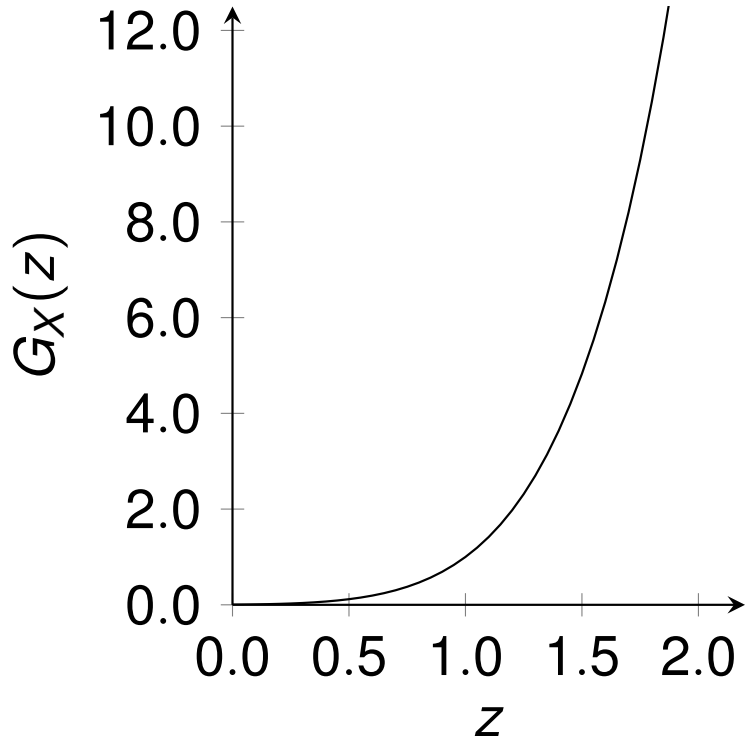


Example (Binomial distribution, Bin(n,p))

$$G_X(z) = \sum_{k=0}^{n} \binom{n}{k} p^k (q)^{n-k} z^k = (q+pz)^n \qquad \text{where } q = 1-p.$$

E.g. $p = 0.6, q = 1-p = 0.4$ and $n = 6$

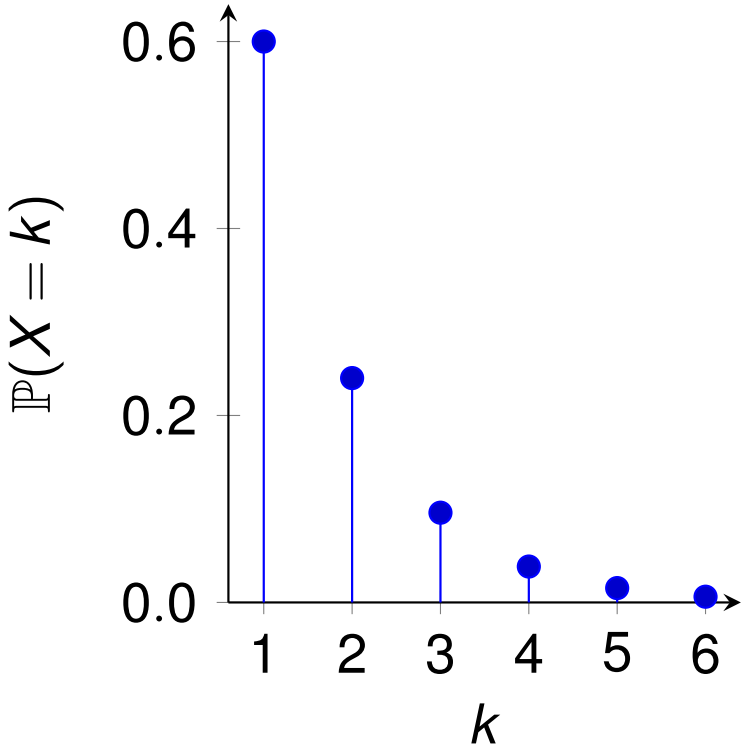


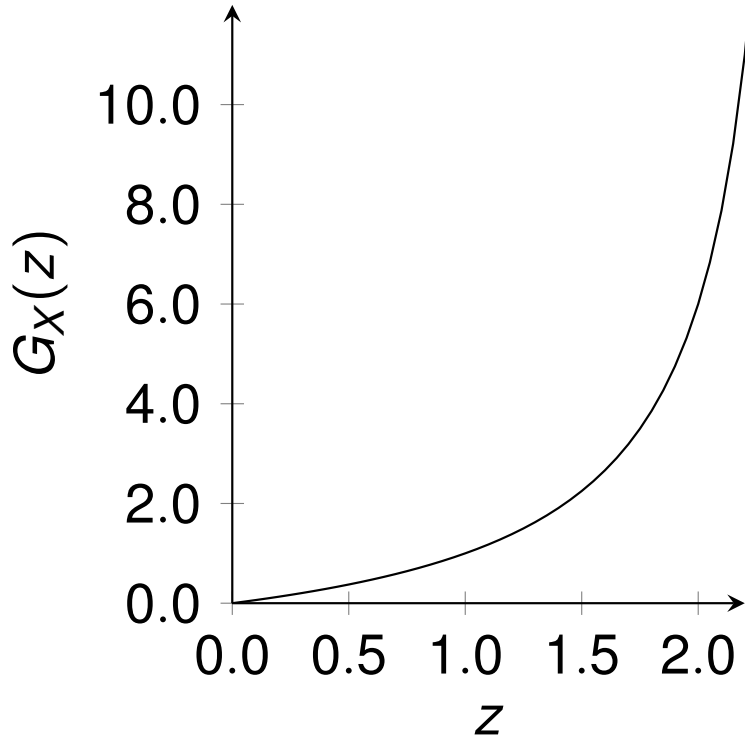


Example (Geometric distribution, Geo(p))

$$G_X(z) = \sum_{k=1}^{\infty} pq^{k-1}z^k = pz\sum_{k=0}^{\infty}(qz)^k = \frac{pz}{1-qz} \text{ if } |z| < q^{-1} \text{ and } q = 1-p.$$

E.g. $p = 0.6, q = 1-p = 0.4$

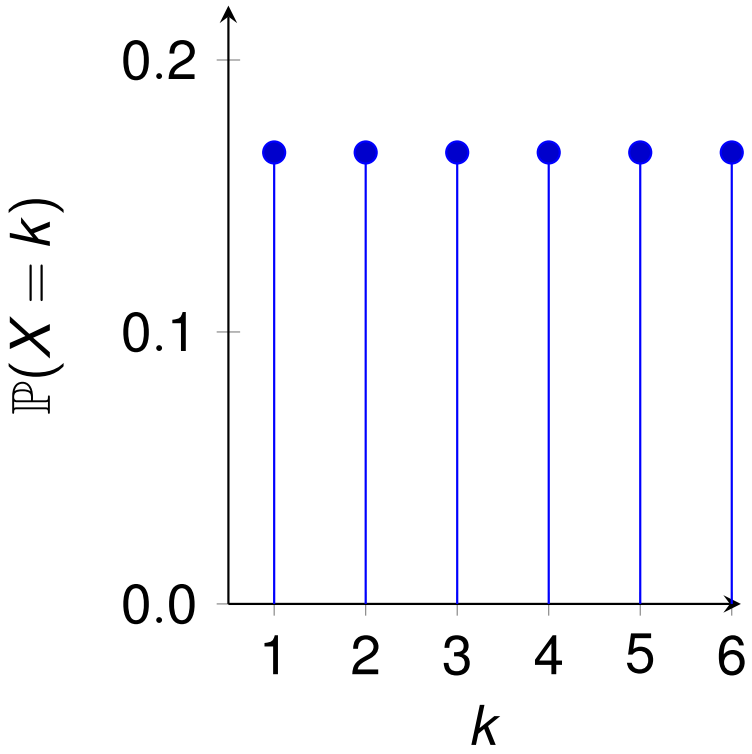


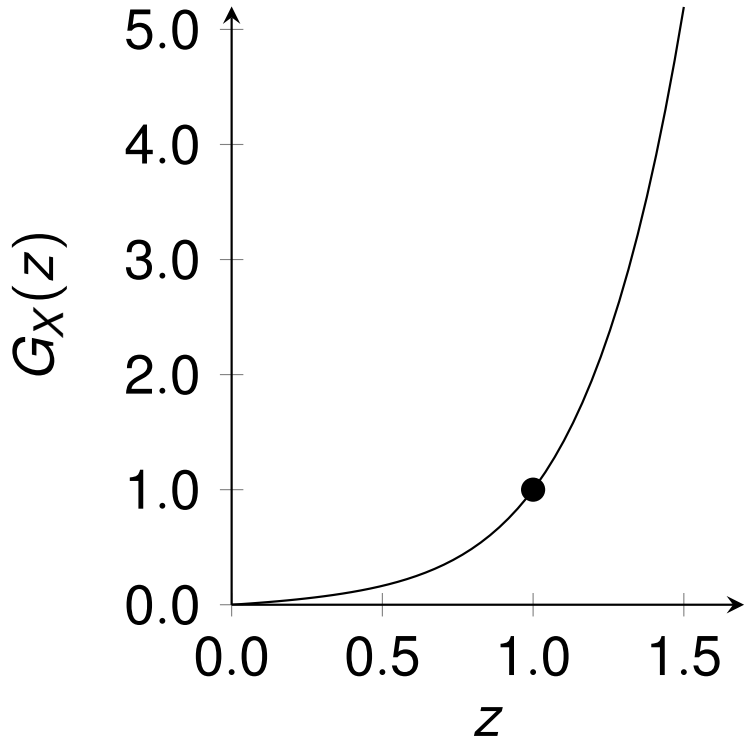


Example (Discrete uniform distribution, $U(1,n)$)

$$G_X(z) = \sum_{k=1}^{n} z^k \frac{1}{n} = \frac{z}{n} \sum_{k=0}^{n-1} z^k = \begin{cases} \frac{z}{n} \frac{(1-z^n)}{(1-z)} & z \neq 1 \\ 1 & z = 1 \end{cases}.$$

E.g. $n = 6$ and so $\mathbb{P}(X = k) = 1/6 \approx 0.167$ for $k = 1, \ldots, 6$

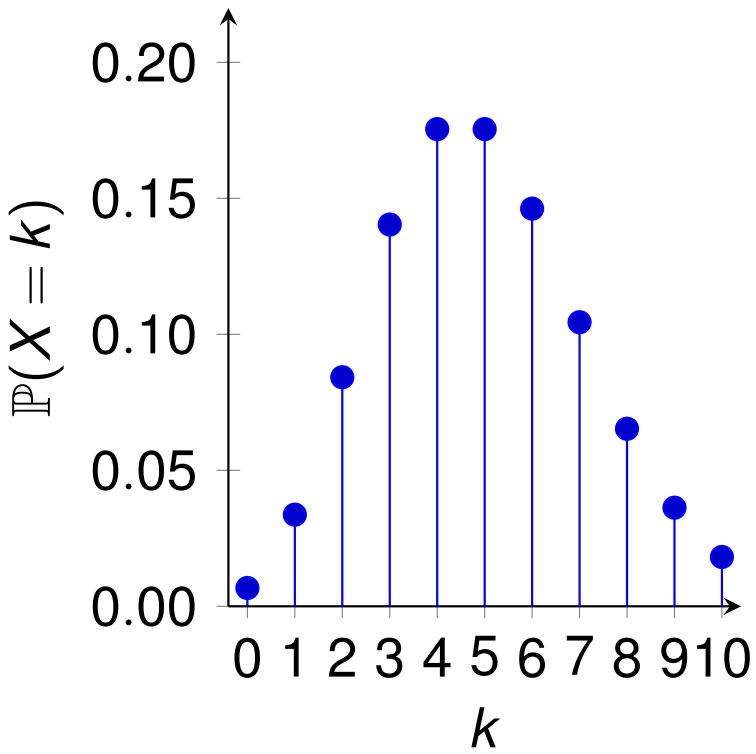


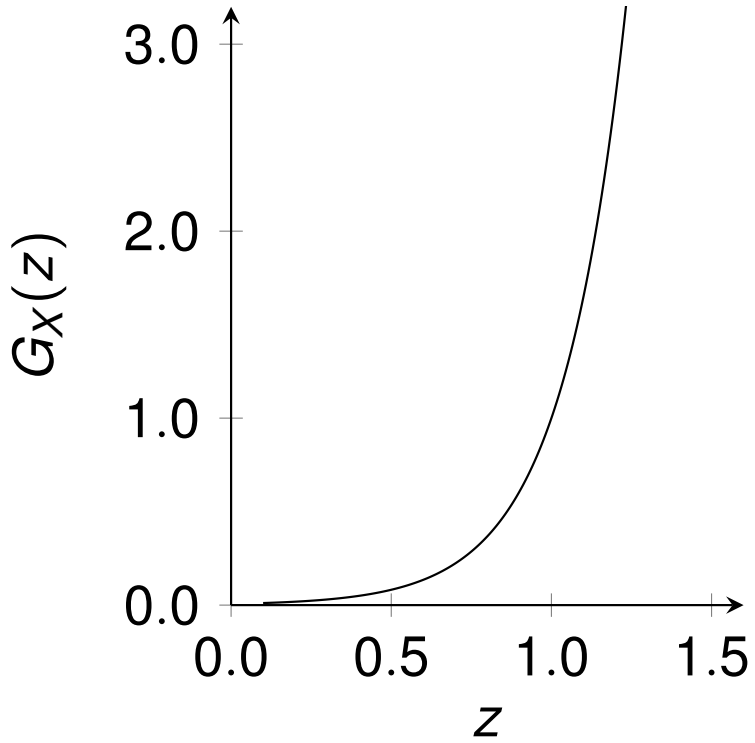


Example (Poisson distribution, Pois(λ))

$$G_X(z) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} z^k = e^{\lambda z} e^{-\lambda} = e^{\lambda(z-1)}.$$

E.g. $\lambda = 5$





Derivatives of the PGF

We can derive a very useful property of the PGF by considering the derivative, $G'_X(z)$, with respect to z. Assume we can interchange the order of differentiation and summation, so that

$$\begin{aligned} G'_X(z) &= \frac{d}{dz}\left(\sum_{k=0}^{\infty} z^k \mathbb{P}(X=k)\right) \\ &= \sum_{k=0}^{\infty} \frac{d}{dz}\left(z^k\right)\mathbb{P}(X=k) \\ &= \sum_{k=0}^{\infty} k z^{k-1}\mathbb{P}(X=k) \end{aligned}$$

then putting $z=1$ we have that

$$G'_X(1) = \sum_{k=0}^{\infty} k\mathbb{P}(X=k) = \mathbb{E}(X)$$

the expectation of the RV X.

Further derivatives of the PGF

Taking the second derivative gives

$$G''_X(z) = \sum_{k=0}^{\infty} k(k-1)z^{k-2}\mathbb{P}(X=k)\,.$$

So that,

$$G''_X(1) = \sum_{k=0}^{\infty} k(k-1)\mathbb{P}(X=k) = \mathbb{E}(X(X-1))$$

Generally, we have the following result.

Theorem

If the RV X has PGF $G_X(z)$ then the r^{th} derivative of the PGF, written $G_X^{(r)}(z)$, evaluated at $z=1$ is such that

$$G_X^{(r)}(1) = \mathbb{E}\left(X(X-1)\cdots(X-r+1)\right).$$

Using the PGF to calculate $\mathbb{E}(X)$ and Var(X)

We have that

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

and

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= [\mathbb{E}(X(X-1)) + \mathbb{E}(X)] - (\mathbb{E}(X))^2 \\ &= G''_X(1) + G'_X(1) - G'_X(1)^2 .\end{aligned}$$

For example, if X is a RV with the Pois(λ) distribution then $G_X(z) = e^{\lambda(z-1)}$. Thus, $G'_X(z) = \lambda e^{\lambda(z-1)}$, $G''_X(z) = \lambda^2 e^{\lambda(z-1)}$ and so $G'_X(1) = \lambda$ and $G''_X(1) = \lambda^2$. So, finally,

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

Sums of independent random variables

The following theorem shows how PGFs can be used to find the PGF of the sum of independent RVs.

Theorem

If X and Y are independent RVs with PGFs $G_X(z)$ and $G_Y(z)$ respectively then

$$G_{X+Y}(z) = G_X(z)G_Y(z)\,.$$

Proof.

Using the independence of X and Y we have that

$$\begin{aligned} G_{X+Y}(z) &= \mathbb{E}(z^{X+Y}) \\ &= \mathbb{E}(z^X z^Y) \\ &= \mathbb{E}(z^X)\mathbb{E}(z^Y) \\ &= G_X(z)G_Y(z) \end{aligned}$$

□

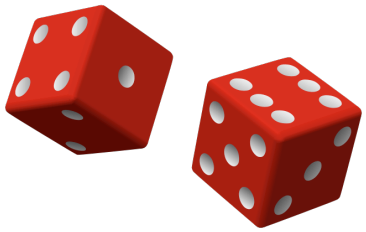
PGF example: Poisson RVs

For example, suppose that X and Y are independent RVs with $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, respectively.
Then

$$\begin{aligned} G_{X+Y}(z) &= G_X(z)G_Y(z) \\ &= e^{\lambda_1(z-1)}e^{\lambda_2(z-1)} \\ &= e^{(\lambda_1+\lambda_2)(z-1)}. \end{aligned}$$

Hence $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$ is again a Poisson RV but with the parameter $\lambda_1 + \lambda_2$.

PGF example: Uniform RVs

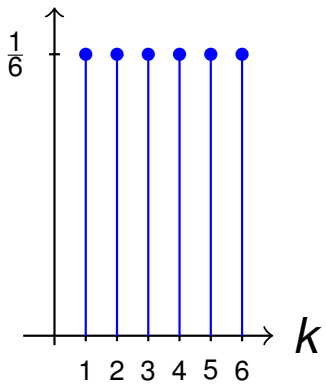


Consider the case of two fair dice with IID outcomes X and Y, respectively, so that $X \sim U(1,6)$ and $Y \sim U(1,6)$. Let the total score be $T = X + Y$ and consider the PGF of T given by $G_T(z) = G_X(z)G_Y(z)$. Then

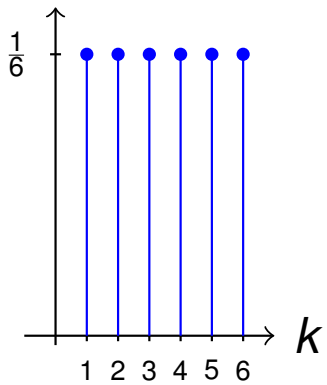
$$G_T(z) = \sum_{k=0}^{\infty} p_k z^k = \frac{1}{6}(z + z^2 + \cdots + z^6)\frac{1}{6}(z + z^2 + \cdots + z^6)$$

$$= \frac{1}{36}\left[z^2 + 2z^3 + 3z^4 + 4z^5 + 5z^6 + 6z^7 + 5z^8 + 4z^9 + 3z^{10} + 2z^{11} + z^{12}\right].$$

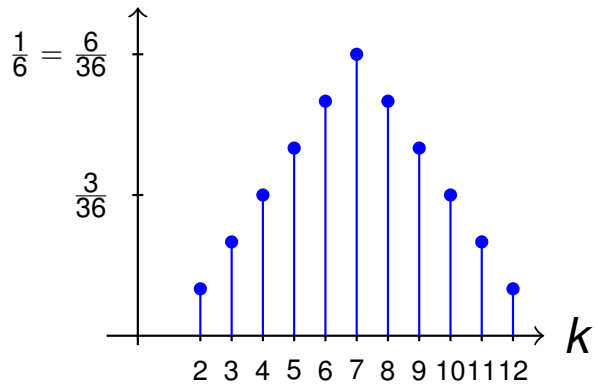
$\mathbb{P}(X = k)$



$\mathbb{P}(Y = k)$



$\mathbb{P}(T = k)$



Limits and inequalities

Limits and inequalities

We are familiar with limits of real numbers. For example, if $x_n = 1/n$ for $n = 1, 2, \ldots$ then $\lim_{n\to\infty} x_n = 0$ whereas if $x_n = (-1)^n$ no such limit exists. Behaviour in the long-run or on average is an important characteristic of everyday life.

We will be concerned with these notions of limiting behaviour when the real numbers x_n are replaced by random variables X_n. As we shall see there are several distinct notions of convergence that can be considered.

To study these forms of convergence and the limiting theorems that emerge we shall also gather a very useful collection of concepts and tools for the probabilistic analysis of models, algorithms and systems.

Probabilistic inequalities

To help assess how close RVs are to each other it is useful to have methods that provide upper bounds on probabilities of the form

$$\mathbb{P}(X \geq a)$$

for fixed constants a.

We shall consider several such bounds and related inequalities.

- Markov's inequality
- Chebyshev's inequality
- Chernoff's inequality

We will use $\mathbb{I}(A)$ for the indicator RV which is 1 if A occurs and 0 otherwise. Observe that we have for such indicator RVs that

$$\mathbb{E}(\mathbb{I}(A)) = \mathbb{P}(A)$$

since

$$\mathbb{E}(\mathbb{I}(A)) = 0 \times \mathbb{P}(A^c) + 1 \times \mathbb{P}(A) = \mathbb{P}(A)\,.$$

Theorem (Markov's inequality)

If $\mathbb{E}(X) < \infty$ *then for any* $a > 0$,

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}\,.$$

Proof.

We have that

$$\mathbb{I}(|X| \geq a) = \begin{cases} 1 & |X| \geq a \\ 0 & \text{otherwise}\,. \end{cases}$$

Clearly,

$$|X| \geq a\mathbb{I}(|X| \geq a)$$

hence

$$\mathbb{E}(|X|) \geq \mathbb{E}(a\mathbb{I}(|X| \geq a)) = a\mathbb{P}(|X| \geq a)$$

which yields the result.

□

Theorem (Chebyshev's inequality)

Let X be a RV with mean $\mu = \mathbb{E}(X)$ and finite variance $\sigma^2 = Var(X)$ then for all $a > 0$

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Proof.

Put $Y = (X - \mu)^2 \geq 0$ then $\mathbb{E}(Y) = \mathbb{E}((X - \mu)^2) = \text{Var}(X) = \sigma^2$. So, by Markov's inequality, for all $b > 0$

$$\mathbb{P}((X - \mu)^2 \geq b) = \mathbb{P}(Y \geq b) \leq \frac{\mathbb{E}(Y)}{b} = \frac{\sigma^2}{b}.$$

Now put $b = a^2$ and noting that $\mathbb{P}((X - \mu)^2 \geq a^2) = \mathbb{P}(|X - \mu| \geq a)$ we have that

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

□

Moment generating function

Definition

The moment generating function (MGF) of a RV X, written $M_X(t)$, is given by

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

and is defined for those values of $t \in \mathbb{R}$ for which this expectation exists.

Using the power series $e^x = 1 + x + x^2/2! + x^3/3! + \cdots$ we see that

$$M_X(t) = \mathbb{E}(e^{tX}) = 1 + \mathbb{E}(X)t + \mathbb{E}(X^2)t^2/2! + \mathbb{E}(X^3)t^3/3! + \cdots$$

and so the n^{th} moment of X, $\mathbb{E}(X^n)$, is given by the coefficient of $t^n/n!$ in the power series expansion of the MGF $M_X(t)$.

Note that for every RV, X, we have that $M_X(0) = 1$ since

$$M_X(0) = \mathbb{E}(e^{0X}) = \mathbb{E}(1) = 1\,.$$

Elementary properties of the MGF

1. If X has MGF $M_X(t)$ then $Y = aX + b$ has MGF $M_Y(t) = e^{bt} M_X(at)$.
2. If X and Y are independent then $X + Y$ has MGF $M_{X+Y}(t) = M_X(t) M_Y(t)$.
3. $\mathbb{E}(X^n) = M_X^{(n)}(0)$ where $M_X^{(n)}$ is the n^{th} derivative of M_X.
4. If X is a discrete RV taking values $0, 1, 2, \ldots$ with PGF $G_X(z) = \mathbb{E}(z^X)$ then $M_X(t) = G_X(e^t)$.

Fundamental properties of the MGF

We will use without proof the following results.

1. Uniqueness: to each MGF there corresponds a unique distribution function having that MGF.
   In fact, if X and Y are RVs with the same MGF in some region $-a < t < a$ where $a > 0$ then X and Y have the same distribution.
2. Continuity: if distribution functions $F_n(x)$ converge pointwise to a distribution function $F(x)$, the corresponding MGFs (where they exist) converge to the MGF of $F(x)$. Conversely, if a sequence of MGFs $M_n(t)$ converge to $M(t)$ which is continuous at $t = 0$, then $M(t)$ is a MGF, and the corresponding distribution functions $F_n(x)$ converge to the distribution function determined by $M(t)$.

Example: exponential distribution

If X has an exponential distribution with parameter $\lambda > 0$ then $f_X(x) = \lambda e^{-\lambda x}$ for $0 < x < \infty$. Hence, for t fixed and for $t < \lambda$,

$$M_X(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \int_0^\infty \lambda e^{-(\lambda - t)x} dx$$
$$= \left[-\frac{\lambda}{(\lambda - t)} e^{-(\lambda - t)x} \right]_0^\infty = \frac{\lambda}{\lambda - t}.$$

For $t < \lambda$

$$\frac{\lambda}{(\lambda - t)} = \frac{1}{1 - t/\lambda} = \left(1 - \frac{t}{\lambda}\right)^{-1} = 1 + \frac{t}{\lambda} + \frac{t^2}{\lambda^2} + \cdots$$

and hence $\mathbb{E}(X) = 1/\lambda$ and $\mathbb{E}(X^2) = 2/\lambda^2$ so that

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 1/\lambda^2.$$

Example: normal distribution

Consider a normal RV $X \sim N(\mu, \sigma^2)$ then $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ so that

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(-2tx\sigma^2+(x-\mu)^2)/2\sigma^2} dx\,.$$

So, by *'completing the square'*,

$$M_X(t) = e^{\mu t+\sigma^2 t^2/2} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-(\mu+t\sigma^2))^2/2\sigma^2} dx \right\}$$
$$= e^{\mu t+\sigma^2 t^2/2}\,.$$

Example: uniform distribution

Consider a uniform RV $X \sim U(a,b)$ for $a < b$. Then

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise}. \end{cases}$$

Hence, for $t \neq 0$,

$$\begin{aligned} M_X(t) &= \int_a^b \frac{e^{tx}}{b-a} dx \\ &= \left[\frac{e^{tx}}{(b-a)t} \right]_a^b \\ &= \frac{e^{bt} - e^{at}}{(b-a)t}. \end{aligned}$$

and $M_X(0) = 1$.

Theorem (Chernoff's bound)

Suppose that X has MGF $M_X(t)$ and $a \in \mathbb{R}$ then for all $t > 0$

$$\mathbb{P}(X \geq a) \leq e^{-ta} M_X(t)\,.$$

Proof.

Using Markov's inequality and noting that for $t > 0$ then e^{tx} is a non-decreasing function of x we have that

$$\begin{aligned}\mathbb{P}(X \geq a) &= \mathbb{P}(e^{tX} \geq e^{ta}) \\ &\leq \frac{\mathbb{E}(e^{tX})}{e^{ta}} \\ &= e^{-ta} M_X(t)\end{aligned}$$

□

Note that the above bound holds for all $t > 0$ so we can select the best such bound by choosing $t > 0$ to minimize $e^{-ta} M_X(t)$.
In fact, the upper bound also holds trivially if $t = 0$ since the RHS is 1.

Notions of convergence: $X_n \to X$ as $n \to \infty$

For a sequence of RVs $(X_n)_{n \geq 1}$, we shall define two distinct notions of convergence to some RV X as $n \to \infty$.

Definition (Convergence in distribution)

$X_n \xrightarrow{D} X$ if $F_{X_n}(x) \to F_X(x)$ for all points x at which F_X is continuous.

Definition (Convergence in probability)

$X_n \xrightarrow{P} X$ if $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ for all $\varepsilon > 0$.

There are further inter-related notions of convergence but two will suffice for our purposes.

Theorem
If $X_n \xrightarrow{P} X$ *then* $X_n \xrightarrow{D} X$.

Proof

We prove this theorem as follows. Fix, $\varepsilon > 0$ then

$$F_{X_n}(x) = \mathbb{P}(X_n \leq x \cap X > x + \varepsilon) + \mathbb{P}(X_n \leq x \cap X \leq x + \varepsilon)$$

since $X > x + \varepsilon$ and $X \leq x + \varepsilon$ form a partition. But if $X_n \leq x$ and $X > x + \varepsilon$ then $|X_n - X| > \varepsilon$
and $\{X_n \leq x \cap X \leq x + \varepsilon\} \subset \{X \leq x + \varepsilon\}$. Therefore,

$$F_{X_n}(x) \leq \mathbb{P}(|X_n - X| > \varepsilon) + F_X(x + \varepsilon)\,.$$

Similarly,

$$\begin{aligned} F_X(x - \varepsilon) &= \mathbb{P}(X \leq x - \varepsilon \cap X_n > x) + \mathbb{P}(X \leq x - \varepsilon \cap X_n \leq x) \\ &\leq \mathbb{P}(|X_n - X| > \varepsilon) + F_{X_n}(x)\,. \end{aligned}$$

Proof, ctd

The proof is completed by noting that together these inequalities show that

$$F_X(x-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_{X_n}(x) \leq \mathbb{P}(|X_n - X| > \varepsilon) + F_X(x+\varepsilon).$$

But $X_n \xrightarrow{P} X$ implies that $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$. So, as $n \to \infty$, $F_{X_n}(x)$ is *'squeezed'* between $F_X(x-\varepsilon)$ and $F_X(x+\varepsilon)$.

Hence, if F_X is continuous at x, $F_{X_n}(x) \to F_X(x)$ and so $X_n \xrightarrow{D} X$. □

Remarks

1. Note that the converse does not hold in general. An exercise on the problem sheet provides a counterexample.
2. Another exercise on the problem sheet shows an important and useful special case where the converse does hold.

Limit theorems

Given a sequence of RVs $(X_n)_{n\geq 1}$, let

$$S_n = X_1 + X_2 + \cdots + X_n \qquad \text{and} \qquad \overline{X}_n = S_n/n.$$

What happens to the sample mean, $\overline{X}_n$, for large n?

Theorem (Weak Law of Large Numbers/WLLN)

Suppose $(X_n)_{n\geq 1}$ are IID RVs with finite mean μ (and finite variance σ^2) then $\overline{X}_n \xrightarrow{P} \mu$.

Note that convergence to μ in the WLLN actually means convergence to a degenerate RV, X, with $\mathbb{P}(X = \mu) = 1$.

Aside: this is referred to as the *weak* law of large numbers since under more restrictive assumptions it holds for a *stronger* form of convergence known as almost sure convergence. Under the strong law of large numbers (SLLN) with almost sure convergence we would have that $\mathbb{P}(\overline{X}_n \to \mu) = 1$.

WLLN

Theorem (Weak Law of Large Numbers/WLLN)

Suppose $(X_n)_{n\geq 1}$ *are IID RVs with finite mean* μ *and finite variance* σ^2 *then* $\overline{X}_n \xrightarrow{P} \mu$.

Proof.

Recall that $\mathbb{E}(\overline{X}_n) = \mu$ and $\text{Var}(\overline{X}_n) = \sigma^2/n$. Hence, by Chebyshev's inequality applied to $\overline{X}_n$ for all $\varepsilon > 0$

$$0 \leq \mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

and so, letting $n \to \infty$,

$$\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \to 0$$

hence $\overline{X}_n \xrightarrow{P} \mu$ as required. □

Applications: estimating probabilities

Suppose we wish to estimate the probability, p, that we succeed when we play some game or perform some experiment. For $i = 1, \ldots, n$, let

$$X_i = \mathbb{I}(\{i^{th}\text{game is success}\}).$$

So $\overline{X}_n = m/n$ if we succeed m times in n attempts.
We have that $\mu = \mathbb{E}(X_i) = \mathbb{P}(X_i = 1) = p$ so then

$$\overline{X}_n \xrightarrow{P} \mu$$

that is

$$m/n \xrightarrow{P} p$$

by the WLLN.
Thus we have shown the important result that the empirical estimate of the probability of some event by its observed sample frequency converges in probability to the correct but usually unknown value as the number of samples grows.
This result forms the basis of all simulation methods.

Monte Carlo simulation and randomized algorithms

Suppose we wish to estimate the value of π. One way to proceed is to perform the following experiment. Select a point $(X, Y) \in [-1,1]^2$ with X and Y chosen independently and uniformly in $[-1,1]$. Now consider those points within unit distance of the origin then

$$\mathbb{P}((X,Y) \text{ lies in unit circle}) = \mathbb{P}(X^2 + Y^2 \leq 1) = \frac{\text{area of circle}}{\text{area of square}} = \frac{\pi}{4}.$$

Suppose we have access to a stream of random variables $U_i \sim U(0,1)$ then $2U_i - 1 \sim U(-1,1)$. Now set $X_i = 2U_{2i-1} - 1$, $Y_i = 2U_{2i} - 1$ and $H_i = \mathbb{I}(\{X_i^2 + Y_i^2 \leq 1\})$ so that

$$\mathbb{E}(H_i) = \mathbb{P}(X_i^2 + Y_i^2 \leq 1) = \frac{\pi}{4}.$$

Hence by the WLLN the proportion of points (X_i, Y_i) falling within the unit circle converges in probability to $\pi/4$. Furthermore, the CLT can be used to form confidence intervals.
This a simple example of a randomized algorithm to solve a deterministic problem.

Central limit theorem

Theorem (Central limit theorem/CLT)

Let $(X_n)_{n\geq 1}$ be a sequence of IID RVs with mean μ, variance σ^2 and whose moment generating function exists in some interval $-a < t < a$ with $a > 0$. Then, as $n \to \infty$

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z$$

where $Z \sim N(0,1)$.

We will now prove this extremely useful result using an approach based on the properties of moment generating functions.

Proof of CLT

Set $Y_i = (X_i - \mu)/\sigma$ then $\mathbb{E}(Y_i) = 0$ and $\mathbb{E}(Y_i^2) = \mathrm{Var}(Y_i) = 1$ so

$$M_{Y_i}(t) = 1 + \frac{t^2}{2} + o(t^2)$$

where $o(t^2)$ refers to terms of higher order than t^2 which will therefore tend to 0 faster than t^2 as $t \to 0$. Also,

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \,.$$

Hence,

$$\begin{aligned} M_{Z_n}(t) &= \left(M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &= \left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \\ &\to e^{t^2/2} \qquad \text{as} \qquad n \to \infty \,. \end{aligned}$$

But $e^{t^2/2}$ is the MGF of the $N(0,1)$ distribution so, together with the continuity property, $Z_n \xrightarrow{D} Z \sim N(0,1)$ and the CLT holds.

CLT example

Suppose $X_1, X_2, \ldots, X_n$ are the IID RVs showing the n sample outcomes of a 6-sided die with common distribution

$$\mathbb{P}(X_i = j) = p_j\,, \qquad j = 1, 2, \ldots, 6$$

Set $S_n = X_1 + X_2 + \cdots + X_n$, the total score obtained, and consider the two cases

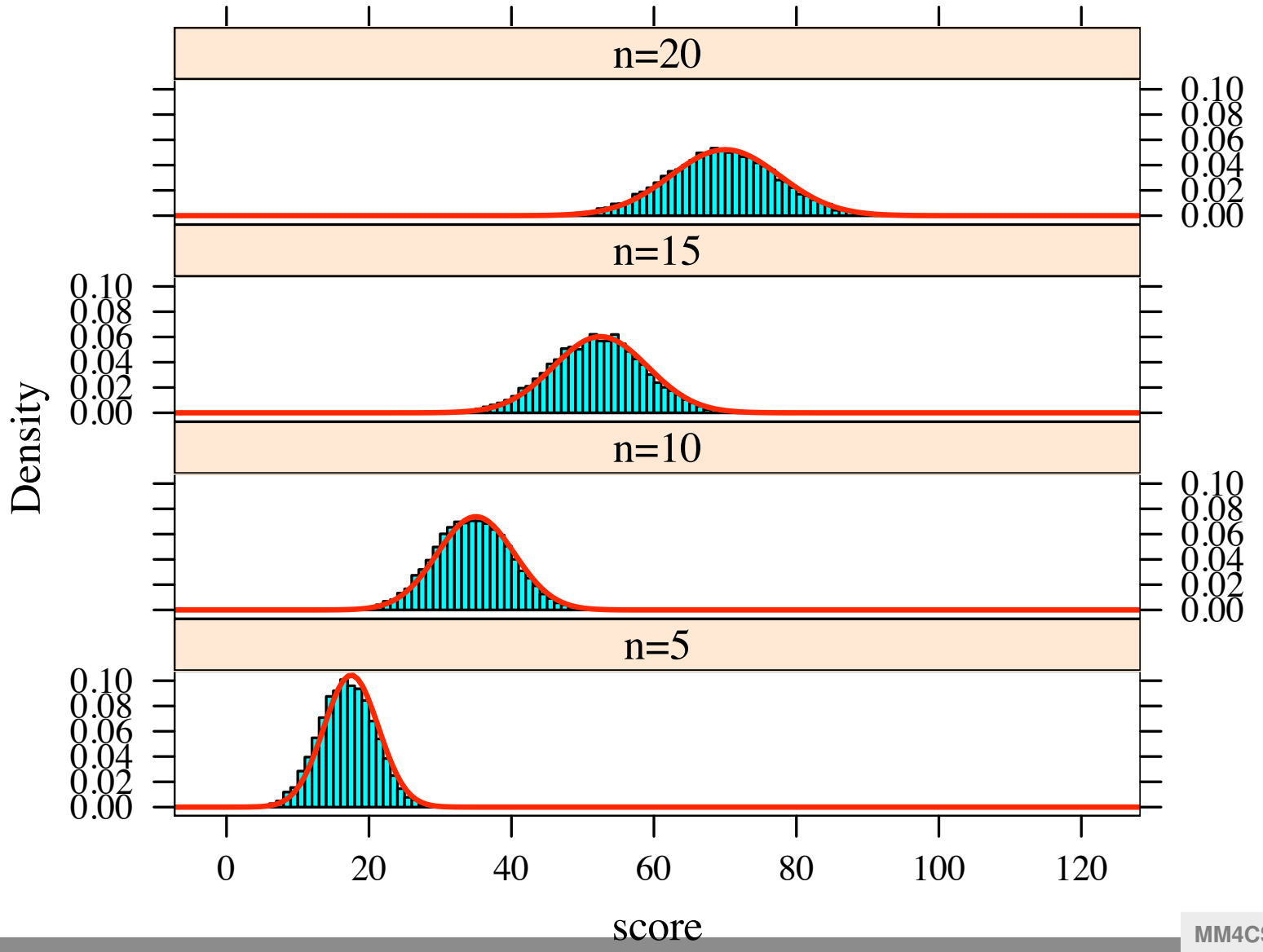
- symmetric: $(p_j) = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$ so that $\mu = \mathbb{E}(X_i) = 3.5$ and $\sigma^2 = \mathrm{Var}(X_i) \approx 2.9$
- asymmetric: $(p_j) = (0.2, 0.1, 0.0, 0.0, 0.3, 0.4)$ so that $\mu = \mathbb{E}(X_i) = 4.3$ and $\sigma^2 = \mathrm{Var}(X_i) \approx 4.0$

for varying sample sizes $n = 5, 10, 15$ and 20.

The CLT tells us that for large n, S_n is approximately distributed as $N(n\mu, n\sigma^2)$ where μ and σ^2 are the mean and variance, respectively, of X_i.

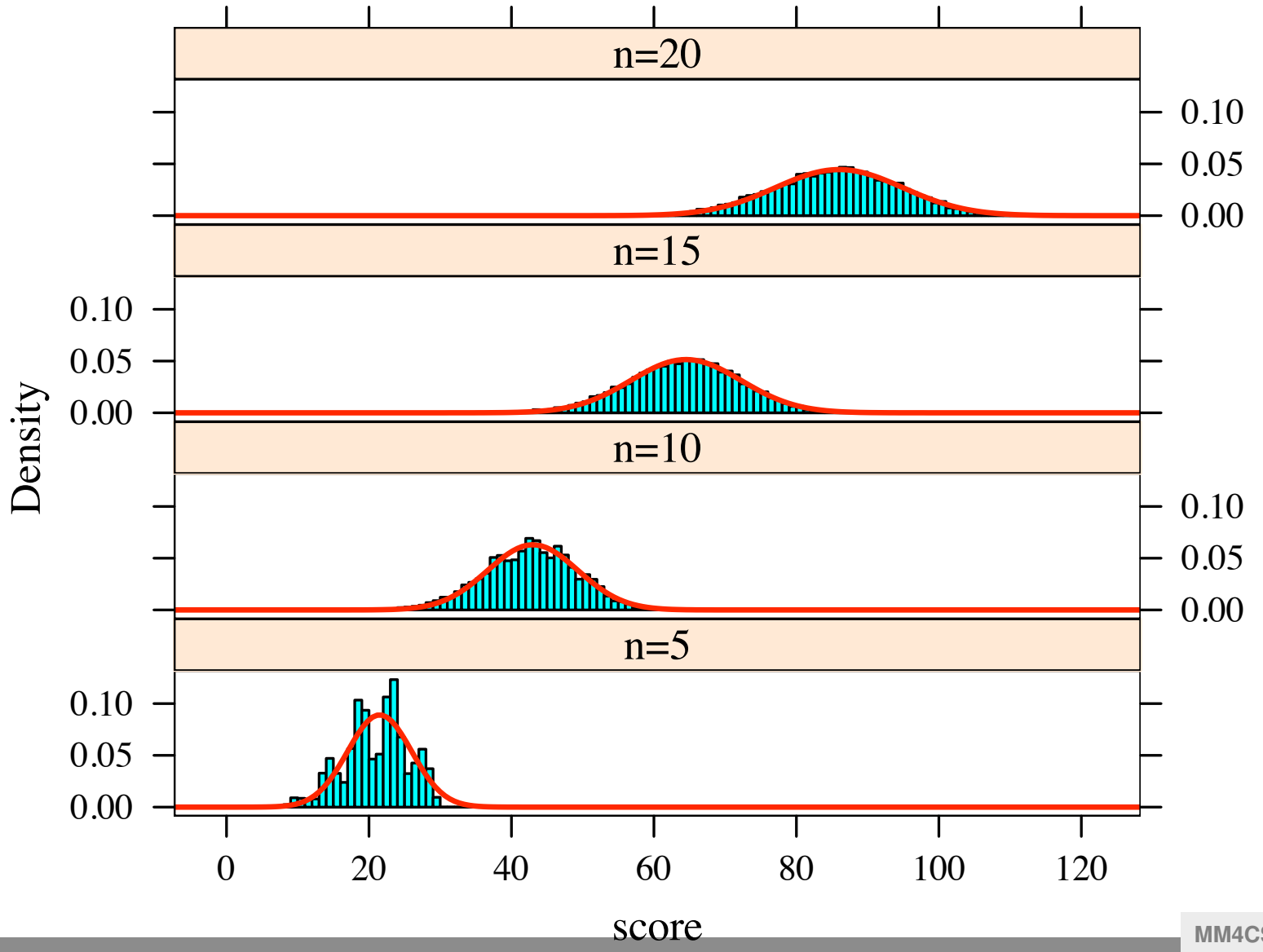
CLT example: symmetric

10,000 replications



CLT example: asymmetric

10,000 replications



Confidence intervals I

One of the major statistical applications of the CLT is to the construction of confidence intervals. The CLT shows that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

is asymptotically distributed as $N(0,1)$. If, the true value of σ^2 is unknown we may estimate it by the sample variance given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 .$$

For instance, it can be shown that $\mathbb{E}(S^2) = \sigma^2$ and then

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}}$$

is approximately distributed as $N(0,1)$ for large n.

Confidence intervals II

Define z_α so that $\mathbb{P}(Z > z_\alpha) = \alpha$ where $Z \sim N(0,1)$ and so

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha\,.$$

Hence,

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\overline{X}_n - \mu}{S/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

$$\mathbb{P}\left(\overline{X}_n - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2}\frac{S}{\sqrt{n}}\right) \approx 1 - \alpha\,.$$

The interval between the pair of end points $\overline{X}_n \pm z_{\alpha/2} S/\sqrt{n}$ is thus an (approximate) $100(1-\alpha)$ percent confidence interval for the unknown parameter μ.

Confidence intervals: example

Consider a collection of n IID RVs, X_i, with common distribution $X_i \sim \text{Pois}(\lambda)$. Hence,

$$\mathbb{P}(X_i = j) = \frac{\lambda^j e^{-\lambda}}{j!} \qquad j = 0, 1, \ldots$$

with mean $\mathbb{E}(X_i) = \lambda$.
Then a 95% confidence interval for the (unknown) mean value λ is given by

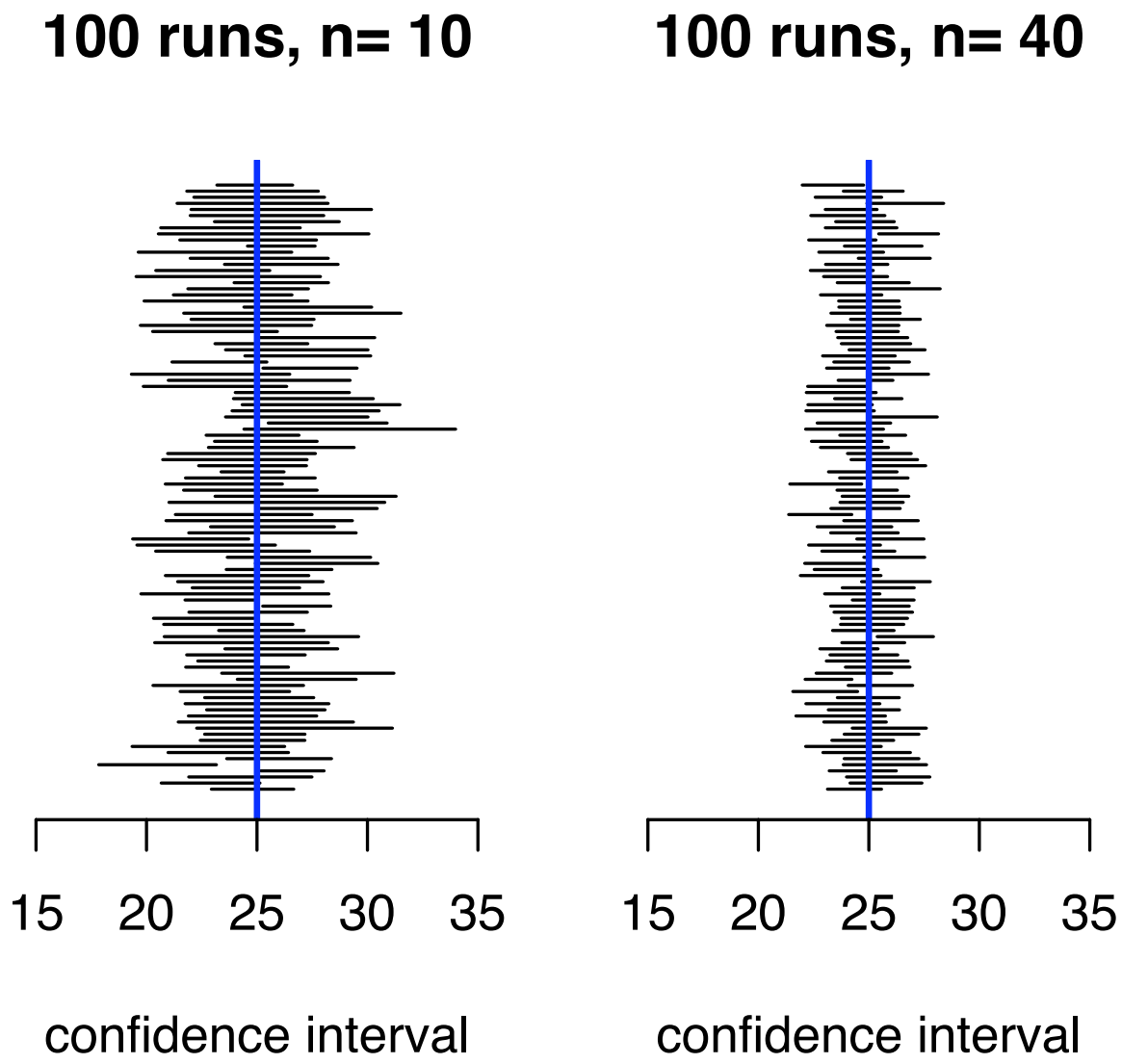
$$\overline{X}_n \pm 1.96 S/\sqrt{n}$$

where $z_{0.025} = 1.96$.
Alternatively, to obtain 99% confidence intervals replace 1.96 by $z_{0.005} = 2.58$ for a confidence interval $\overline{X}_n \pm 2.58 S/\sqrt{n}$.

α	$\alpha/2$	$z_{\alpha/2}$	$100(1-\alpha)\%$
0.05	0.025	1.96	95%
0.01	0.005	2.58	99%

95% confidence intervals

Illustration with $\lambda = 25$ and $\alpha = 5\%$



Stochastic processes

Random walks

Consider a sequence $Y_1, Y_2, \ldots$ of IID RVs with $\mathbb{P}(Y_i = 1) = p$ and $\mathbb{P}(Y_i = -1) = 1 - p$ with $p \in [0,1]$.

Definition (Simple random walk)

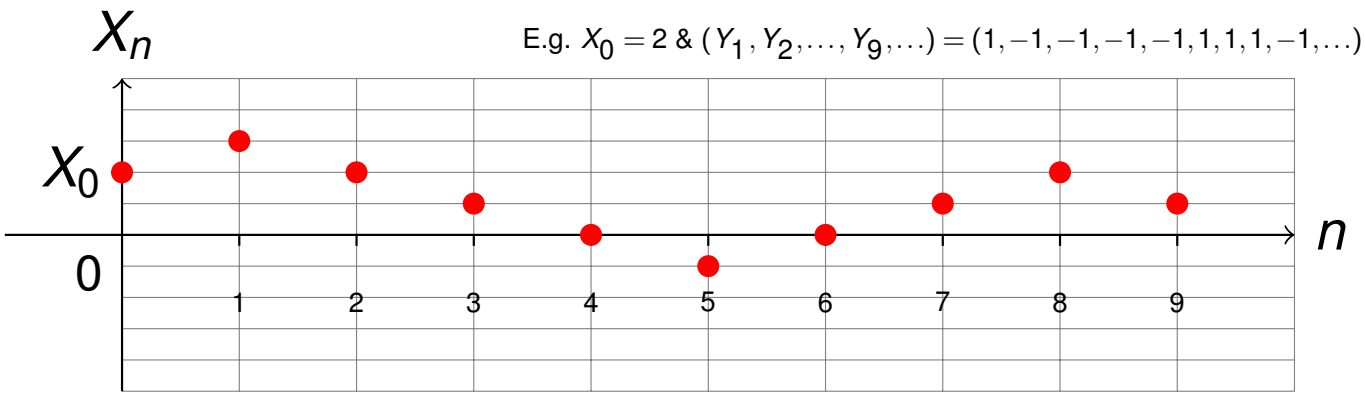
The simple random walk is a sequence of RVs $\{X_n \,|\, n \in \{1,2,\ldots\}\}$ defined by

$$X_n = X_0 + Y_1 + Y_2 + \cdots + Y_n$$

where $X_0 \in \mathbb{R}$ is the starting value.

Definition (Simple symmetric random walk)

A simple symmetric random walk is a simple random walk with the choice $p = 1/2$.



Examples

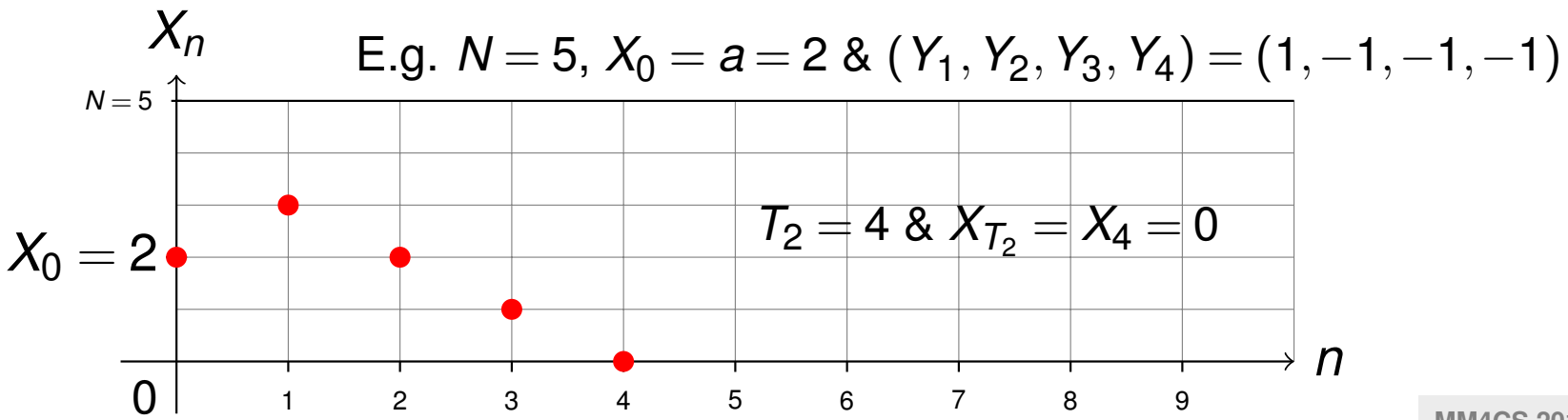
Practical examples of random walks abound across the physical sciences (motion of atomic particles) and the non-physical sciences (epidemics, gambling, asset prices, PageRank, cryptocurrencies (Bitcoin)).

The following is a simple model for the operation of a casino. Suppose that a gambler enters with a capital of £X_0. At each stage the gambler places a stake of £1 and with probability p wins the gamble otherwise the stake is lost. If the gambler wins the stake is returned together with an additional sum of £1.
Thus at each stage the gambler's capital increases by £1 with probability p or decreases by £1 with probability $1-p$.
The gambler's capital X_n at stage n thus follows a simple random walk except that the gambler is bankrupt if X_n reaches £0 and then can not continue to any further stages.

The Gambler's ruin problem

We now consider a variant of the simple random walk. Consider two players A and B with a joint capital between them of £N. Suppose that initially A has $X_0 = £a$ $(0 \leq a \leq N)$.
At each time step player B gives A £1 with probability p and with probability $q = (1-p)$ player A gives £1 to B instead. The outcomes at each time step are independent and fix $p \in (0,1)$.
The game ends at the first time T_a if either $X_{T_a} = £0$ or $X_{T_a} = £N$ for some $T_a \in \{0,1,\ldots\}$.
We can think of A's wealth, X_n, at time n as a simple random walk on the states $\{0,1,\ldots,N\}$ with absorbing barriers at 0 and N.
Define the probability of ruin, ρ_a, for gambler A as

$$\rho_a = \mathbb{P}(\text{A is ruined}) = \mathbb{P}(\text{B wins}) \quad \text{for } 0 \leq a \leq N.$$



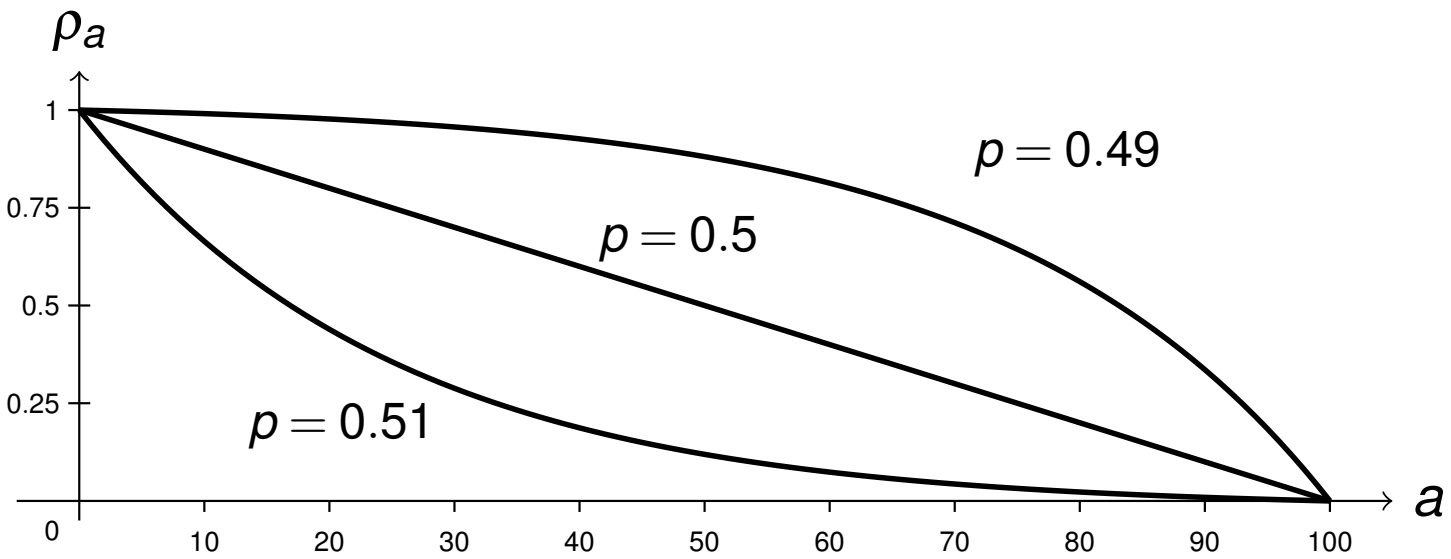
Solution of the Gambler's ruin problem

Theorem

The probability of ruin when A starts with an initial capital of £a is given by

$$\rho_a = \begin{cases} \dfrac{\left(\frac{q}{p}\right)^a - \left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N} & \text{if } p \neq q \\ 1 - \dfrac{a}{N} & \text{if } p = q = 1/2. \end{cases}$$

For illustration here is a set of graphs of ρ_a for $N = 100$ and three possible choices of p.



Proof

It will be convenient to work with

$$\alpha_a = \mathbb{P}(\text{A wins} \,|\, X_0 = a) = 1 - \mathbb{P}(\text{A is ruined} \,|\, X_0 = a) = 1 - \rho_a$$

and set $\theta = q/p$ (recall that $p \neq 0$) and so $\theta = 1$ if and only if $p = 0.5$.

Consider what happens at the first time step then for $0 < a < N$

$$\begin{aligned}
\alpha_a &= \mathbb{P}(\text{A wins} \cap Y_1 = +1 \,|\, X_0 = a) + \mathbb{P}(\text{A wins} \cap Y_1 = -1 \,|\, X_0 = a) \\
&= p\mathbb{P}(\text{A wins} \,|\, X_0 = a+1) + q\mathbb{P}(\text{A wins} \,|\, X_0 = a-1) \\
&= p\alpha_{a+1} + q\alpha_{a-1} \,.
\end{aligned}$$

We now proceed to solve this set of difference equations with the boundary conditions $\alpha_0 = 0$ and $\alpha_N = 1$.

Proof, ctd

Since $p+q=1$ we can rewrite the difference equations as

$$\begin{aligned}
\alpha_a = (p+q)\alpha_a &= p\alpha_{a+1} + q\alpha_{a-1} \\
p(\alpha_{a+1} - \alpha_a) &= q(\alpha_a - \alpha_{a-1}) \\
\alpha_{a+1} - \alpha_a &= \theta(\alpha_a - \alpha_{a-1}) \\
&= \theta^2(\alpha_{a-1} - \alpha_{a-2}) \\
&= \cdots \\
&= \theta^a(\alpha_1 - \alpha_0) = \theta^a \alpha_1
\end{aligned}$$

using in the last step the boundary condition that $\alpha_0 = 0$.

Now consider

$$\alpha_a = \alpha_a - \alpha_0 = \sum_{i=1}^{a} (\alpha_i - \alpha_{i-1}) = \sum_{i=1}^{a} \theta^{i-1} \alpha_1 \,.$$

Proof, ctd

But

$$\sum_{i=1}^{a} \theta^{i-1} = (1 + \theta + \theta^2 + \cdots + \theta^{a-1}) = \begin{cases} \frac{1-\theta^a}{1-\theta} & \theta \neq 1 \\ a & \theta = 1 \end{cases}$$

and so using $\alpha_N = 1$ gives

$$\alpha_N = 1 = \sum_{i=1}^{N} \theta^{i-1} \alpha_1 = \begin{cases} \frac{(1-\theta^N)\alpha_1}{1-\theta} & \theta \neq 1 \\ N\alpha_1 & \theta = 1 \end{cases}$$

and so for $0 < a < N$

$$\alpha_a = \begin{cases} \frac{1-\theta^a}{1-\theta^N} & \theta \neq 1 \\ \frac{a}{N} & \theta = 1\,. \end{cases}$$

and the theorem holds after the substitutions $\theta = q/p$ and $\rho_a = 1 - \alpha_a$.

Mean duration time

Set T_a as the time to be absorbed at either 0 or N starting from the initial state a and write $\mu_a = \mathbb{E}(T_a)$.

Then conditioning on the first step as before leads to the difference equations

$$\mu_a = 1 + p\mu_{a+1} + q\mu_{a-1} \quad \text{for } 0 < a < N$$

and boundary conditions $\mu_0 = \mu_N = 0$.

It can be shown that the solution μ_a is given by

$$\mu_a = \begin{cases} \frac{1}{p-q}\left(N\frac{\left(\frac{q}{p}\right)^a - 1}{\left(\frac{q}{p}\right)^N - 1} - a\right) & \text{if } p \neq q \\ a(N-a) & \text{if } p = q = 1/2\,. \end{cases}$$

We skip the derivation here but an exercise on the problem sheet invites you to check that this solution obeys the stated difference equations.

Case study: Bitcoin

Bitcoin is based around a proof-of-work mechanism which uses a decentralised peer-to-peer network of workers (known as miners) to ensure (with high probability) that bitcoins are not double-spent. In order to achieve double spending of a bitcoin the attacker would need to create a longer block chain than the honest chain.
Suppose that the honest workers can produce blocks on average every T/p time units while the attacker can do so on average every T/q time units with $q = 1 - p < p$. If the (honest) seller waits for a given number n of blocks to be created then this would take on average nT/p time units. Thus the average number of blocks that the attacker could create, m, would be such that $nT/p = mT/q$.
Thus $m = nq/p$ independent of T.
If $q > p$ then surely the attacker can always catch up the honest workers however large a head start, n, is considered. What is the chance that the attacker could still catch up the honest chain when $q < p$?

Satoshi Nakamato
Bitcoin: A peer-to-peer electronic cash system.
http://bitcoin.org/bitcoin.pdf, 2008.

Bitcoin analysis using Gambler's ruin problem

The Bitcoin white paper proposes the simple probabilistic model that the random number of blocks, X, that the attacker could produce as the honest workers produce their fixed number, n, of blocks has a Poisson distribution with mean $\lambda = nq/p$. Thus,

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

for $k = 0, 1, \ldots$.
What is the chance that the attacker could then overtake the honest workers? This is precisely the Gambler's ruin problem starting from initial assets of $n - k$ with $\theta = q/p < 1$ and in the limit that the total wealth $N \to \infty$.
Recalling our expression for the ruin probabilities we have that

$$\mathbb{P}(\text{attacker catches up} \mid k \text{ blocks}) = \begin{cases} (q/p)^{n-k} & k \leq n \\ 1 & k > n. \end{cases}$$

Bitcoin: calculations

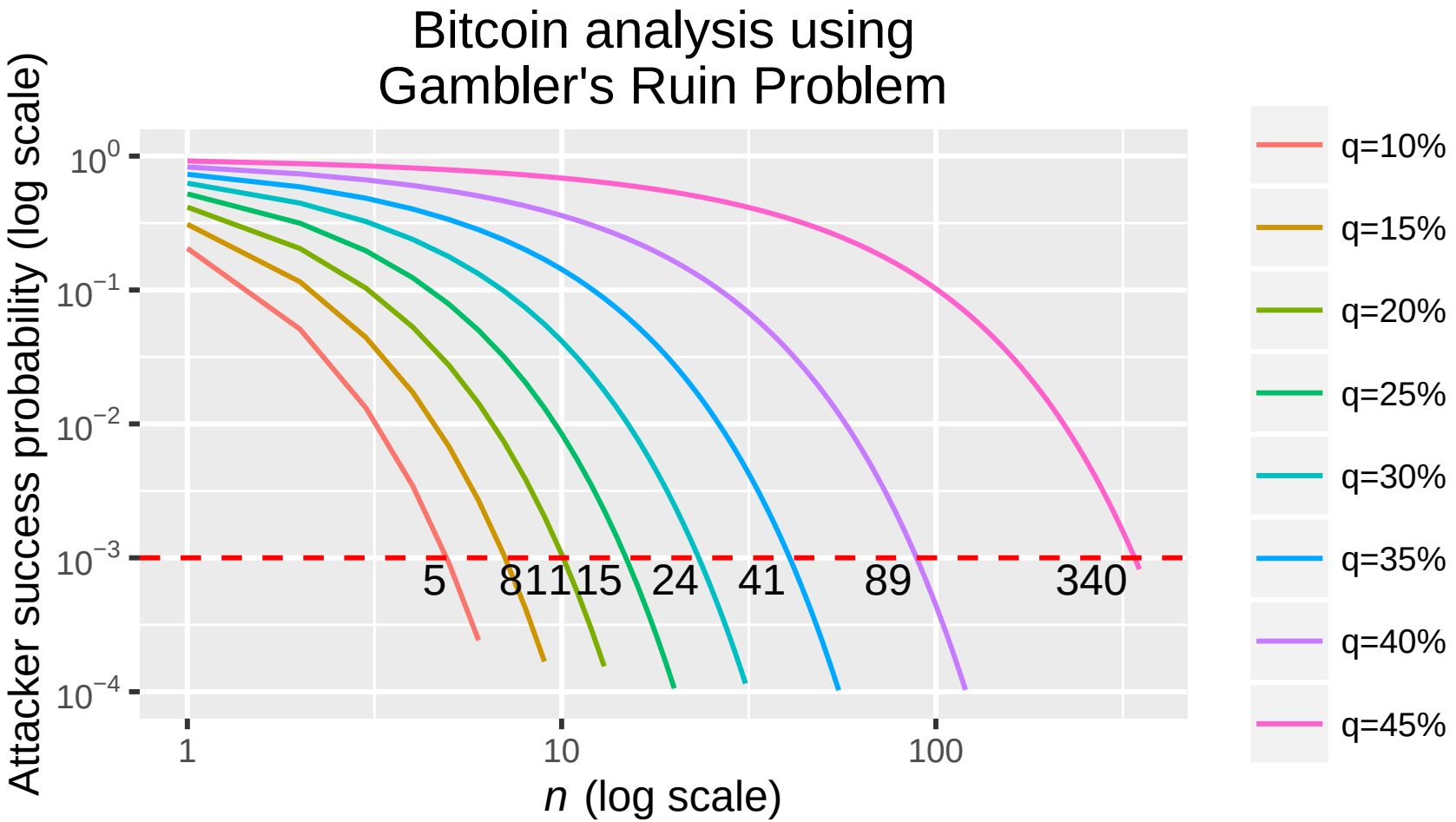
Hence, using the law of total probability,

$$\begin{aligned}
\mathbb{P}(\text{attacker catches up}) &= \sum_{k=0}^{\infty} \mathbb{P}(X=k)\mathbb{P}(\text{attacker catches up} \mid k \text{ blocks}) \\
&= \sum_{k=0}^{n} \mathbb{P}(X=k)(q/p)^{n-k} + \sum_{k=n+1}^{\infty} \mathbb{P}(X=k) \\
&= \sum_{k=0}^{n} \mathbb{P}(X=k)(q/p)^{n-k} + \left(1 - \sum_{k=0}^{n} \mathbb{P}(X=k)\right) \\
&= 1 - \sum_{k=0}^{n} \mathbb{P}(X=k)\left(1-(q/p)^{n-k}\right) \\
&= 1 - \sum_{k=0}^{n} \frac{\lambda^k e^{-\lambda}}{k!}\left(1-(q/p)^{n-k}\right)
\end{aligned}$$

where $\lambda = nq/p$ is the mean number of blocks that an attacker can produce while the honest workers produce n blocks.

Some illustrative computations: try these yourself!

The Bitcoin white paper (section 11) provides a C program to compute this probability for fixed $q = 1 - p$ and n. Observe that the probability of catching up the honest workers drops off rapidly with n.



Markov chains

Definition (Markov chain)

Suppose that (X_n) $n \geq 0$ is a sequence of discrete random variables taking values in some countable state space S. The sequence (X_n) is a Markov chain (MC) if

$$\mathbb{P}(X_n = x_n \,|\, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_n = x_n \,|\, X_{n-1} = x_{n-1})$$

for all $n \geq 1$ and for all $x_0, x_1, \ldots, x_n \in S$.

Since, S is countable we can always choose to label the possible values of X_n by integers and say that when $X_n = i$ the Markov chain is in the "i^{th} state at the n^{th} step" or "visits i at time n".

Occasionally, we shall restrict our results to the case of finite state spaces.

Transition probabilities

The dynamics of the Markov chain are governed by the transition probabilites $\mathbb{P}(X_n = j \,|\, X_{n-1} = i)$.

Definition (time-homogeneous MC)

A Markov chain (X_n) is time-homogeneous if

$$\mathbb{P}(X_n = j \,|\, X_{n-1} = i) = \mathbb{P}(X_1 = j \,|\, X_0 = i)$$

for all $n \geq 1$ and states $i, j \in S$.

- We shall assume that our MCs are time-homogeneous unless explicitly stated otherwise.

Transition matrix

Definition (Transition matrix)

The transition matrix, P, of a MC (X_n) is given by $P = (p_{ij})$ where for all $i, j \in S$

$$p_{ij} = \mathbb{P}(X_n = j \,|\, X_{n-1} = i)\,.$$

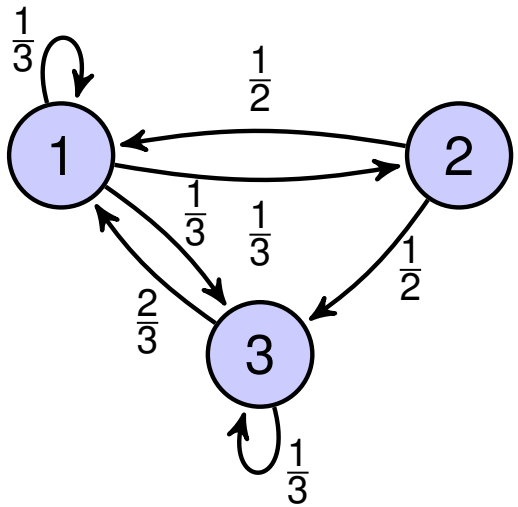
- Note that P is a stochastic matrix, that is, it has non-negative entries ($p_{ij} \geq 0$) and the row sums all equal one ($\sum_j p_{ij} = 1$).
- The transition matrix completely characterizes the dynamics of the MC.

Example

Suppose the states of the MC are $S = \{1,2,3\}$ and that the transition matrix is given by

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

- Thus, in state 1 we are equally likely to be in any of the three states at the next step.
- In state 2, we can move with equal probabilities to 1 or 3 at the next step.
- Finally in state 3, we either move to state 1 with probability $2/3$ or remain in state 3 at the next step.



n-step transition matrix

Definition (n-step transition matrix)

For $n \geq 0$ the n-step transition matrix is $P^{(n)} = (p_{ij}^{(n)})$ where

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i)\,.$$

Thus $P^{(1)} = P$ and we also set $P^{(0)} = I_{|S|}$, the $|S| \times |S|$-identity matrix.

Chapman-Kolmogorov equations

Theorem (Chapman-Kolmogorov)

For all states i, j and for all steps m, n

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

Hence, $P^{(m+n)} = P^{(m)} P^{(n)}$ and $P^{(n)} = P^n$, the n^{th} power of P.

Proof.

$$\begin{aligned} p_{ij}^{(m+n)} &= \mathbb{P}(X_{m+n} = j | X_0 = i) = \sum_{k \in S} \mathbb{P}(X_{m+n} = j, X_m = k | X_0 = i) \\ &= \sum_{k \in S} \mathbb{P}(X_{m+n} = j | X_m = k, X_0 = i) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_{k \in S} \mathbb{P}(X_{m+n} = j | X_m = k) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_{k \in S} p_{kj}^{(n)} p_{ik}^{(m)} \end{aligned}$$

The Chapman-Kolmorgorov equations tell us how the long-term evolution of the MC depends on the short-term evolution specified by the transition matrix.
If we let $\lambda_i^{(n)} = \mathbb{P}(X_n = i)$ be the elements of a row vector $\lambda^{(n)}$ specifying the distribution of the MC at the n^{th} time step then the follow holds.

Lemma

If m, n are non-negative integers then $\lambda^{(m+n)} = \lambda^{(m)} P^{(n)}$ and so, in particular, if $m = 0$

$$\lambda^{(n)} = \lambda^{(0)} P^{(n)}$$

where $\lambda^{(0)}$ is the initial distribution $\lambda_i^{(0)} = \mathbb{P}(X_0 = i)$ of the MC.

Proof.

$$\lambda_j^{(m+n)} = \mathbb{P}(X_{m+n} = j) = \sum_i \mathbb{P}(X_{m+n} = j | X_m = i)\mathbb{P}(X_m = i)$$

$$= \sum_i \lambda_i^{(m)} p_{ij}^{(n)} = \left(\lambda^{(m)} P^{(n)}\right)_j$$

Classification of states

Definition (Accessibility)

If, for some $n \geq 0$, $p_{ij}^{(n)} > 0$ then we say that state j is accessible from state i, written $i \rightsquigarrow j$.
If $i \rightsquigarrow j$ and $j \rightsquigarrow i$ then we say that i and j communicate, written $i \leftrightsquigarrow j$.

Observe that the relation communicates $\leftrightsquigarrow$ is

- reflexive
- symmetric
- transitive

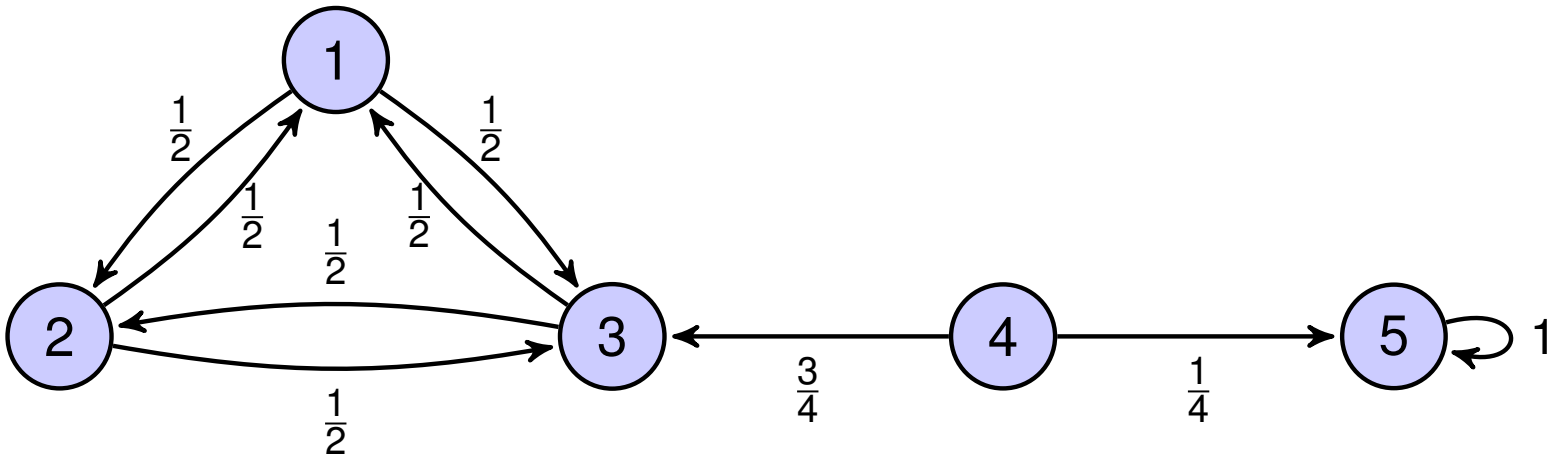
and hence is an equivalence relation. The corresponding equivalence classes partition the state space into subsets of states, called communicating classes.

Irreducibility

- A communicating class, C, that once entered can not be left is called closed, that is $p_{ij} = 0$ for all $i \in C, j \notin C$.
- A closed communicating class consisting of a single state is called absorbing.
- When the state space forms a single communicating class, the MC is called irreducible and is called reducible otherwise.

Example (5 states)

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$



- State space is reducible into three classes
- Classes: $\{1,2,3\}$, $\{4\}$ and $\{5\}$
- $\{1,2,3\}$ and $\{5\}$ are closed; $\{5\}$ is absorbing.

Recurrence and transience of MCs

Write for $n \geq 1$

$$f_{ij}^{(n)} = \mathbb{P}(X_n = j, X_{n-1} \neq j, \ldots, X_1 \neq j \,|\, X_0 = i)$$

so that $f_{ij}^{(n)}$ is the probability starting in state i that we visit state j for the first time at time n. Also, let

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

the probability that we ever visit state j, starting in state i.

Definition

- If $f_{ii} < 1$ then state i is transient
- If $f_{ii} = 1$ then state i is recurrent.

Recurrence and transience, ctd

- Observe that if we return to a state i at some time n then the evolution of the MC is independent of the path before time n. Hence, the probability that we will return at least N times is f_{ii}^N.
- Now, if i is recurrent $f_{ii}^N = 1$ for all N and we are sure to return to state i infinitely often.
- Conversely, if state i is transient then $f_{ii}^N \to 0$ as $N \to \infty$ and so there is zero probability of returning infinitely often.

Solidarity property

It may be shown that all states within a communicating class share the property of either all being transient or all being recurrent. It is usual to refer to a communicating class as being either a transient or recurrent class.

Recurrent classes are closed

Theorem

Suppose that C is a recurrent class then C is closed.

Proof.

Let C be a class that is not closed. Then there exists some states $i \in C$ and $j \notin C$ such that $p_{ij} > 0$. Thus $j \not\rightsquigarrow i$ as $j \notin C$ and so

$$\mathbb{P}(X_n \neq i \text{ for all } n \geq 1 \,|\, X_0 = i) \geq \mathbb{P}(X_1 = j \,|\, X_0 = i) = p_{ij} > 0$$

which contradicts that i is recurrent. □

Mean recurrence time

First, let

$$T_i = \min\{n \geq 1 : X_n = i\}$$

be the time of the first visit to state i and set $T_i = \infty$ if no such visit ever occurs.

Thus, $\mathbb{P}(T_i = \infty | X_0 = i) > 0$ if and only if i is transient in which case $\mathbb{E}(T_i | X_0 = i) = \infty$.

Definition (Mean recurrence time)

The mean recurrent time, μ_i, of a state i is defined as

$$\mu_i = \mathbb{E}(T_i | X_0 = i) = \begin{cases} \sum_n n f_{ii}^{(n)} & \text{if } i \text{ is recurrent} \\ \infty & \text{if } i \text{ is transient.} \end{cases}$$

- Note that μ_i may still be infinite when i is recurrent.

Positive and null recurrence

Definition

A recurrent state i is

- positive recurrent if $\mu_i < \infty$ and
- null recurrent if $\mu_i = \infty$.

Finite state Markov Chains

The state space S is a countable set. When S is also finite there are several useful simplifications that occur.

Theorem

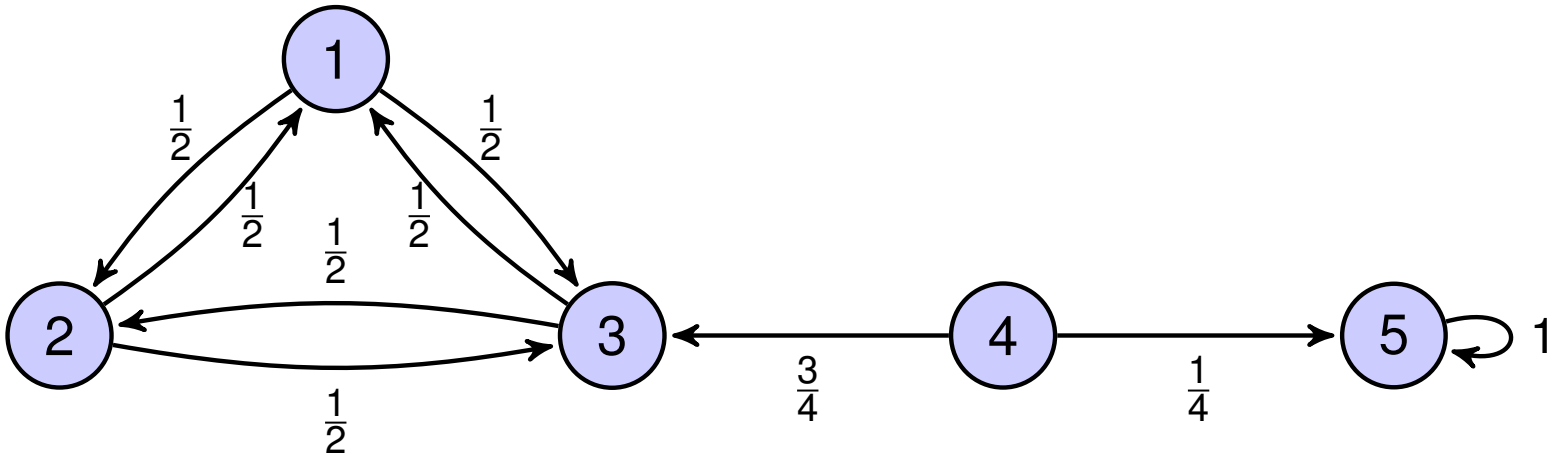
If S is finite then

1. *there is always at least one recurrent state*
2. *all recurrent states are positive recurrent.*

Proof.

Omitted. □

Example (5 states), ctd



Recall that

- State space is reducible into three classes
- Classes: $\{1,2,3\}$, $\{4\}$ and $\{5\}$
- $\{1,2,3\}$ and $\{5\}$ are closed; $\{5\}$ is absorbing.

Furthermore

- Classes $\{1,2,3\}$ and $\{5\}$ are positive recurrent
- Class $\{4\}$ is transient.

Later we will consider an example with a null recurrent class but where the state space is necessarily countably infinite.

Stationary distributions

Definition

The vector $\pi = (\pi_j; j \in S)$ is a stationary distribution for the MC with transition matrix P if

1. $\pi_j \geq 0$ for all $j \in S$ and $\sum_{j \in S} \pi_j = 1$
2. $\pi = \pi P$, or equivalently, $\pi_j = \sum_{i \in S} \pi_i p_{ij}$.

Such a distribution is stationary in the sense that $\pi P^2 = (\pi P)P = \pi P = \pi$ and for all $n \geq 0$

$$\pi P^n = \pi .$$

Thus if X_0 has distribution π then X_n has distribution π for all $n \geq 0$.

Existence of a stationary distribution

Theorem

1. *An irreducible Markov Chain has a stationary distribution π if and only if all states are positive recurrent.*
2. *If this is the case then π is the unique stationary distribution and it is given by*

$$\pi_i = \frac{1}{\mu_i}$$

where μ_i is the mean recurrence time of state $i \in S$.

Intuition

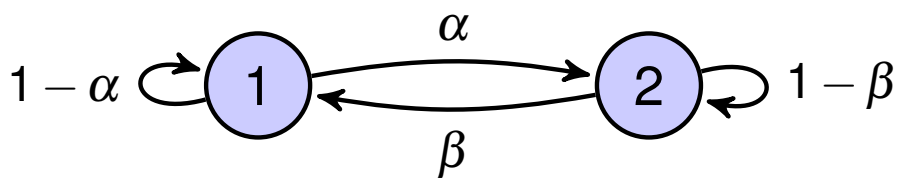
Suppose the current distribution for a MC is given by a stationary distribution π and consider the evolution of the MC for a further period of T steps (with T large). Since π is stationary the probability of being in any state i remains π_i, so we will make around $T\pi_i$ visits to i. Consequently, the mean recurrence time of state i would be $T/(T\pi_i) = 1/\pi_i$.

Stationary distributions for 2-state MCs

Consider the MC with transition matrix

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

with $0 \leq \alpha, \beta \leq 1$.



Now if $\alpha = \beta = 0$ then the state state is reducible and $P = I_2$, the 2×2 identity matrix, and so any distribution on 2 states, π, is stationary with $\pi = \pi P = \pi I_2 = \pi$. Thus, we have non-uniqueness of stationary distributions if the state space is reducible.

Stationary distributions for 2-state MCs, ctd

Now suppose $\alpha+\beta>0$ and write $\pi=(\pi_1,\pi_2)$ with $\pi_1,\pi_2\geq 0$ and $\pi_1+\pi_2=1$. Hence $\pi_2=1-\pi_1$ and the stationary equations become

$$\begin{pmatrix}\pi_1 & 1-\pi_1\end{pmatrix}=\begin{pmatrix}\pi_1 & 1-\pi_1\end{pmatrix}\begin{pmatrix}1-\alpha & \alpha\\ \beta & 1-\beta\end{pmatrix}$$

Thus

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

and $\pi=(\frac{\beta}{\alpha+\beta},\frac{\alpha}{\alpha+\beta})$ is the unique stationary distribution.

Note that the case $\alpha+\beta>0$ includes the possibilities that either $\alpha=0$ or $\beta=0$ (but not both) in which case the state space is again reducible.

Periodicity

Let d_i be the greatest common divisor of $\{n : p_{ii}^{(n)} > 0\}$.

Definition

- If $d_i = 1$ then i is aperiodic.
- If $d_i > 1$ then i is periodic with period d_i.

Remark

It may be shown that the period is a class property, that is, if C is a communicatiing class and $i, j \in C$ then $d_i = d_j$.

Limiting behaviour as $n \to \infty$

Theorem

For an irreducible, aperiodic Markov Chain

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_j}$$

for all states $i, j \in S$.

Proof.

Omitted. □

Markov's example

Markov was led to the notion of a Markov chain by studying the patterns of vowels and consonants in text. In his original example he estimated a transition matrix for the states $\{\text{vowel}, \text{consonant}\}$ as

$$P = \begin{pmatrix} 0.128 & 0.872 \\ 0.663 & 0.337 \end{pmatrix}.$$

Taking successive powers of P we find

$$P^2 = \begin{pmatrix} 0.595 & 0.405 \\ 0.308 & 0.692 \end{pmatrix} \qquad P^3 = \begin{pmatrix} 0.345 & 0.655 \\ 0.498 & 0.502 \end{pmatrix} \qquad P^4 = \begin{pmatrix} 0.478 & 0.522 \\ 0.397 & 0.603 \end{pmatrix}.$$

As $n \to \infty$, all rows become identical in the limit

$$P^n \to \begin{pmatrix} 0.432 & 0.568 \\ 0.432 & 0.568 \end{pmatrix}.$$

Check that $\pi = (0.432, 0.568)$ is a stationary distribution (that is, $\pi = \pi P$) using the earlier 2-state example with $\alpha = 0.872$ and $\beta = 0.663$.

Furthermore, the mean recurrence times for the two states are $1/0.432 \approx 2.315$ and $1/0.568 \approx 1.761$.

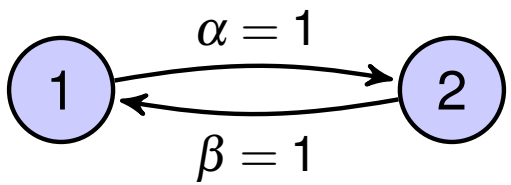
Remark on aperiodicity

What happens if we do not have aperiodicity?

Consider the 2-state example with $\alpha = \beta = 1$ then the MC just deterministically alternates between the two states and the period is 2. We can see that

$$p_{11}^{(n)} = p_{22}^{(n)} = \begin{cases} 1 & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases}$$

and so the limits $\lim_{n\to\infty} p_{ij}^{(n)}$ do not exist.



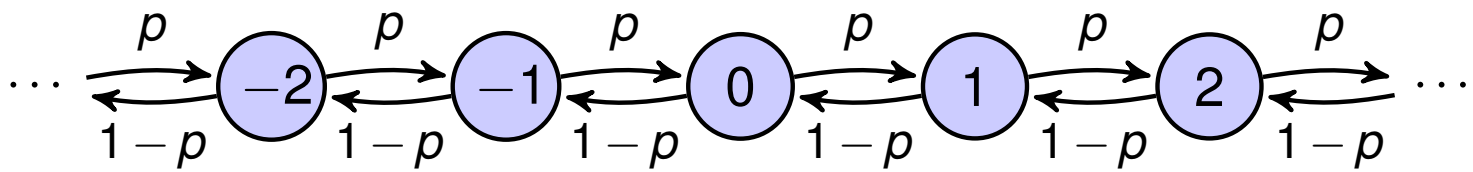
Random walks revisited

Recall the simple random walk starting at the origin (that is, $X_0 = 0$) given by

$$X_n = Y_1 + Y_2 + \cdots + Y_n$$

where $Y_1, Y_2, \ldots$ are IID RVs with $\mathbb{P}(Y_i = 1) = p$ and $\mathbb{P}(Y_i = -1) = 1 - p$ for $p \in [0, 1]$. Note here that the state space is countably infinite, namely the set of integers $\mathbb{Z}$.

Intutitively, if $p > \frac{1}{2}$ the MC will drift to the right and so eventually will never return to the origin and every state is transient. Likewise, if $p < \frac{1}{2}$ the drift will be to the left and again every state is transient.



Random walks revisited, ctd

Again intuitively, if $p = \frac{1}{2}$ then the origin is a recurrent state. Is it positive recurrent or null recurrent?

We know that if it is positive recurrent then there would be a unique stationary distribution, π, since the MC is also clearly irreducible in this case.

However such a distribution would statisfy $\pi = \pi P$ which gives for all i that

$$\pi_i = \frac{1}{2}\pi_{i-1} + \frac{1}{2}\pi_{i+1}$$

and so π_i is constant which contradicts that $\pi = (\pi_i : i \in \mathbb{Z})$ is a distribution with $\sum_{i\in\mathbb{Z}} \pi_i = 1$.

Thus, the simple symmetric random walk is null recurrent.

The problem sheet explores a related example where the random walk has states $\{0, 1, 2, \ldots\}$ but $p_{00} = p$ so there is a self-loop at the origin. In this example for a certain range of p values the random walk is positive recurrent and has a unique stationary distribution.

Time-reversibility

Suppose now that $(X_n : -\infty < n < \infty)$ is an irreducible, positive recurrent MC with transition matrix P and unique stationary distribution π. Suppose also that X_n has the distribution π for all $-\infty < n < \infty$. Now define the reversed chain by

$$Y_n = X_{-n} \qquad \text{for } -\infty < n < \infty$$

Then (Y_n) is also a MC and Y_n has the distribution π.

Definition (Reversibility)

A MC (X_n) is reversible if the transition matrices of (X_n) and (Y_n) are equal.

Theorem

A MC (X_n) is reversible if and only if

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \text{for all } i, j \in S.$$

Proof.

Consider the transition probabilities q_{ij}, say, of the MC (Y_n) then

$$\begin{aligned} q_{ij} &= \mathbb{P}(Y_{n+1} = j \,|\, Y_n = i) \\ &= \mathbb{P}(X_{-n-1} = j | X_{-n} = i) \\ &= \mathbb{P}(X_m = i | X_{m-1} = j)\mathbb{P}(X_{m-1} = j)/\mathbb{P}(X_m = i) \qquad \text{where } m = -n \\ &= p_{ji}\pi_j/\pi_i \,. \end{aligned}$$

Hence, $p_{ij} = q_{ij}$ if and only if $\pi_i p_{ij} = \pi_j p_{ji}$. □

Theorem

For an irreducible chain, if there exists a vector π such that

1. $0 \leq \pi_i \leq 1$ *and* $\sum_i \pi = 1$
2. $\pi_i p_{ij} = \pi_j p_{ji}$ *for all* $i, j \in S$

then the MC is reversible with stationary distribution π.

Proof.

Suppose that π satisfies the conditions of the theorem then

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$$

and so $\pi = \pi P$ and the distribution is stationary. □

The conditions $\pi_i p_{ij} = \pi_j p_{ji}$ for all $i, j \in S$ are known as the local balance (or detailed balance) conditions.

Ehrenfest model

Suppose we have two containers A and B containing a total of m balls. At each time step a ball is chosen uniformly at random and switched between containers. Let X_n be the number of balls in container A after n units of time. Thus, (X_n) is a MC with transition matrix given by

$$p_{i,i+1} = 1 - \frac{i}{m}, \qquad p_{i,i-1} = \frac{i}{m}.$$

Instead of solving the equations $\pi = \pi P$ we look for solutions to

$$\pi_i p_{ij} = \pi_j p_{ji}$$

which yields $\pi_i = \binom{m}{i}(\frac{1}{2})^m$, a binomial distribution with parameters m and $\frac{1}{2}$.

Random walk on an undirected graph

Consider a graph G consisting of a countable collection of vertices $i \in N$ and a finite collection of edges $(i,j) \in E$ joining (unordered) pairs of vertices. Assume also that G is connected. A natural way to construct a MC on G uses a random walk through the vertices. Let v_i be the number of edges incident at vertex i. The random walk then moves from vertex i by selecting one of the v_i edges with equal probability $1/v_i$. So the transition matrix, P, is

$$p_{ij} = \begin{cases} \frac{1}{v_i} & \text{if } (i,j) \text{ is an edge} \\ 0 & \text{otherwise}. \end{cases}$$

Since G is connected, P is irreducible. The local balance conditions for $(i,j) \in E$ are

$$\begin{aligned} \pi_i p_{ij} &= \pi_j p_{ji} \\ \pi_i \frac{1}{v_i} &= \pi_j \frac{1}{v_j} \\ \frac{\pi_i}{\pi_j} &= \frac{v_i}{v_j}. \end{aligned}$$

Hence,

$$\pi_i \propto v_i$$

and the normalization condition $\sum_{i \in N} \pi_i = 1$ gives

$$\pi_i = \frac{v_i}{\sum_{j \in N} v_j}$$

and P is reversible.

Ergodic results

Ergodic results refer to the limiting behaviour of time averages. In the case of Markov Chains we shall consider the long-run proportion of time spent in a given state.

Let $N_j^k(n)$ be the number of visits to k starting from state j before time n then consider the time average conditional on $X_0 = j$

$$\frac{1}{n}\mathbb{E}(N_j^k(n)) = \frac{1}{n}\mathbb{E}\left(\sum_{r=1}^{n}\mathbb{I}(X_r = k)\right) = \frac{1}{n}\sum_{r=1}^{n}\mathbb{E}(\mathbb{I}(X_r = k)) = \frac{1}{n}\sum_{r=1}^{n} p_{jk}^{(r)}$$

Now if the MC is irreducible and aperiodic we know that $\lim_{r\to\infty} p_{jk}^{(r)} = \frac{1}{\mu_k}$ for all states $j, k \in S$ and hence

$$\frac{1}{n}\mathbb{E}(N_j^k) = \frac{1}{n}\sum_{r=1}^{n} p_{jk}^{(r)} \to \frac{1}{\mu_k}\,.$$

Thus time averages also converge in the same way as limiting n-step transition probabilities.

Example: random surfing on web graphs

Consider a web graph, $G = (V, E)$, with vertices given by a finite collection of web pages $i \in V$ and (directed) edges given by $(i, j) \in E$ whenever there is a hyperlink from page i to page j.

Random walks through the web graph have received much attention in the recent years.

Consider the following model, let $X_n \in V$ be the location (that is, web page visited) by the surfer at time n and suppose we choose X_{n+1} uniformly from the, $L(i)$, outgoing links from i, in the case where $L(i) > 0$ and uniformly among all pages in V if $L(i) = 0$ (the dangling page case).

Pagerank: transition matrix

Hence, the transition matrix, $\hat{P}_{ij}$, say, has non-zero entries given by

$$\hat{p}_{ij} = \begin{cases} \frac{1}{L(i)} & \text{if } L(i) > 0 \text{ and } (i,j) \in E \\ \frac{1}{|V|} & \text{if } L(i) = 0 \end{cases}$$

Note that the resulting MC may not be irreducible and may be periodic.

We need to find a variant of the MC that is irreducible, aperiodic and positive recurrent so that we can proceed to determine a unique stationary distribution and exploit the limiting probabilities.

Recall that irreducible finite state MCs are always positive recurrent.

For our web graph model V, the set of web pages, is finite (though very large) so it will be sufficient to ensure that the MC is just irreducible and aperiodic.

Easily bored web surfer model

We will make a further adjustment to ensure irreducibility and aperiodicity as follows.

For $0 < \alpha \leq 1$ set

$$p_{ij} = (1 - \alpha)\hat{p}_{ij} + \alpha \frac{1}{|V|}.$$

We can interpret this as an "easily bored web surfer" model and see that the transitions take the form of a mixture of two distributions.

With probability $1 - \alpha$ we follow the randomly chosen outgoing link (unless the page is dangling in which case we move to a randomly chosen page) while with probability α we jump to a random page selected uniformly from the entire set of pages V.

PageRank

Brin *et al* (1999) used this approach to define PageRank through the limiting distribution of this Markov Chain, that is π_i where the vector π satisfies

$$\pi = \pi P$$

They report typical values for α of between 0.1 and 0.2.

The ergodic results now tells us that the random surfer in this model spends a proportion π_i of the time visiting page i — a notion in some sense of the *'importance'* of page i.

Thus, two pages i and j can be ranked according to the total order defined by

$$i \geq j \quad \text{if and only if} \quad \pi_i \geq \pi_j \,.$$

Sergey Brin, Lawrence Page, Rajeev Motwani and Terry Winograd
The PageRank Citation Ranking: Bring Order to the Web
Technical Report, Computer Science Department, Stanford University. (1999).
http://dbpubs.stanford.edu:8090/pub/1999-66

Computing PageRank: the power method

We seek a solution to the system of equations

$$\pi = \pi P$$

that is, we are looking for an eigenvector of P (with corresponding eigenvalue of one). Google's computation of PageRank is one of the world's largest matrix computations.

The power method starts from some initial distribution $\pi^{(0)}$, updating $\pi^{(k-1)}$ by the iteration

$$\pi^{(k)} = \pi^{(k-1)} P = \cdots = \pi^{(0)} P^k$$

Advanced methods from linear algebra can be used to speed up convergence of the power method and there has been much study of related MCs to include web browser back buttons and many other properties as well as alternative notions of the *'importance'* of a web page.

The power method lends itself well to large-scale parallel computation using the MapReduce approach.

Hidden Markov Models

An extension of Markov Chains is provided by Hidden Markov Models (HMM) where a statistical model of observed data is constructed from an underlying but usually hidden Markov Chain.

Such models have proved very popular in a wide variety of fields including

- speech and optical character recognition
- natural language processing
- bioinformatics and genomics.

We shall not consider these applications in any detail but simply introduce the basic ideas and questions that Hidden Markov Models address.

A Markov model with hidden states

Suppose we have a MC with transition matrix P but that the states i of the chain are not directly observable. Instead, we suppose that on visiting any state i at time n there is a randomly chosen output value or token, Y_n, that is observable.

The probability of observing the output token t when in state i is given by some distribution b_i, depending on the state i that is visited.

Thus,

$$\mathbb{P}(Y_n = t | X_n = i) = (b_i)_t$$

where $(b_i)_t$ is the t^{th} component of the distribution b_i.

For an excellent introduction to HMM, see:

Lawence R. Rabiner
A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition
Proceedings of the IEEE, Vol 77, No 2, February (1988).

Three central questions

There are many variants of this basic setup but three central problems are usually addressed.

Definition (Evaluation problem)

Given a sequence $y_1, y_2, \ldots, y_n$ of observed output tokens and the parameters of the HMM (namely, P, b_i and the distribution for the initial state X_0) how do we compute

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \,|\, \text{HMM parameters})$$

that is, the probability of the observed sequence given the model?

Such problems are solved in practice by the forward algorithm.

A second problem that may occur in an application is the decoding problem.

Definition (Decoding problem)

Given an observed sequence of output tokens $y_1, y_2, \ldots, y_n$ and the full description of the HMM parameters, how do we find the best fitting corresponding sequence of (hidden) states $i_1, i_2, \ldots, i_n$ of the MC?

Such problems are solved in practice by a dynamic programming approach called the Viterbi algorithm.

The third important problem is the learning problem.

Definition (Learning problem)

Given an observed sequence of output tokens $y_1, y_2, \ldots, y_n$, how do we adjust the parameters of the HMM to maximize

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \,|\, \text{HMM parameters})$$

The observed sequence used to adjust the model parameters is called a training sequence. Learning problems are crucial in most applications since they allow us to create the "best" models in real observed processes.

Iterative procedures, known as the Baum-Welch method, are used to solve this problem in practice.

Applications of Markov Chains

These and other applications of Markov Chains are important topics in a variety of Part II courses, including

- Artificial Intelligence II
- Bioinformatics
- Computer Systems Modelling

Properties of discrete RVs

RV, X	Parameters	$\mathrm{Im}(X)$	$\mathbb{P}(X=k)$	$\mathbb{E}(X)$	$\mathrm{Var}(X)$	$G_X(z)$
Bernoulli	$p \in [0,1]$	$\{0,1\}$	$(1-p)$ if $k=0$ or p if $k=1$	p	$p(1-p)$	$(1-p+pz)$
Bin(n,p)	$n \in \{1,2,\ldots\}$ $p \in [0,1]$	$\{0,1,\ldots,n\}$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	$(1-p+pz)^n$
Geo(p)	$0 < p \leq 1$	$\{1,2,\ldots\}$	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pz}{1-(1-p)z}$
$U(1,n)$	$n \in \{1,2,\ldots\}$	$\{1,2,\ldots,n\}$	$\frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{z(1-z^n)}{n(1-z)}$
Pois(λ)	$\lambda > 0$	$\{0,1,\ldots\}$	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	$e^{\lambda(z-1)}$

Properties of continuous RVs

RV, X	Parameters	$\mathrm{Im}(X)$	$f_X(x)$	$\mathbb{E}(X)$	$\mathrm{Var}(X)$
$U(a,b)$	$a,b \in \mathbb{R}$ $a < b$	(a,b)	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathrm{Exp}(\lambda)$	$\lambda > 0$	$\mathbb{R}_+$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$N(\mu,\sigma^2)$	$\mu \in \mathbb{R}$ $\sigma^2 > 0$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2