Modeling and Statistical Inference

Simon R. Arridge

April 29, 2004

† *Department of Computer Science, University College London, Gower Street London, WC1E 6BT*

Theory III

October-December 2003

1 Random Variables and Probability Density Functions

A *Random variable* (r.v.) is a variable which can take on certain values, each of which has an associated *probability*

The *range* of the random variable is the set of values that it can take. This may be continuous (C) or discrete (D), infinite (I) or finite(F).

Example	Type	Range($\boldsymbol{X}$)
Face value on die	D,F	$\{1, 2, 3, 4, 5, 6\}$
No. people entering a building since opening	D,I	$\{0, 1, \ldots, \infty\}$
No. mms rainfall	C,I	$\mathbb{R}^+$
Time spent by a student in an exam	C,F	$[0, 2.5\text{hours}]$

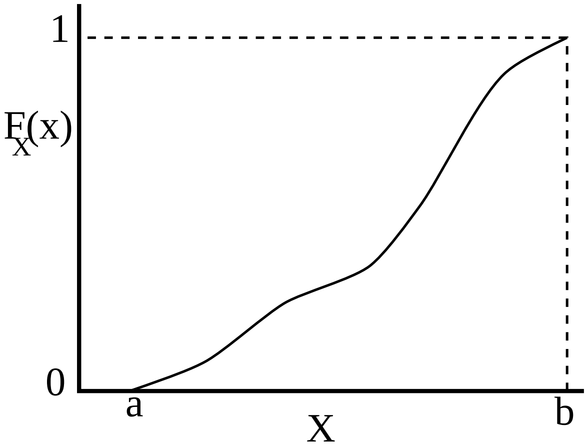
1.1 Distribution Functions

Suppose $\boldsymbol{X} \in (a, b)$ is a r.v. The probability that $\boldsymbol{X}$ takes on a value less than a specified value is called the *distribution function* for $\boldsymbol{X}$:

$$F_{\boldsymbol{X}}(x) \equiv Pr(\boldsymbol{X} \leq x)$$

The following should be obvious :

$$\begin{aligned} F_{\boldsymbol{X}}(a) &= 0 \\ F_{\boldsymbol{X}}(b) &= 1 \\ F_{\boldsymbol{X}}(x+h) &\geq F_{\boldsymbol{X}}(x) \quad \forall\, h \geq 0 \end{aligned}$$



$F_{\boldsymbol{X}}$ is therefore a monotonically increasing function

Suppose $\boldsymbol{X}$ is discrete. Then $F_{\boldsymbol{X}}(x)$ is a discontinuous function.

In general if $\boldsymbol{X} \in \{a, a+1, \ldots, b\}$, what is the probability it takes a specific value $\boldsymbol{X} = x$?

By combination of probabilities,

$$Pr(x||y) = Pr(x) + Pr(y) - Pr(x\&\&y)$$

we have

$$\begin{aligned} Pr(\boldsymbol{X} \leq x) &= Pr(X \leq x-1||\boldsymbol{X} = x) \\ &= Pr(\boldsymbol{X} \leq x-1) + Pr(\boldsymbol{X} = x) \\ => \quad Pr(\boldsymbol{X} = x) &= F_{\boldsymbol{X}}(x) - F_{\boldsymbol{X}}(x-1) \end{aligned}$$

1.2 Probability Density Functions

The function

$$f_{\boldsymbol{X}} = F_{\boldsymbol{X}}(x) - F_{\boldsymbol{X}}(x-1) \equiv Pr(\boldsymbol{X} = x)$$

is the *Probability Density Function* (PDF) for $\boldsymbol{X}$. Suppose $\boldsymbol{X}$ is continuous. The quantity $F_{\boldsymbol{X}}(x) - F_{\boldsymbol{X}}(x-1)$ represents the probability that $\boldsymbol{X}$ takes a value in the range $x - 1 < \boldsymbol{X} \leq x$. In the limit

$$f_X(x) = \lim_{h \to 0} \frac{F_{\boldsymbol{X}}(x+h) - F_{\boldsymbol{X}}(x)}{h} = F'_{\boldsymbol{X}}(x)$$

This can be interpreted as saying

$$f_X(x)\delta x \equiv Pr(\boldsymbol{X} \in (x, x + \delta x])$$

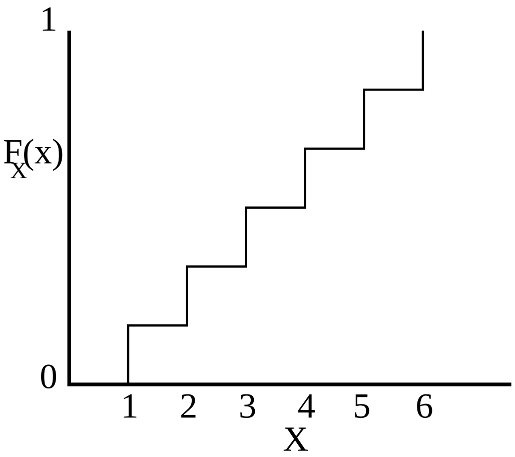


Figure 1: Pr(face value of a die $\leq x$)

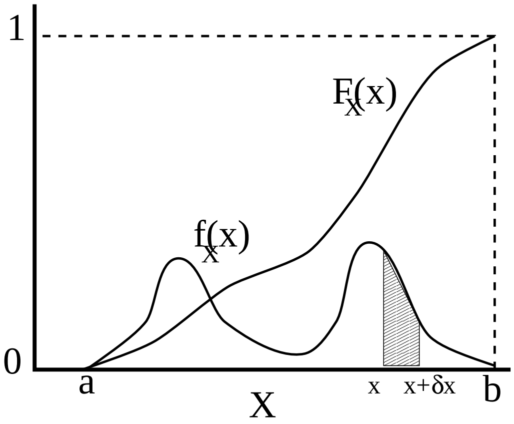
The following properties of the PDF follow where Ω is the range of $\boldsymbol{X}$

$$\begin{aligned} f_{\boldsymbol{X}}(x) &\geq 0 \quad \forall\, x \in \Omega \\ \sum_{x \in \Omega} f_{\boldsymbol{X}}(x) &= 1 \quad \text{discrete case} \\ \int_{\Omega} f_{\boldsymbol{X}}(x)\mathrm{d}x &= 1 \quad \text{continuous case} \end{aligned}$$

Let $A \subset \Omega$ be a subset of the range of $\boldsymbol{X}$ then

$$\begin{aligned} Pr(\boldsymbol{X} \in A) &= \sum_{x \in A} f_{\boldsymbol{X}}(x) \quad \text{discrete case} \\ &= \int_{A} f_{\boldsymbol{X}}(x)\mathrm{d}x \quad \text{continuous case} \end{aligned}$$

The function $F_{\boldsymbol{X}}$ is also called the *Cumulative Probability Density Function.*



2 Example Distributions

2.1 Binomial (Bernouilli)

Suppose we have a system in just two states A, B, with a random process giving the result with probability a, b respectively. Clearly

$$b = 1 - a$$

An example is the toss of a coin with

$$Pr(\text{Heads}) = Pr(\text{Tails}) = \frac{1}{2}$$

suppose we repeat the experiment n times and ask for the probability of getting x heads (a so called *Bernouilli Trial*).

Represent a Head by 1, Tail by 0, then an example sequence is

$$1001001110\ldots10$$

Suppose there are x '1's.

We can slot the first 1 into 1 of n slots,

We can slot the second 1 into 1 of $n-1$ slots,

...

We have a total of

$$n(n-1)(n-2)\ldots(n-x+1) = \frac{n!}{(n-x)!}$$

choices. But we can also rearrange the '1's in $x!$ ways that are indistinguishable. This implies there are a total of

$$\binom{n}{x} = \frac{n!}{(n-x)!x!}$$

ways of getting x heads (i.e. there are $\binom{n}{x} = \frac{n!}{(n-x)!x!}$ ways of reordering x heads and $(n-x)$ tails)

The required PDF is the Binomial Distribution

$$f_{\boldsymbol{X}} = \binom{n}{x} a^x (1-a)^{n-x}$$

By inspection we find that each value of $f_{\boldsymbol{X}}(x)$ is a coefficient of expansion of

$$(a + (1-a))^n = a^n + na^{n-1}(1-a) + \ldots + \binom{n}{k} a^k (1-a)^{n-k} + \ldots + (1-a)^n$$

Clearly the left hand side is 1 which satisfies the requirement for $\sum_{x\in\Omega} f_{\boldsymbol{X}}(x) = 1$

2.2 The Poisson Distribution

The Poisson model assumes that the average number of events per unit time is constant, but that events may occur at any time

Examples

Patients arrive at a surgery

An atom decays radioactively

Buses arrive at a bus stop

Process resource requests arrive at CPU

Postulates

1. Events occur singly along a continuous axis (time)
2. Events occur *uniformly* in that the average is constant
3. Events are independent

Arrange for 2. and 3. to be true by choosing short enough interval δt.

P_2 :-	Pr(2+ occurences in δt)	$= \epsilon_2$
P_1 :-	Pr(1 occurence in δt)	$= \lambda\delta t - \epsilon_1$
P_0 :-	Pr(0 occurences in δt)	$= 1 - \lambda\delta t + (\epsilon_1 - \epsilon_2)$

Note:
ϵ_1 and ϵ_2 are assumed small probabilities $O(\delta t^2)$.
Since ϵ_2 is not zero, neither can be ϵ_1. This is because the *average* number of occurences would be slightly greater than $\lambda\delta t$.

Put

$$\begin{aligned}\epsilon_1 &= K_1\delta t^2\\ \epsilon_2 &= K_2\delta t^2\end{aligned}$$

where K_1, K_2 are constants. Then

$$\epsilon_1 - \epsilon_2 = (K_1 - K_2)\delta t^2 = K_0\delta t^2$$

Definition

$$Pr(i,t) := Pr(\text{event occurs } i \text{ times in time } t)$$

Consider the change in probability with change in time $t \to t + \delta t$

$$\begin{aligned}
Pr(i, t+\delta t) &= Pr(i \text{ occurences in } t \,\&\&\, 0 \,\text{in}\, \delta t) \\
&+ Pr(i-1 \text{ occurences in } t \,\&\&\, 1 \,\text{in}\, \delta t) \\
&+ Pr(i-2 \text{ occurences in } t \,\&\&\, 2 \,\text{in}\, \delta t) \\
&+ \ldots \\
&= \sum_{n=0}^{n=i} Pr(i-n, t) Pr(n, \delta t)
\end{aligned}$$

Using definitions of $P_0 \rightarrow P_2$:

$$\begin{aligned}
Pr(i, t+\delta t) &= Pr(i,t)(1 - \lambda\delta t + K_0 \delta t^2) \\
&+ Pr(i-1,t)(\lambda\delta t - K_1 \delta t^2) \\
&+ Pr(i-2,t) K_2 \delta t^2 \\
&+ O(\delta t)
\end{aligned}$$

Rearranging

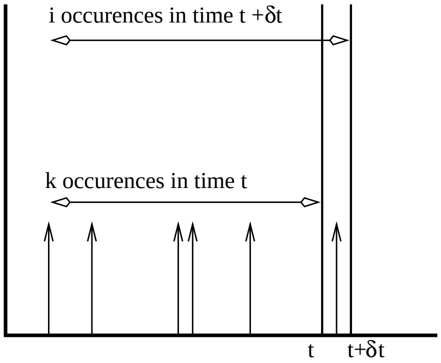
$$\begin{aligned}
\frac{Pr(i,t+\delta t) - Pr(i,t)}{\delta t} &= \lambda \left(Pr(i-1,t) - Pr(i,t)\right) \\
&+ O(\delta t^2)
\end{aligned}$$

Taking the limit as $\delta t \rightarrow 0$

$$\begin{aligned}
\frac{\partial Pr(i,t)}{\partial t} &= -\lambda \left(Pr(i,t) - Pr(i-1,t)\right) \quad & i \geq 1 \\
\frac{\partial Pr(0,t)}{\partial t} &= -\lambda Pr(0,t) \quad & i = 0
\end{aligned}$$

For $i = 0$, and using $Pr(0,0) = 1$, we have

$$Pr(0,t) = \mathrm{e}^{-\lambda t}$$



For $i = 1$, and using $Pr(1,0) = 0$:

$$\begin{aligned} \frac{\mathrm{d}e^{\lambda t} Pr(1,t)}{\mathrm{d}t} &= \lambda \\ \Rightarrow \quad Pr(1,t) &= \lambda t \mathrm{e}^{-\lambda t} \end{aligned}$$

By induction and using $Pr(i,t) = 0$, we get

$$Pr(i,t) = \frac{(\lambda t)^i}{i!} \mathrm{e}^{-\lambda t}$$

Poisson model as limit of Binomial

Consider the Binomial model in the case when a becomes very small and n becomes large, such that na is a constant.

In fact $\mu = na$ is the *mean* or *expected value* of both distributions

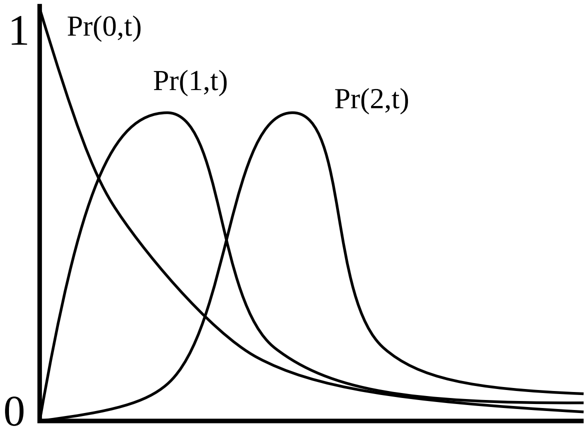
$$\begin{aligned} f_{\boldsymbol{X}}(x) &= \binom{n}{x} a^x (1-a)^{n-x} \\ &= \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x} \frac{n(n-1)\dots(n-x+1)}{x!} \\ &= \frac{\mu^x}{x!} \left(\frac{n}{n}\right) \left(\frac{n-1}{n}\right) \dots \left(\frac{n-x+1}{n}\right) \left(1 - \frac{\mu}{n}\right)^{n-x} \end{aligned}$$

In the limit $n \to \infty$

$$\lim_{n\to\infty} \left(\frac{n-k}{n}\right) \to 1\,, \quad \lim_{n\to\infty} \left(1 - \frac{\mu}{n}\right)^{n-x} \to \mathrm{e}^{-\mu}$$

which leads to

$$\lim_{n\to\infty} f_{\boldsymbol{X}}(x) \to \frac{\mu^x}{x!} \mathrm{e}^{-\mu}$$



2.3 Exercises

1. The arrival of patients at a doctor's surgery may be regarded as a Poisson process. Twenty patients arrive in an hour; what are the Poisson probabilities for the number of arrivals in a given 3-minute period.

2. A 500-page book contains 50 misprints. What is the probability that there will be more than two misprints on a particular page?

3. Calculate the Poisson probabilities for $\mu = \frac{1}{2}$, and calculate the first few binomial probabilities for $n = 50, a = 0.01$. Why is $0.99^{50} \simeq \mathrm{e}^{-\frac{1}{2}}$?

4. The road accidents in a certain area occur at the rate of one every two days. Calculate the probability of $0, 1, 2, \ldots 6$ accidents per week in this district. What is the most likely number of accidents per week ? How many days in a week are expected to be free of accidents ?

5. Find the probability function for the length of time we have to wait before the first occurence of a random event which happens on average once every six seconds, and compare it with the waiting time before a six appears on a die thrown once a second. Repeat for the time before the second occurence.

2.4 Answers

1. $\{0.37, 0.37, 0.18, 0.06 \ldots\}$

2. 0.0015

3. Poisson $\{0.6065, 0.3032, 0.0758, 0.0126, 0.0016\}$
   Binomial $\{0.6051, 0.3056, 0.0756, 0.0122, 0.0015\}$

4. $0.03, 0.106, 0.185, 0.216, 0.189, 0.132, 0.055$. Most likely : 3. Accident free-days : 4.2.

5. $\frac{1}{6}\mathrm{e}^{\frac{1}{6}t}$; $\frac{1}{6}(1-\frac{5}{6}t)^{-1}$; $\frac{1}{36}\mathrm{e}^{\frac{1}{36}t}$; $\frac{1}{36}(1-\frac{5}{6}t)^{-2}$;

3 Expectation Algebra

3.1 Expectation

The expected value of a r.v. is the integral (or sum in discrete case) over all possible values, weighted by probability :

$$E[\boldsymbol{X}] = \begin{cases} \int_\Omega x f_{\boldsymbol{X}}(x)\mathrm{d}x & \text{continuous} \\ \sum_{i\in\Omega} x_i f_i & \text{discrete} \end{cases}$$

More generally , for any function $g(\boldsymbol{X})$ we define

$$E[g(\boldsymbol{X})] = \begin{cases} \int_\Omega g(x) f_{\boldsymbol{X}}(x)\mathrm{d}x & \text{continuous} \\ \sum_{i\in\Omega} g_i f_i & \text{discrete} \end{cases}$$

where $g_i = g(x_i)$. The following properties follow

1. $E[K] = K$ for any constant K
2. $E[g(\boldsymbol{X}) + h(\boldsymbol{X})] = E[g(\boldsymbol{X})] + E[h(\boldsymbol{X})]$
3. $E[Kg(\boldsymbol{X})] = KE[g(\boldsymbol{X})]$

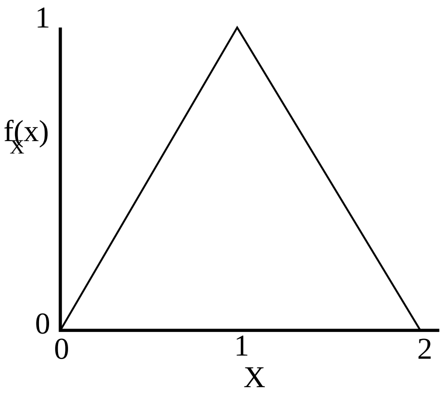
Example 1
A game is played with three dice. You pay £1 to play. You choose any number 1 → 6 and win

£4	if you get three of that number
£3	if you get two of that number
£2	if you get one of that number

No other results wins. What are your expected winnings?

$$\begin{aligned} Pr(\text{gain£}3) &= \frac{1}{216} \\ Pr(\text{gain£}2) &= \frac{15}{216} \\ Pr(\text{gain£}1) &= \frac{75}{216} \\ Pr(\text{lose£}1) &= \frac{125}{216} \\ E(\text{winnings}] &= 3\frac{1}{216} + 2\frac{15}{216} + 1\frac{75}{216} - 1\frac{125}{216} \\ &= \frac{-17}{216} \end{aligned}$$

The house wins about $\sim 8p$ on average **Example 2**
$\boldsymbol{X}$ is the sum of two numbers randomly chosen in $[0 \to 1]$. What are $E[\boldsymbol{X}]$ and $E[\boldsymbol{X}^2]$?



The PDF for $\boldsymbol{X}$ rises linearly from $0 \rightarrow 1$ and falls linearly from $1 \rightarrow 2$

$$\begin{aligned} E[\boldsymbol{X}] &= \int_0^1 x^2 \mathrm{d}x + \int_1^2 x(2-x)\mathrm{d}x \\ &= \frac{1}{3} + \left(3 - \frac{7}{3}\right) = 1 \\ E[\boldsymbol{X}^2] &= \int_0^1 x^3 \mathrm{d}x + \int_1^2 x^2(2-x)\mathrm{d}x \\ &= \frac{1}{4} + \left(\frac{14}{3} - \frac{15}{4}\right) = \frac{7}{6} \end{aligned}$$

3.2 Variance

The variance of a distribution is a measure of the dispersion away from the expectation. It is the integral sum of the squared differences from the expectation.

$$\begin{aligned} Var[X] &= \int_\Omega (x - E(\boldsymbol{X}))^2 f_{\boldsymbol{X}}(x)\mathrm{d}x \\ &= E[(\boldsymbol{X} - E(\boldsymbol{X}))^2] \end{aligned}$$

Variance is measured in squared units
Standard Deviation $\sigma = \sqrt{Variance}$ is measured in the same units as $\boldsymbol{X}$.
Let $\mu = E[\boldsymbol{X}]$, then

$$\begin{aligned} \sigma^2 = Var[\boldsymbol{X}] &= E[(\boldsymbol{X} - \mu)^2] \\ &= E[\boldsymbol{X}^2 - 2\mu\boldsymbol{X} + \mu^2] \\ &= E[\boldsymbol{X}^2] - 2\mu E[\boldsymbol{X}] + \mu^2 \\ &= E[\boldsymbol{X}^2] - 2\mu^2 + \mu^2 \\ &= E[\boldsymbol{X}^2] - \mu^2 \end{aligned}$$

3.3 Higher Moments

We can straightforwardly define the higher order moments around the expectation

$$\mu_r := E[(\boldsymbol{X} - \mu)^r]$$

Usually only μ_3 ("skewness"), and μ_4 ("kurtosis") are considered relevent.

It is possible to express moments around the expectation in terms of moments around the origin $\mu'_r := E[\boldsymbol{X}^r]$. For example

$$\begin{aligned} \mu_3 &= \mu'_3 - 3\mu'_2\mu'_1 + 2{\mu'_1}^3 \\ \mu_4 &= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_3{\mu'_1}^2 - 3{\mu'_1}^4 \end{aligned}$$

which may be computationally cheaper. E.g.

$$\begin{aligned} \mu_3 &= \mu'_3 - 3(\mu_2 + {\mu'_1}^2)\mu'_1 + 2{\mu'_1}^3 \\ \mu_4 &= \mu'_3 - 3\mu_2\mu'_1 - {\mu'_1}^3 \end{aligned}$$

Note It is important to realise that some distributions may not have some moments. This is because the integrals may not always converge. However, for most of the distributions that we will meet, all the moments are finite.

In principle, if all the moments of a distributions are known, then the distribution is known. This follows from *Taylor's Theorem*.

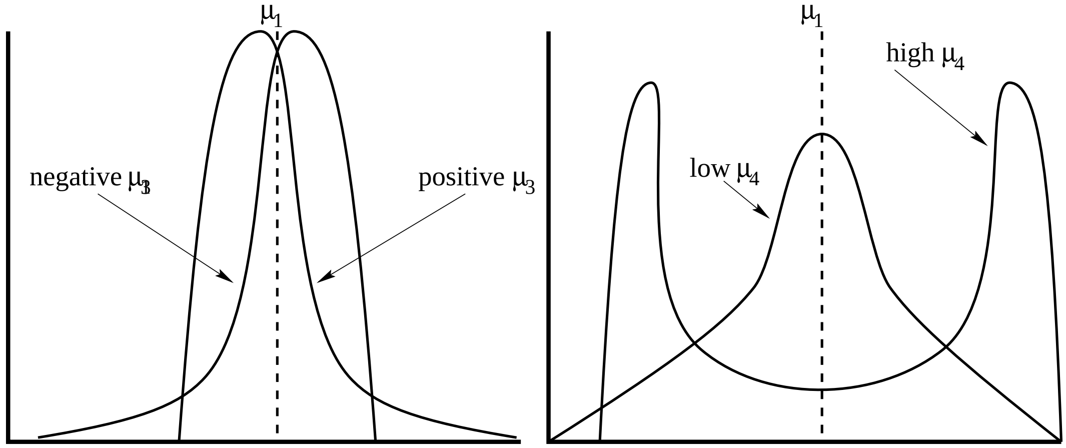


Figure 2: Left : Two distributions with the same expectation μ_1, and variance μ_2 but different skew μ_3. Right:Two distributions with the same expectation μ_1, variance μ_2, and skew μ_3 but different kurtosis μ_4.

3.4 Generating Functions

A probability generator is a function of an arbitrary parameter t such that the coefficient of t^i is the probability of event i

This works best for discrete probability distributions

$$G(t) = E[t^i] = \sum_{i \in \Omega} f_i t^i$$

Example
The Binomial distribution, for n trials with probability a, has G.F.

$$G_{\text{Binomial}}(t; a, n) = (1 - a + at)^n = (1 + a(t-1))^n$$

Consider what happens as $n \to \infty$, with $\mu = an$ a constant:

$$G_{\text{Binomial}}(t; a, n) \to \left(1 + \frac{\mu(t-1)}{n}\right)^n$$

as $n \to \infty$ we have $\left(1 + \frac{x}{n}\right)^n \to \mathrm{e}^x$,

$$\Rightarrow \quad G_\infty(t) := \lim_{n \to \infty} G_{\text{Binomial}}(t; a, n) = \mathrm{e}^{\mu(t-1)}$$

But this is just the Poisson distribution

$$G_\infty(t) = \mathrm{e}^{-\mu}\mathrm{e}^{\mu t} = \mathrm{e}^{-\mu}\left(1 + \mu t + \frac{(\mu t)^2}{2!} + \dots\right)$$

The probability generator has a useful property

$$\begin{aligned} G(t) &= E[t^i] \\ \Rightarrow \quad G'(t) &= E[it^{(i-1)}] \\ G''(t) &= E[i(i-1)t^{(i-2)}] \end{aligned}$$

since t is an arbitrary function, take it to be 1 :

$$G'(1) = \mu'_1, \quad G''(1) = \mu'_2 - \mu'_1$$

Thus we get

$$\begin{aligned} \mu = \mu'_1 &= G'(1) \\ \sigma^2 = \mu_2 &= G''(1) + G'(1) - G''(1)^2 \end{aligned}$$

3.5 Moment Generating Functions

It is difficult to define probability generators for continuous functions. Instead we introduce the *moment generating function* (MGF)

$$\begin{aligned} M_{\boldsymbol{X}}(t) &= E\left[e^{tx}\right] = E\left[1 + xt + \frac{(xt)^2}{2!} + \ldots\right] \\ &= E\left[\sum_i \frac{(xt)^i}{i!}\right] \\ &= \sum_i \frac{t^i}{i!}\mu_i' \end{aligned}$$

The MGF contains "all the moments"; in particular the i^{th} moment is the ceofficient of $\frac{t^i}{i!}$ in the expansion of the MGF in powers of t - i.e. the Taylor series.

$$M_{\boldsymbol{X}}(t) = M_X(0) + tM_X'(0) + \frac{t^2}{2!}M_X''(0) + \ldots + \frac{t^i}{i!}M_X^{(i)}(0) + \ldots$$

So we have formally

$$\mu_i' = M_{\boldsymbol{X}}^{(i)}(0)$$

Example 1
Find the MGF when 1 and -1 are equally likely

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

Differentiation of this i times shows that all even moments are 1, and all odd moments are 0

Example 2
Find the MGF for a uniform PDF in the interval $\boldsymbol{X} \in [-1, 1]$

$$\begin{aligned} M_{\boldsymbol{X}}(t) &= \int_{-1}^{1} \frac{1}{2}\mathrm{e}^{tx}\mathrm{d}x \\ &= \frac{1}{2t}\left(e^t - e^{-t}\right) \\ &= \frac{\sinh(t)}{t} \end{aligned}$$

This has all odd moments zero, and even moments $\mu_i' = \frac{1}{1+i}$
Additional Properties of MGFs

For any function $g(\boldsymbol{X})$ of the random variable

$$M_{g(\boldsymbol{X})} = E\left[\mathrm{e}^{tg(\boldsymbol{X})}\right]$$

and for any constant A, B

$$M_{a+b\boldsymbol{X}}(t) = \mathrm{e}^{ta}M_{\boldsymbol{X}}(tb)$$

3.6 Exercises

1. Find the Moment Generating Function (MGF) in the following cases
   a) When 0, 1 are equally likely
   b) When -1,0,1 are equally likely
   c) For the negative exponenetial function $P(x) = k\mathrm{e}^{-kx}$
   d) For the function $P(x) = 2x\, for\, 0 \leq x \leq 1$
   e) For the binomial probabilities with parameters n, a
   f) For the Poisson probabilities with parameter μ

2. Find $E[x]$, $E[y]$,$E[x+y]$,$E[xy]$, for the following ranges and PDFS :
   a) $\Omega =$ the unit square, $P(x, y) = 4(1 - x - y + xy)$
   b) $\Omega =$ the unit square, $P(x, y) = 6(x - y)^2$
   c) $\Omega =$ the quarter plane x, y positive, $P(x, y) = \mathrm{e}^{-(x+y)}$
   d) $\Omega =$ a quarter circle radius 3, x, y positive, $P(x, y)$ a quarter cone radius 3, with volume 1

3. If $y = mx + c$ and $M(t)$ is the MGF for x, what is the MGF for y ?

3.7 Answers

1. a) $\frac{1}{2}(1+\mathrm{e}^t)$
   b) $\frac{1}{3}(\mathrm{e}^{-t}+1+\mathrm{e}^t)$
   c) $k(1-\frac{t}{k})$
   d) $\frac{2}{t^2}(1-\mathrm{e}^t = t\mathrm{e}^t)$
   e) $[1+a(\mathrm{e}^t-1)]^n$
   f) $\mathrm{e}^{m(\mathrm{e}^t-1)}$

2. a)1/3,1/3,2/3,1/9
   b)1/2,1/2,1,1/6
   c)1,1,2,1
   d)$3/\pi$, $3/\pi$, $6/\pi$, $27/10\pi$

3. $\mathrm{e}^{ct}M(mt)$.

4 Some Distributions

4.1 Discrete Uniform Distribution

$\boldsymbol{X}$ is a r.v. with domain $\Omega = \{1, 2, \ldots n\}$ with all values equally likely.

$$\begin{aligned} f_{\boldsymbol{X}}(x) &= \frac{1}{n} \quad \forall\, x \in \Omega \\ \Rightarrow \quad E[\boldsymbol{X}] &= \sum_{x=1}^{n} \frac{x}{n} = \frac{n+1}{2} \\ E[\boldsymbol{X}^2] &= \sum_{x=1}^{n} \frac{x^2}{n} = \frac{(n+1)(2n+1)}{6} \\ \Rightarrow \quad Var[\boldsymbol{X}] &= \frac{(n+1)(n-1)}{12} \end{aligned}$$

4.2 Continuous Uniform Distribution

$\boldsymbol{X}$ ranges continuously in the (closed) domain $\Omega = [a, b]$. "Nothing is known" about $\boldsymbol{X}$ apart from this

$$f_{\boldsymbol{X}}(x) = \frac{1}{b-a} \quad x \in [a, b]$$

Find MGF from

$$M_{\boldsymbol{X}}(t) = \int_a^b \frac{\mathrm{e}^{tx}}{b-a} \mathrm{d}x = \frac{1}{b-a} \frac{\mathrm{e}^{tb} - \mathrm{e}^{ta}}{t}$$

Expanding $M_{\boldsymbol{X}}(t)$ we find

$$\begin{aligned} M_{\boldsymbol{X}}(t) &= \frac{1}{t(b-a)} \left[t(b-a) + \frac{t^2}{2}(b^2 - a^2) + \frac{t^6}{6}(b^3 - a^3) + \ldots \right] \\ &= 1 + t \frac{b^2 - a^2}{2(b-a)} + \frac{t^2}{2} \frac{b^3 - a^3}{3(b-a)} + \ldots \end{aligned}$$

This gives $\mu_1' = \frac{b+a}{2}$, $\mu_2' = \frac{b^2+ab+a^2}{3}$

4.3 Binomial Distribution

We have already seen that

$$f_{\boldsymbol{X}}(x) = \binom{n}{x} a^x (1-a)^{n-x} \quad x \in \{0, 1, 2, \ldots n\}$$

The (discrete) MGF is

$$\begin{aligned}
M_{\boldsymbol{X}}(t) &= \sum_{x=0}^{n} \mathrm{e}^{tx} \binom{n}{x} a^x (1-a)^{n-x} \\
&= \sum_{x=0}^{n} \mathrm{e}^{tx} \binom{n}{x} \left(a\mathrm{e}^t\right)^x (1-a)^{n-x} \\
&= \left(a\mathrm{e}^t + (1-a)\right)^n \\
M'_{\boldsymbol{X}}(t) &= n\left(a\mathrm{e}^t + (1-a)\right)^{n-1} a\mathrm{e}^t \\
M''_{\boldsymbol{X}}(t) &= n(n-1)\left(a\mathrm{e}^t + (1-a)\right)^{n-2} a^2\mathrm{e}^{2t} + \\
&\qquad n\left(a\mathrm{e}^t + (1-a)\right)^{n-1} a\mathrm{e}^t
\end{aligned}$$

This gives

$$\begin{aligned}
\mu &= M'_{\boldsymbol{X}}(0) = na \\
\sigma^2 &= M''_{\boldsymbol{X}}(0) - \mu^2 = na(1-a)
\end{aligned}$$

4.4 Beta Distribution

The Beta distribution arises when there is a prior probability that a distribution is *not* uniform, for example that it is closer to 1 than 0, or that its most probable value is known.

The Beta *function* is defined

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t \quad a > 0, b > 0$$

The Gamma function is defined

$$\Gamma(a,b) = \int_0^\infty t^{a-1}\mathrm{e}^{-t}\mathrm{d}t \quad a > 0$$

It is easy to show :

$$\Gamma(a+1) = a\Gamma(a) \quad \forall\, a; \quad \Gamma(n+1) = n! \quad \text{for integer}\, n$$

So the Gamma function is like a generalisation of factorial for non integers.

The Gamma and Beta functions are related like this

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

The Beta probability distribution for r.v. $\boldsymbol{X}$ with $\Omega = [0,1]$ is defined by

$$f_{\boldsymbol{X}}(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \quad x \in [0,1]$$

We get the mean and variance directly

$$\begin{aligned}
E[\boldsymbol{X}] &= \frac{1}{B(a,b)}\int_0^1 x^a(1-x)^{b-1}\mathrm{d}x = \frac{B(a+1,b)}{B(a,b)} \\
&= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{a}{a+b} \\
E[\boldsymbol{X}^2] &= \frac{1}{B(a,b)}\int_0^1 x^{a+1}(1-x)^{b-1}\mathrm{d}x = \frac{B(a+2,b)}{B(a,b)} \\
&= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{a(a+1)}{(a+b+1)(a+b)}
\end{aligned}$$

This leads to

$$Var[\boldsymbol{X}] = \frac{ab}{(a+b)^2(a+b+1)}$$

Note $Beta(0,1) \equiv Uniform(0,1)$

4.5 Poisson Distribution

We have seen that

$$f_{\boldsymbol{X}}(x) = \frac{\mu}{x!}\mathrm{e}^{-\mu} \quad x \in \{0,1,2,\ldots\}$$

The MGF is

$$\begin{aligned}
M_{\boldsymbol{X}}(t) &= \mathrm{e}^{-\mu}\sum_{x=0}^{\infty}\frac{(\mu\mathrm{e}^t)^x}{x!} = \mathrm{e}^{\mu(\mathrm{e}^t-1)} \\
M'_{\boldsymbol{X}}(t) &= \mu\mathrm{e}^t\mathrm{e}^{\mu(\mathrm{e}^t-1)} \\
M''_{\boldsymbol{X}}(t) &= \left(\mu\mathrm{e}^t + \mu^2\mathrm{e}^{2t}\right)\mathrm{e}^{\mu(\mathrm{e}^t-1)}
\end{aligned}$$

From this we get

$$\begin{aligned}
\mu &= M'_{\boldsymbol{X}}(0) = \mu \\
\sigma^2 &= M''_{\boldsymbol{X}}(0) - \mu^2 = \mu
\end{aligned}$$

I.e.The Poisson distribution has variance equal to the mean

4.6 The Exponential Distribution

The Poisson Distribution gave the expected number of events in a fixed time interval.

Now consider the time between events

Let $\boldsymbol{Y}$ be the r.v. denoting time to the next event, and consider $Pr(\boldsymbol{Y} \leq y)$, with $\Omega = [0,\infty)$

Divide time into intervals $\delta t = \frac{1}{n}$.

$$
\begin{array}{rl}
Pr(\boldsymbol{Y} \leq y) = Pr(\text{event in next interval}) & OR \\
Pr(\text{no event in } \delta t) \cdot Pr(\text{event in next interval}) & OR \\
Pr(\text{no event in } 2\delta t) \cdot Pr(\text{event in next interval}) & OR \\
\vdots &
\end{array}
$$

$$
\begin{aligned}
\Rightarrow Pr(\boldsymbol{Y} \leq y) &= a\left(1 + (1-a) + \ldots (1-a)^{ny-1}\right) \\
&= a\left(\frac{1-(1-a)^{ny}}{1-(1-a)}\right) \\
&= 1 - \left(1 - \frac{\mu}{n}\right)^{ny}
\end{aligned}
$$

where $a = \mu\delta t = \frac{\mu}{n}$ Now let $n \to \infty$

$$Pr(\boldsymbol{Y} \leq y) = 1 - \mathrm{e}^{\mu y} \quad y \geq 0$$

This is the distribution function $F_{\boldsymbol{Y}}(y)$ so the PDF is

$$f_{\boldsymbol{Y}}(y) = F'_{\boldsymbol{Y}}(y) = \mu\mathrm{e}^{-\mu y} \quad y \geq 0$$

This is the *exponential distribution* and is a model for the "time between random events". The MGF is

$$
\begin{aligned}
M_{\boldsymbol{Y}}(t) &= \int_0^\infty \mathrm{e}^{ty}\mu\mathrm{e}^{-\mu y}\mathrm{d}y = \mu\int_0^\infty \mathrm{e}^{-y(\mu-t)}\mathrm{d}y \\
&= \frac{\mu}{\mu - t} = \left(1 - \frac{t}{\mu}\right)^{-1}
\end{aligned}
$$

Taking the Taylor series explicitly

$$M_{\boldsymbol{Y}}(t) = 1 + \frac{t}{\mu} + \frac{t^2}{\mu^2} + \ldots$$

From this we get

$$
\begin{aligned}
\mu &= M'_{\boldsymbol{Y}}(0) = \frac{1}{\mu} \\
\sigma^2 &= M''_{\boldsymbol{Y}}(0) - \mu^2 = \frac{1}{\mu^2}
\end{aligned}
$$

4.7 The Normal Distribution

The normal distribution is defined

$$f_{\boldsymbol{X}}(x) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad x \in (\infty, \infty)$$

The MGF is

$$M_{\boldsymbol{X}}(t) = \mathrm{e}^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

from which we get

$$E[\boldsymbol{X}] = \mu\,, \quad Var[\boldsymbol{X}] = \sigma^2$$

Defining a new variable $\boldsymbol{Z} = \frac{\boldsymbol{X}-\mu}{\sigma}$ we have

$$M_{\boldsymbol{Z}}(t) = \mathrm{e}^{\frac{1}{2}t^2}$$

from which we get

$$E[\boldsymbol{Z}] = 0\,, \quad Var[\boldsymbol{Z}] = 1$$

This is called the "standard Normal Distribution".

The Normal Distribution is of great importance, because in most cases, in the limit of sufficiently large populations, all distributions tend to this distribution. (this is the *Central Limit Theorm*).

5 Multivariate Distributions

if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two random variables, the event $(\boldsymbol{X} = x \,\&\, \boldsymbol{Y} = y)$ has probability

$$Pr(\boldsymbol{X}, \boldsymbol{Y} \in [x, x+\delta x], [y, y+\delta y]) = f_{\boldsymbol{X},\boldsymbol{Y}}(x, y)\delta x \delta y$$

which is the probabilty that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are in the neighbourhood $[x, x+\delta x], [y, y+\delta y]$ around x, y. $f_{\boldsymbol{X},\boldsymbol{Y}}(x, y)$ is a *multivariate PDF*

$$Pr(\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{N}(x, y)) = \int_{\mathcal{N}(x,y)} f_{\boldsymbol{X},\boldsymbol{Y}}(x, y)\mathrm{d}x\mathrm{d}y$$

and by definition

$$\int_{\Omega} f_{\boldsymbol{X},\boldsymbol{Y}}(x, y)\mathrm{d}x\mathrm{d}y = 1$$

5.1 Joint PDFs and Marginalisation

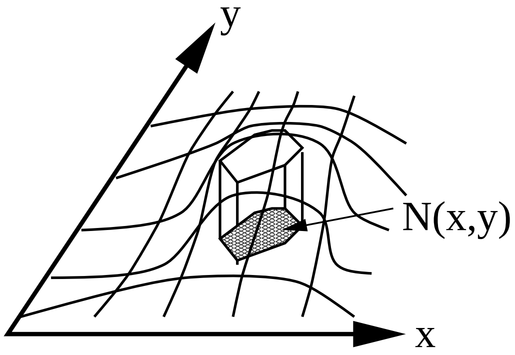
The generalisation to more variables is obvious. We define a *vector random variable*

$$\vec{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_N \end{pmatrix}$$

with associated PDF $f_{\vec{\boldsymbol{X}}}(\vec{x})$, which is the *joint PDF* of $\{x_1, x_2, \ldots x_N\}$.

Give a joint PDF the distribution of any subset of the r.v.s can be found by integrating out the other r.v.s. This is called *Marginalisation* and the integrated out variibles are sometimes called *nuisance parameters*. E.g.

$$\begin{aligned} f_{\boldsymbol{X}}(x) &= \int_{-\infty}^{\infty} f_{\boldsymbol{X},\boldsymbol{Y}}(x, y)\mathrm{d}y \\ f_{\boldsymbol{X}_i}(x) &= \int_{\Pi_{j\neq i}\Omega_j} f_{\vec{\boldsymbol{X}}}(\vec{x})\mathrm{d}^{n-1}x_{j\neq i} \end{aligned}$$



More generally if we define a dependency

$$\boldsymbol{G} = a_1\boldsymbol{X}_1 + a_2\boldsymbol{X}_2 + \ldots a_n\boldsymbol{X}_n = \vec{a}\cdot\vec{\boldsymbol{X}}$$

then $\boldsymbol{G} = K$ represents a hyper-plane in $\vec{x}$-space. Integration over the orthogonal space to this hyperplane gives the PDF of $\boldsymbol{G}$

$$f_{\boldsymbol{G}}(g) = \int_{\Omega\perp(\vec{a}\cdot\vec{\boldsymbol{X}}=K)} f_{\vec{\boldsymbol{X}}}(\vec{x})\mathrm{d}^{n-1}\vec{x}$$

5.2 Conditional PDFs

Joint PDFs and marginalisation can be related to the *conditional probability*

$$f_{\boldsymbol{X}}(x|y) := pr(\boldsymbol{X} = x \text{ given } \boldsymbol{Y} = y)$$

In fact we have

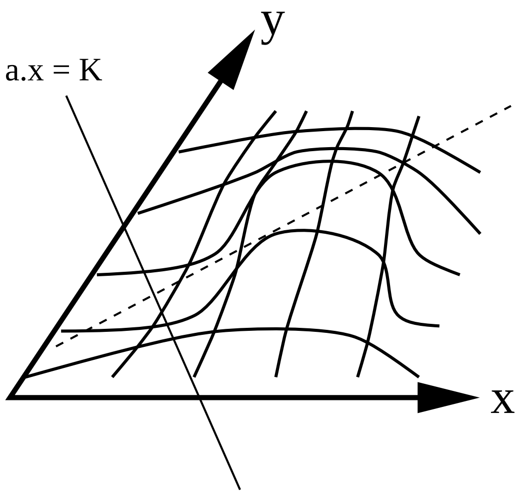
$$f_{\boldsymbol{X},\boldsymbol{Y}}(x,y) = f_{\boldsymbol{X}}(x|y)f_{\boldsymbol{Y}}(y)$$

By induction this extends to any number of variables, for example

$$\begin{aligned} f_{\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3}(x_1,x_2,x_3) &= f_{\boldsymbol{X}_1,\boldsymbol{X}_2}(x_1,x_2|x_3)f_{\boldsymbol{X}_3}(x_3) \\ &= f_{\boldsymbol{X}_1}(x_1|x_2,x_3)f_{\boldsymbol{X}_2}(x_2|x_3)f_{\boldsymbol{X}_3}(x_3) \end{aligned}$$

The random variables are *independent* iff the joint PDF (and therefore each marginalisation) *factorises* into the product of the approriate set of marginal univariate distributions

$$f_{\vec{\boldsymbol{X}}}(\vec{x}) = f_{\boldsymbol{X}_1}(x_1)f_{\boldsymbol{X}_2}(x_2)\cdots_{\boldsymbol{X}_N}(x_N) = \Pi_{i=1}^{n} f_{\boldsymbol{X}_i}(x_i)$$



5.3 Summary Measures and Relationships

The notions of expected value and moment generating functions apply directly to multivariate distributions. In particular

$$E[\boldsymbol{X}_i] = \int_\Omega x_i f_{\vec{\boldsymbol{X}}}(\vec{x}) \mathrm{d}^n \vec{x}$$

and if $G(\vec{\boldsymbol{X}})$ is any function of the r.v.s then the MGF of $\boldsymbol{G}$ is

$$M_{\boldsymbol{G}}(t) = E\left[e^{tG(\vec{\boldsymbol{X}})}\right]$$

One measure of relationship between variables is the *covariance*

$$\begin{aligned} Cov[\boldsymbol{X},\boldsymbol{Y}] &= E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{Y} - E[\boldsymbol{Y}])] \\ &= E[\boldsymbol{X}\boldsymbol{Y}] - E[\boldsymbol{X}]E[\boldsymbol{Y}] \end{aligned}$$

Note :

$$Cov[\boldsymbol{X},\boldsymbol{X}] = Var[\boldsymbol{X}]$$

The Cauchy-Schwartz inequality shows that

$$\begin{aligned} &\left(\int_\infty^\infty \int_\infty^\infty (x-\mu_x)(y-\mu_y) f_{\boldsymbol{X},\boldsymbol{Y}}(x,y)\mathrm{d}x\mathrm{d}y\right)^2 \leq \\ &\left(\int_\infty^\infty (x-\mu_x)^2 f_{\boldsymbol{X}}(x)\mathrm{d}x\right)\left(\int_\infty^\infty (y-\mu_y)^2 f_{\boldsymbol{Y}}(y)\mathrm{d}y\right) \\ &\Rightarrow \quad Covar[\boldsymbol{X},\boldsymbol{Y}]^2 \leq Var[\boldsymbol{X}]Var[\boldsymbol{Y}] \end{aligned}$$

From this we define the *correlation coefficient*

$$\rho(\boldsymbol{X},\boldsymbol{Y}) = \frac{Covar[\boldsymbol{X},\boldsymbol{Y}]}{\sqrt{Var[\boldsymbol{X}]Var[\boldsymbol{Y}]}}$$

The correlation coefficient is always in the range $[-1,1]$.

The C-S inequality becomes equality when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are linearly related and zero if they are independent. **But** zero covariance does **not** imply that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.

5.4 The Multivariate Normal Distribution

let $\boldsymbol{X}$ be a vector of r.v.s with mean vector $\vec{\mu}$ and (symmetric positive definite) covariance matrix C. The multivariate normal distribution has PDF

$$f_{\vec{\boldsymbol{X}}}(\vec{x}) = \frac{\mathrm{e}^{-\frac{1}{2}(\vec{x}-\vec{\mu})^{\mathrm{T}}\mathsf{C}^{-1}(\vec{x}-\vec{\mu})}}{\sqrt{(2\pi)^n det(\mathsf{C})}}$$

Consider the case $n = 2$

$$\begin{aligned}
\mathsf{C} &= \begin{pmatrix} \sigma_x^2 & cov_{xy} \\ cov_{xy} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \\
det(\mathsf{C}) &= (1-\rho^2)\sigma_x^2\sigma_y^2 \\
\mathsf{C}^{-1} &= \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x\sigma_y} \\ -\frac{\rho}{\sigma_x\sigma_y} & \frac{1}{\sigma_y^2} \end{pmatrix}
\end{aligned}$$

We can write the 2D PDF explicitly:

$$f_{\boldsymbol{X},\boldsymbol{Y}}(x,y) = \frac{\mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

Any sym.pos.def. matrix may be diagonalised:

$$\mathsf{C} = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix}\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \end{pmatrix}$$

where $\left\{\sigma_k^2, \vec{u}_k\right\}$ is the eigensystem of C:

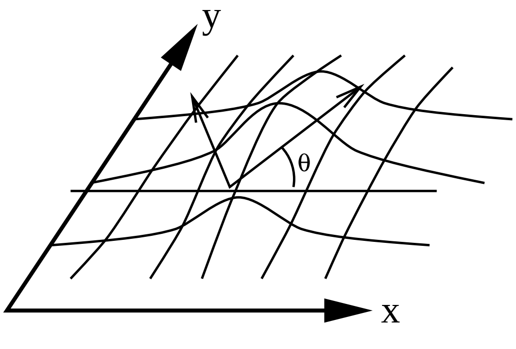
$$\mathsf{C}\begin{pmatrix} u_{k1} \\ u_{k2} \end{pmatrix} = \sigma_k^2\begin{pmatrix} u_{k1} \\ u_{k2} \end{pmatrix}$$

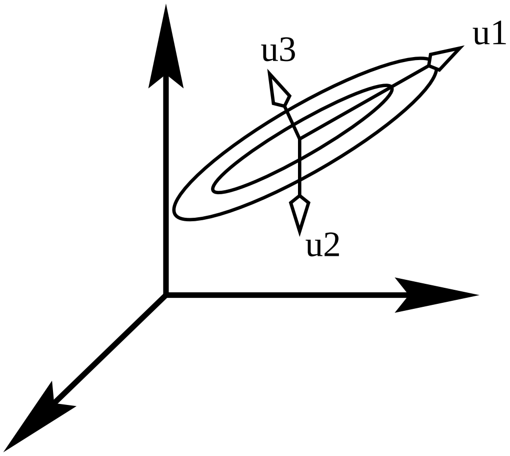
$\{\vec{u}_1, \vec{u}_2\}$ forms an orthogonal coordinate system:

$$f_{\boldsymbol{U}_1,\boldsymbol{U}_2}(u_1,u_2) = \frac{\mathrm{e}^{-\frac{1}{2}\left[\frac{\vec{e}_1\cdot\vec{e}_1}{\sigma_1^2} + \frac{\vec{e}_2\cdot\vec{e}_2}{\sigma_2^2}\right]}}{2\pi\sigma_1\sigma_2} = \frac{\mathrm{e}^{-\frac{1}{2}\frac{\vec{e}_1\cdot\vec{e}_1}{\sigma_1^2}}}{\sqrt{2\pi}\sigma_1}\frac{\mathrm{e}^{-\frac{1}{2}\frac{\vec{e}_2\cdot\vec{e}_2}{\sigma_2^2}}}{\sqrt{2\pi}\sigma_2}$$

where $\vec{e}_1 = \vec{u}_1\cdot(\vec{\boldsymbol{X}} - \vec{\mu}), \vec{e}_2 = \vec{u}_2\cdot(\vec{\boldsymbol{X}} - \vec{\mu})$. The new coordinate system is in terms of r.v.s that are *independent*

$\vec{e}_1$ is normal with variance σ_1^2





$\vec{e}_2$ is normal with variance σ_2^2

This is easily generalised to n dimensions

Note: It is always possible to find a covariance matrix, but this will *not* guarantee independent r.v.s

6 Goodness of Fit

6.1 Hypothesis Testing

The concept of *goodness of fit* or *hypothesis testing* is at the heart of applications of probability theory and statistics.

We can state the principles as

1. Write down precisely the hypothesis to be tested [the *Null Hypothesis*]
2. Set up a mathematical model of the situation
3. Define a measure of badness of fit
4. Find the probability with which, on the hypothesis, this measure takes a value less than the one observed
5. Discuss whether the *evidence* supports or contradicts the hypothesis

Example Median annual wheat yield is quoted as 2.5 tonnes per hectare.

A sample of farms produce the following values

2.26	2.66	1.86	2.14	2.82	2.44	2.32	2.36

do these figures support or contradict the quoted figure ?

1 Hypothesis : *The actual yield is a r.v with a PDF whose median is 2.5*

By definition of median, the probability of being above or below the median is the same.

Without any further knowledge or assumptions about the underlying PDF we can set up model of *Bernouilli Trials:*

2 Model : *the number of fields with yield over median is a r.v. with PDF Binomial(a,n)*

$$\begin{aligned} Pr(\text{yield} > \text{median}) &= a = \frac{1}{2} \\ Pr(k \text{ fields have yield }) > \text{median} &= \binom{n}{k} a^k (1-a)^{n-k} \end{aligned}$$

3 Badness of Fit

The number of yields above the median is a r.v. $\boldsymbol{X}$, and $E[\boldsymbol{X}] = na =: \mu$

For 8 samples $\mu = 4$, so $|x - 4|$ is an appropriate measure (absolute distance from the mean)

4 Probability of obtaining measure value

Observed value of x is 2.

Required probability is

$$\begin{aligned} Pr(|x-4| \geq 2) &= Pr(x \in \{0,1,2,6,7,8\}) \\ &= \left\{ 2\left(\frac{1}{2}\right)^8 + 2\binom{8}{1}\left(\frac{1}{2}\right)^8 + 2\binom{8}{2}\left(\frac{1}{2}\right)^8 \right\} \\ &\simeq 0.29 \end{aligned}$$

5 Discussion

The chance of getting two (or more) out of eight yields above the median is about three in ten. There is nothing unusual about this, so the evidence does not contradict the hypothesis

6.2 Levels of Significance

Suppose we had 10 samples, all above the mean, we would have

$$Pr(|x-5| = 5) = Pr(x \in \{0, 10\}) = 2\left(\frac{1}{2}\right)^{10} \simeq 0.002$$

This is very unlikely, so is strong evidence that the hypothesis is wrong.

If we had 15 out of 20 successes we would get

$$Pr(|x-10| \geq 5) = 2\left(\frac{1}{2}\right)^{20} \sum_{k=0}^{5} \binom{20}{k} \simeq 0.041$$

Is this significant or not?

Significance is quoted in terms of a *level*, e.g.

10/10 is significant at 0.1% level

15/20 is significant at 5% level

FALSE NEGATIVE : The significance level tells us the probability of *rejecting* the hypothesis when it was actually *true.*

FALSE POSITIVE : Alternatively, we can assess the probability of *accepting* the hypothesis when it was actually *false.*

Under different circumstances, one type of error may be more critical than the other.

For example, diagnosis of cancer would rather accept false positives than false negatives

6.3 Standardised Measures

The *normal distribution* is frequently used as a standardised measure.

If a random variable is assumed normal with mean μ and variance σ^2, then the probability of obtaining a value x is given by

$$Pr(|x-\mu| \geq |x_k-\mu|) = 2\int_{x_k}^{\infty} \frac{\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\mathrm{d}x$$

By transformation of variables we can express this in terms of the standardised variable $\boldsymbol{Z} = \frac{\boldsymbol{X}-\mu}{\sigma}$ with mean zero, variance 1.

$$\begin{aligned}Pr(|x-\mu| \geq |x_k-\mu|) &= Pr(|z| \geq z_k) \\ &= 2\int_{z_k}^{\infty} \frac{\mathrm{e}^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\mathrm{d}z \\ &= 2(1-erf(z))\end{aligned}$$

where $erf(z)$ is the error function.

The z variable is "distance from mean in units of standard deviation".

It is tabulated, or can be calculated by standard packages such as Matlab or Mathematica.

We find

P	99	95	10	5	1	0.1
z	0.0125	0.0628	1.64	1.96	2.58	3.29

For example a value of $z = 2.5$ is "almost significant at the 1% level".

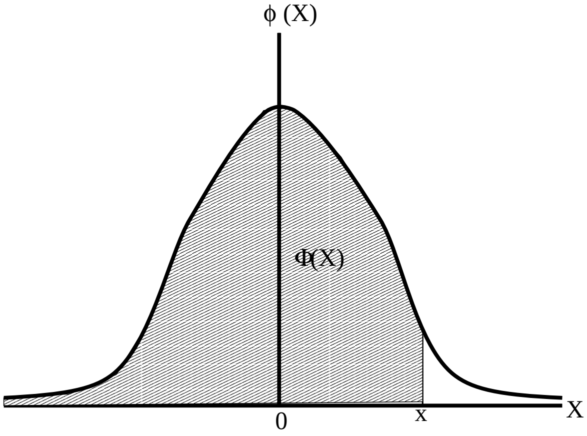
6.4 χ^2 on 1 degree of freedom

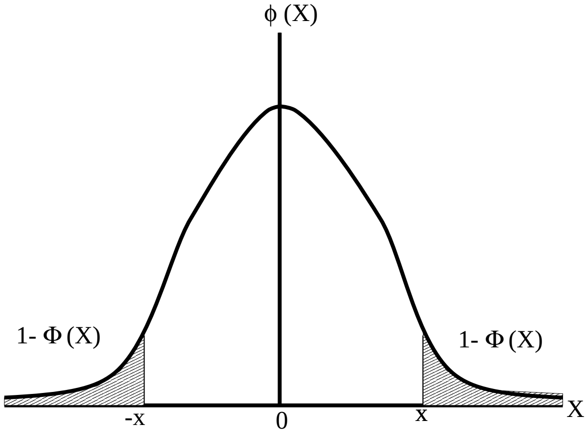
Suppose we used the variable

$$\boldsymbol{U} = \frac{(\boldsymbol{X}-\mu)^2}{\sigma^2} = \boldsymbol{Z}^2$$

The probabilities remain unchanged:

$$Pr(\boldsymbol{U} \geq u) = Pr(\boldsymbol{Z} \geq z) = Pr(|\boldsymbol{X}-\mu| \geq |x-\mu|)$$


ϕ (X)
Φ(X)
0
x
X


ϕ (X)
1- Φ (X)
1- Φ (X)
-x
0
x
X

The distribution of the square of a standardised normal r.v. is called χ^2 with one degree of freedom. It has MGF

$$\begin{aligned} M_{\chi^2}(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{tz^2-\frac{1}{2}z^2}\mathrm{d}z \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{\frac{1}{2}z^2(1-2t)}\mathrm{d}z \\ &= (1-2t)^{-\frac{1}{2}} \end{aligned}$$

6.5 χ^2 on n degrees of freedom

The sum of n independent standard normally distributed random variables is a r.v. with mean zero and variance $\sigma^2 = n$.

To see this, change variables :

$$\begin{aligned} y = e_0 &= \frac{1}{\sqrt{n}}(x_1 + x_2 + \ldots x_n) \\ e_1 &= \frac{1}{\sqrt{2}}(x_1 - x_2) \\ &\vdots \\ e_{n-1} &= \frac{1}{\sqrt{2}}(x_{n-1} - x_n) \end{aligned}$$

The new variables $\{e_0 \ldots e_{n-1}\}$ are all orthogonal, and standard normally distributed . The PDF for z is found by marginalising over the other variables

$$\begin{aligned} f_{\mathbf{Y}}(y) &= \frac{1}{\sqrt{2\pi^n}}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\left(y^2+e_1^2+\ldots e_{n-1}^2\right)}\Pi_{j=1}^{n-1}\mathrm{d}e_j \\ &= \frac{1}{\sqrt{2n\pi}}\mathrm{e}^{-\frac{1}{2n}(x_1+x_2+\ldots x_n)^2} \quad = N(0,\sqrt{N}) \end{aligned}$$

The sum of random variables is not a sufficient statistic itself to test significance.

Testing this statistic only tells us if the mean of a distribution is close to the hypothetical mean.

Example
The life of light bulbs is supposed to be $N(900, 80)$ (in hours). Three bulbs are tested and fail at 1060, 700 and 920 hours. Put

$$y = \sum x_i = 2680$$

which should be distributed $N(2700, 80\sqrt{3})$

in normalised terms, the samples are

$$\begin{aligned} z &= 2, -2.5, 0.25 \\ \Rightarrow \tilde{y} &= 0.25 \end{aligned}$$

Table 1: Percentage Points of the χ^2 probability function

P	99	95	10	5	1	0.1
$\nu = 1$	0.000157	0.00393	2.71	3.84	6.63	10.83
2	0.0201	0.102	4.61	5.99	9.21	13.75
3	0.115	0.352	6.25	7.81	11.34	16.27
4	0.297	0.711	7.78	9.49	13.28	18.47
5	0.554	1.15	9.24	11.07	15.09	20.51
6	0.873	1.64	10.64	12.59	16.81	22.46
7	1.24	2.17	12.02	14.07	18.47	24.32
8	1.65	2.73	13.36	15.51	20.09	26.12
9	2.09	3.33	14.68	16.92	21.67	27.88
10	2.56	3.94	15.99	18.31	23.21	29.59
12	3.57	5.23	18.55	21.03	26.21	32.91
16	5.81	7.96	23.54	26.30	32.00	39.25
20	8.26	10.85	28.41	31.41	37.57	45.31
30	14.96	18.49	40.26	43.77	50.89	59.66
40	22.17	26.51	51.81	55.76	63.69	73.40

Under the assumption that $\tilde{y}$ has PDF $N(0, \sqrt{3})$, these results are not at all significant, even though in individual terms two out of three are significant at 5% level.

The sum of squares of k independent normally distributed random variables is called "χ^2 on k degrees of freedom".

It is tabulated or can be calculated

In the above case we have

$$\chi^2 = 2^2 + (2.5)^2 + (0.25)^2 \simeq 10.3$$

and this is significant above the 5% level.

from the MGF for χ_1^2 and the property of independent r.v.s we have the MGF for $\chi^2_{\nu=n}$

$$M_{\chi^2_n}(t) = (1 - 2t)^{-\frac{n}{2}}$$

6.6 Application of χ^2

Whenever data are *independent* samples of r.v.s the χ^2 statistic can be applied to each sample (as 1 d.o.f) or to their sum (as n d.o.f). The appropriate sample variances should be used.

Example : Mendel (1865), the founder of the theory of genetics, did experiments on cross breeding plants and examining characteristics of the resultant hybrids. In one experiment on pea-plants he found that some seeds were angular and some round in the following table.

Plant No.	1	2	3	4	5	6	7	8	9	10	Total
Round	45	27	24	19	32	26	88	22	28	25	5474
Angular	12	8	7	10	11	6	24	10	6	7	1850

Based on the total $n = 7324$ the total number of round seeds is in a $3 : 1$ ratio to the number of angular. His hypothesis is that the number of seeds that are angular is a binomial PDF with $a = 0.25$. We can analyse each plant separately as a $\chi^2_{\nu=1}$ with one degree of freedom, assuming a Bernouilli trial with seperate n. Or we can treat the sum as $\chi^2_{\nu=10}$ with $n = 7324$. The statistics come out like this (taking $\boldsymbol{X}$ as "angular seed")

No.	1	2	3	4	5	6	7	8	9	10
Round	45	27	24	19	32	26	88	22	28	25
Angular	12	8	7	10	11	6	24	10	6	7
n	57	35	31	29	43	32	112	32	34	32
na	14.5	8.75	7.75	7.25	10.75	8	28	8	8.5	8
$na(1-a)$	10.7	6.56	5.19	5.44	8.06	6	21	6	6.33	6
$\frac{(\boldsymbol{X}-na)^2}{na(1-a)}$	0.48	0.08	0.1	1.4	0.01	0.7	0.76	0.7	0.99	0.17

None of these is significant at the normal levels, meaning that none contradict the hypothesis. In addition, for the sum

$$\boldsymbol{Z}^2 = \sum_{i=1}^{10} \frac{(\boldsymbol{X}_i - n_i a)^2}{n_i a(1-a)} \rightarrow 5.30$$

This is also not significant for the $\chi^2_{\nu=10}$ statistic.

6.7 Linear Constraints

Often the measured results are *not* independent, but are restricted to having a given total, so that at least one can be inferred from the others.

Example A die is thrown 180 times and 40 sixes are obtained. Is this unusual?

A_1 (sixes)	A_2 (non-sixes)	Total
$x_1 = 40$	$x_2 = 140$	$n = 180$

We have : $a_1 = \frac{1}{6}, a_2 = \frac{5}{6}, E[A_1] = na_1 = 30, E[A_2] = na_2 = 150, \sigma^2 = na_1 a_2 = 25$. We can take either the sixes or non-sixes as a $\chi^2_{\nu=1}$ statistic

$$\boldsymbol{Z}^2 = \frac{(x_1 - E[A_1])^2}{\sigma^2} = \frac{(x_2 - E[A_2])^2}{\sigma^2} = \frac{10^2}{25} = 4$$

But we also have the symmetrical relation

$$\boldsymbol{Z}^2 = \sum_{i=1}^{2} \frac{(x_i - E[A_i])^2}{E[A_i]} = \frac{(x_1 - E[A_1])^2}{\sigma^2} = \frac{(x_2 - E[A_2])^2}{\sigma^2}$$

I.e. two *independent* Poisson variables The usual rule when we have n frequencies with m constraints is to calculate

$$\boldsymbol{Z}^2 = \sum_{i=1}^{n} \frac{(x_i - E[A_i])^2}{E[A_i]}$$

and treat this as $\chi^2_{\nu=(n-m)}$ because each constraint removes one degree of freedom.

Example A die is thrown 180 times with results

1	2	3	4	5	6	Total
40	35	30	30	25	20	$n = 180$

If the die is fair then each number has $a_i = \frac{1}{6}$ and we have

$$\boldsymbol{Z}^2 = \sum_{i=1}^{6} \frac{(x_i - E[A_i])^2}{E[A_i]} = \frac{(x_i - 30)^2}{30} = 8.33$$

which is not significant for $\chi^2_{\nu=5}$.

6.8 Contingency Tables

When data is presented as a contingency table, the null hypothesis is that the two analyses represented by the rows and columns are statistically independent.

Example The following data are obtained for symptoms of a disease vs. concentration of a certain drug

	no symptoms	mild	severe
High concentration	8	8	6
Low concentration	15	5	3

We can analyse the 45 patients as : i) 22 with high concentrations vs 23 with low, ii) 23 with no symptoms, 13 with mild, and 9 with severe.
Null Hypothesis : the drug concentration is statistically independent of the symptoms Taking each entry in the table as Poisson, and assuming independence we get the joint probabilities

$$\begin{aligned} Pr(\text{conc.} = x_i \,\&\, \text{symp.} = y_j) &= \\ Pr(\text{conc.} = x_i) Pr(\text{symp.} = y_j) &= \frac{x_i}{\sum_i x_i} \frac{y_j}{\sum_j y_j} \end{aligned}$$

Then we get the expected frequencies

11.25	6.35	4.40	22
11.75	6.65	4.60	23
23	13	9	45

Finally we get the statistic

$$\boldsymbol{Z}^2 = 0.94 + 0.43 + 0.58 + 0.9 + 0.41 + 0.56 = 3.82$$

The number of constraints are the row and column sums minus one (since the total number of patients is given by only one of the total row or column sums). In this case $3 + 2 - 1 = 4$. So the no. d.o.f is 2. (In general for n rows and $ncolumns$ the no. d.o.f is $(n-1)(m-1)$). For this example, 3.82 is not significant for $\chi^2_{\nu=2}$. **Example.** The number of people attacked by typhoid, and dependence on state of inoculation are :

	Attacked	Not Attacked
Inoculated	56	6759
Not inoculated	272	11396

The null hypothesis is that attack is statistically independent of inoculation. The expectations and totals are

121	6694	6815
207	11461	11668
328	18155	18483

$$\boldsymbol{Z}^2 = \frac{65}{121} + \frac{65}{207} + \frac{65}{6694} + \frac{65}{11461} = 56.3$$

This is assessed on $\chi^2_{\nu=1}$ and is highly significant.

The implication is that these data are strong evidence that inoculation lowers the chance of being attacked by typhoid.

6.9 Rules

1. **Use for frequencies only** : The χ^2 statistic assumes that individual frequencies, constrained by linear relations, behave as binomial or Poisson variables, which tend to Normal. Cannot be used generally for arbitrary r.v.s.
2. **Do not use for comparison** : It is not true that a hypothesis that gives a lower χ^2 is a better fit. For example 212 heads out of 400 coin tosses has $\boldsymbol{Z}^2 = 1.44$ which is not significant. But we can reduce the statistic to zero by using $a = 0.53$ (a slightly biased coin). But this does not make the ccoin biased.
3. **Too good a fit** if a coin is tossed 400 times and 200 heads come up, the probability is 100% "too good to be true". This can be used as evidence that data has been "cooked"

6.10 The mathematics of χ^2

The χ^2 on 1 d.o.f is easily derived from the normal function

$$\phi(x) = f_{\boldsymbol{X}}(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}$$

Consider $\boldsymbol{U} = \boldsymbol{X}^2$. Then

$$\begin{array}{rcl} Pr(u < \boldsymbol{U} < u + \delta u) & = & Pr(\chi^2 < \boldsymbol{U} < \chi^2 + 2\chi\delta\chi) \\ 2\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \delta x & \simeq & \frac{2}{\sqrt{2\pi}} u^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}u} \delta u \end{array}$$

For the general $\chi^2_{\nu=n}$ distribution we consider $\boldsymbol{U} = \boldsymbol{X}_1^2 + \boldsymbol{X}_2^2 + \ldots \boldsymbol{X}_n^2$. To take all the contributions to $Pr(u < \boldsymbol{U} < u + \delta u)$ into account we consider the hypersphere

$$\boldsymbol{X}_1^2 + \boldsymbol{X}_2^2 + \ldots \boldsymbol{X}_n^2 = \boldsymbol{U}$$

The tabulated function is then

$$\frac{\int_{\chi^2_{\nu=n}}^{\infty} u^{\frac{1}{2}(n-1)} \mathrm{e}^{-\frac{1}{2}u} \mathrm{d}u}{\int_0^{\infty} u^{\frac{1}{2}(n-1)} \mathrm{e}^{-\frac{1}{2}u} \mathrm{d}u}$$

The density function for $\chi^2$is

$$f_{\chi^2_{\nu=n}} = \frac{u^{\frac{1}{2}(n-1)} \mathrm{e}^{-\frac{1}{2}u}}{2^{\frac{1}{2}n} \Gamma(\frac{n}{2})}$$

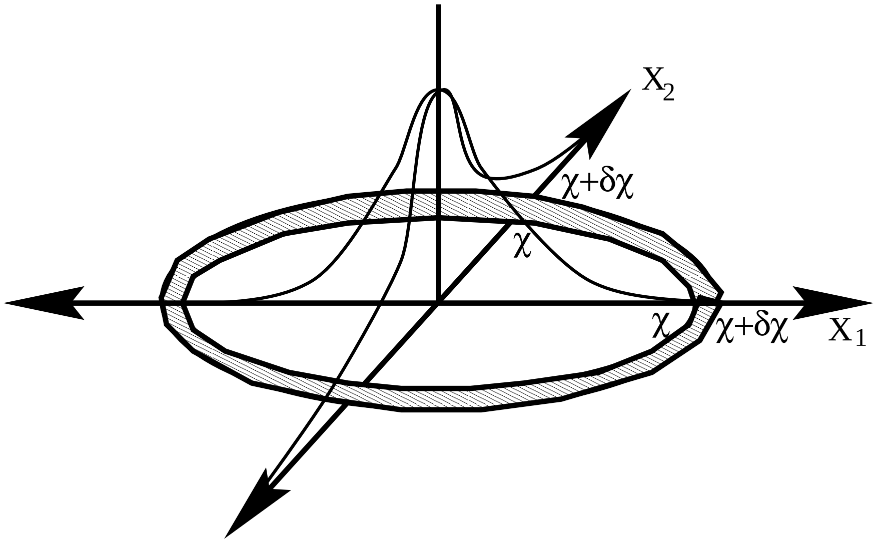


Figure 3: $\chi^2_{\nu=2}$ is the integral over a hypersphere of the product of normal univariate PDFs

6.11 Exercises

1. You are told that one person in three is left-handed. You need to find a left-handed person to ask questions for a survey. You ask seven people before finding a left-handed person. Should this cause you to doubt the original statement?

2. Light bulbs are supposed to have a life whose PDF is Normal with mean 900 hours and standard deviation 80 hours. You test a bulb to destruction and find that it lasts a) 1060 hours, b) 700 hours. Do either of these tests cast doubt on the original statement?

3. I.Q.'s are supposed to form a Normal population with mean 100 and standard deviation 15. A group of five people have I.Q.s $\{95, 105, 124, 130, 133\}$. Are these a particularly unrepresentative group?

4. In an experiment with two dice, the following are observed

Score	2	3	4	5	6	7	8	9	10	11	12
Frequency	3	11	11	9	12	16	19	11	8	5	2

is this result indication of any unfairness in the dice?

5. In a survey of the maintenance habits of car owners, 1890 car tyres were inspected. Examine the following data for evidence as to whether or not the spare wheel is exchanged regularly with other wheels on the car.

	new	part-worn	worn	badly worn
wheels in use	350	902	288	73
spare wheel	107	79	59	32

6.12 Answers

1. No, $\left(\frac{2}{3}\right)^7 = 0.0585$ so this event is not significant at 5% level (just)
2. a) $\chi_1^2 = 4 => P = 0.046$ b) $\chi_1^2 = 6.25 => P = 0.012$. Both of these are significant at 5% level, so the statement is probably wrong.
3. $Z = 11.6$ this is just significant for χ_5^2 at 5% level.
4. group together the figures for 2,3, 11, and 12, and use χ^2 with 8 degrees of freedom ($\rightarrow Z = 6.3$). This is not significant.
5. 4×2 table with 5 constraints $\Rightarrow \chi^2$ with 3 degrees of freedom.$Z = 85.2$ which is very significant.

7 Correlation and Regression

Frequently we measure data in pairs, either with one of the pair *independent* and the other *dependent*, or both dependent. If the data match a predictive model, we will expect the dependent data to be wholly determined by the model, i.e. the hypothesis that they are independent will be disproved. In general the model will be non-linear

$$y_i = F(x_i) + e_i$$

where y_i is a sample of the dependent variable, x_i is a sample of the independent variable, and e_i is noise. The simplest case is linear where we will state

$$y_i = mx_i + c + e_i$$

and we effectively are looking for a straight line fit of slope m and offset c. In both the linear and non-linear case we need an assessment of how well the data is fit by the model

7.1 Least Square Criterion

A widely used criterion is to minimise the errors

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - mx_i - c)^2$$

To simplify the analysis, shift to coordinates centred at the mean

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \hat{\mu}_y \; ; \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}_x$$

In the (m, c) search space the minimiser of S^2 is also the minimiser of

$$\begin{aligned} \tilde{S}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left((y_i - \overline{y}) - m(x_i - \overline{x}) - c\right)^2 \\ &= s_y^2 + m^2 s_x^2 - 2ms_{xy} + c^2 \end{aligned}$$

where

$$\begin{aligned} s_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \hat{\sigma}_y^2 \\ s_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \hat{\sigma}_x^2 \\ s_{xy}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) = c\hat{o}v[xy] \end{aligned}$$

are estimates of the variances and covariances of (x, y). The minimum of $\tilde{S}^2$ is

$$c = 0\,, m = \frac{s_{xy}}{s_x^2}$$

In the original coordinate system the line of best fit is therefore

$$y = \frac{s_{xy}}{s_x^2} x + \left(\overline{y} - \frac{s_{xy}}{s_x^2} \overline{x} \right)$$

which gives a total error of

$$S^2 = s_y^2 \left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2} \right)$$

7.2 Correlation

We define

$$r = \frac{s_{xy}}{s_x s_y} = \hat{\rho}$$

as the estimate of the *correlation coefficient*, and we have

$$S^2 = s_y^2 (1 - r^2)$$

If the sum squared error S^2 were zero the $r = \pm 1$ and the data are completely correlated, i.e. y_i is completely determined by x_i.

We also interpret S^2 as the part of the variance unaccounted for by the linear relation.

We can use r to test significance. If x and y are independent and drawn from normal distribution then

$$\frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \text{ is } N \left(0, \frac{1}{\sqrt{n-3}} \right)$$

Example

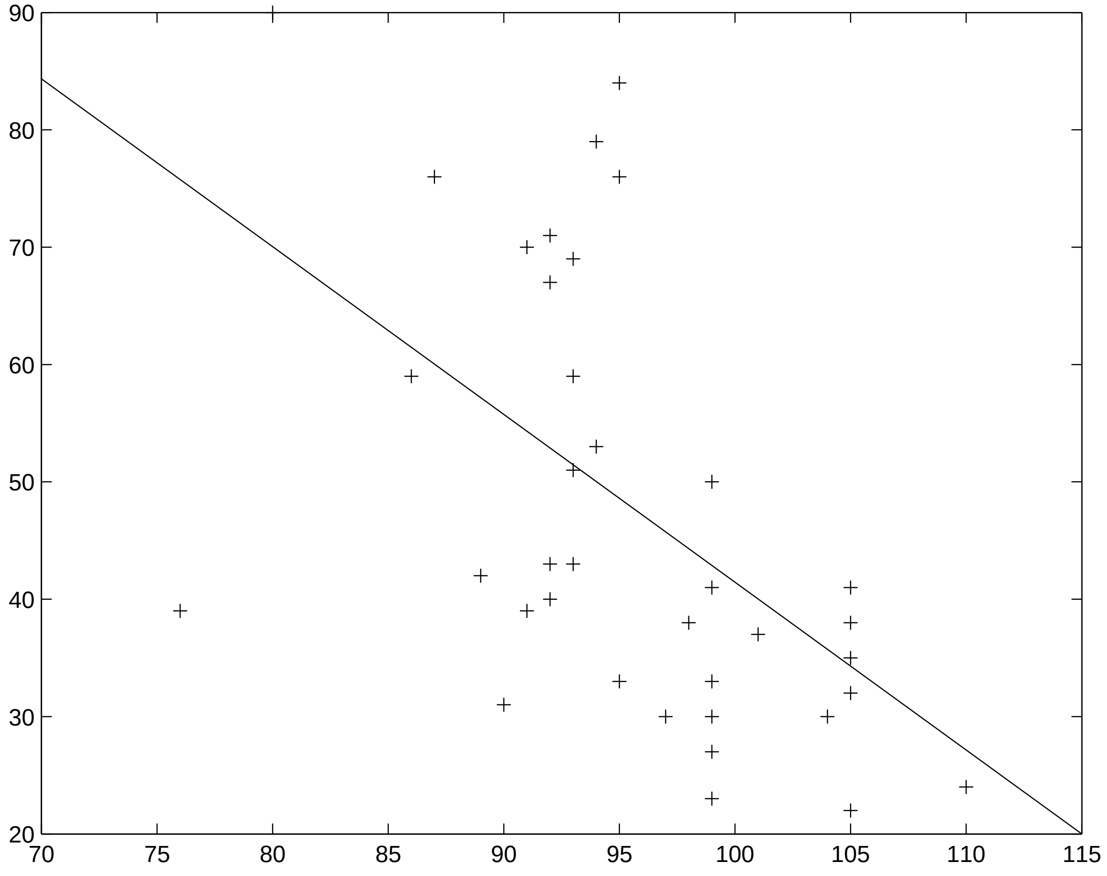
To assess the significance of r we test

$$\frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) = -0.600$$

against normal

$$N \left(0, \frac{1}{\sqrt{n-3}} \right) = N(0, 0.1715)$$

This is highly significant and strong evidence that the data are correlated, despite the wide distribution of the y values.



x = (76 87 92 92 93 93 95 98 99 99 105 105 101 80 89 91 92 93 94 95 99 99 104 105 110 86 90 91 92 93 95 97 99 99 105 105 94)
y = (39 76 71 40 43 69 84 38 41 30 22 38 37 90 42 39 43 51 79 33 33 50 30 32 24 59 31 70 67 59 76 30 27 23 41 35 53)

number of data	$n = 37$
mean of x	$\overline{x} = 95.46$
mean of x	$\overline{y} = 47.19$
variance in x	$s_x = 7.01$
variance in y	$s_y = 18.64$
covariance	$s_{xy} = -70.19$
Gradient	$m = -1.43$
Correlation	$r = -0.537$

7.3 More than two dimensions

Suppose we have data in N dimensions, and a linear model

$$x_N = a_0 + a_1 x_1 + \ldots a_{n-1} x_{N-1}$$

Taking the mean as the origin of coordinates as before, the *line of best fit* is the line through the origin $\mathbf{u}(\lambda) = \lambda\hat{\mathbf{u}}$, that minimises the sum of squares of perpendicular distances from each datum $\mathbf{x}_i = (x_{1,i}, x_{2,i}, \ldots x_{N,i})$ to this line.

Let the unit vector for this line be

$$\hat{\mathbf{u}} = u_1\mathbf{e}_1 + u_2\mathbf{e}_2 + \ldots u_N\mathbf{e}_N$$

where $\{\mathbf{e}_k\}$ are the orthonormal basis vectors of the coordinate systm. Then the perpendicular distance of $\mathbf{x}_i$ to $\mathbf{u}$ is given by

$$x^2_{\perp,i} = |\mathbf{x}_i|^2 - (\mathbf{x}_i \cdot \hat{\mathbf{u}})^2$$

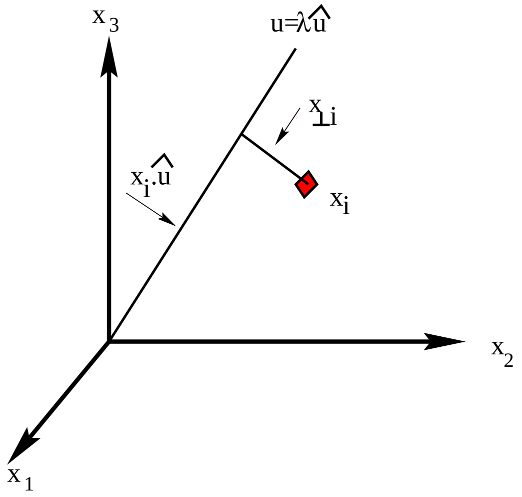
where $|\mathbf{x}_i|$ is the length of $\mathbf{x}_i$. Since this length does not depend on the orientation of $\hat{\mathbf{u}}$, minimisation of $\sum_i x^2_{\perp,i}$ is equivalent to *maximisation* of $\sum_i (\mathbf{x}_i \cdot \hat{\mathbf{u}})^2$.

Suppose that we choose a coordinate system so that $\mathbf{e}_1 = \hat{\mathbf{u}}$. Then the quantity

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i \cdot \hat{\mathbf{u}})^2$$

is the variance of the first coordinate of the data, whereas the quantity

$$S_T^2 = s_1^2 + s_2^2 + \ldots s_N^2 = \frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i|^2$$



which is *invariant to the choice of* $\hat{\mathbf{u}}$ is the *Total Variance* of the data around the origin.

We can interpret the choice of the line of best fit as the choice of a coordinate system so that the variance around this line represents the maximum part of the total variance. This part s_1^2 is given by

$$\begin{pmatrix} u_1 & u2 & \dots & u_N \end{pmatrix} \begin{pmatrix} s_1^2 & s_{1,2}^2 & \dots & s_{1,N}^2 \\ s_{1,2}^2 & s_2^2 & \dots & s_{2,N}^2 \\ \vdots & \vdots & \dots & \vdots \\ s_{1,N}^2 & s_{2,1}^2 & \dots & s_N^2 \end{pmatrix} \begin{pmatrix} u_1 \\ u2 \\ \vdots \\ u_N \end{pmatrix} =: \hat{\mathbf{u}}^{\mathrm{T}} \mathsf{S} \hat{\mathbf{u}}$$

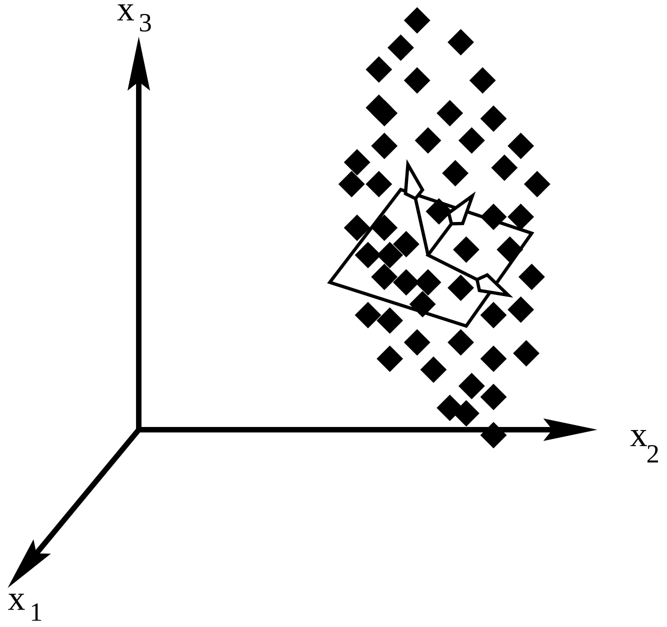
If we chose a coordinate system to be the eigenvectors of S we can diagonalise the system

$$\mathsf{E}^{\mathrm{T}}\mathsf{S}\mathsf{E} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \lambda_N \end{pmatrix}$$

The first eigenvalue $\lambda_1 = s_1^2$ is the principal part of the variance, the next $\lambda_2 = s_2^2$ gives the second most significant part of the variance and so on.

Frequently, the higher eigenvalues fall to zero. or nearly zero, after a certain number of terms, which indicates that there are only a limited number of independent components in the data.

This technique is called *Principal Component Analysis* and the eigenvectors themeselves are called the *Principal Components.*



Note Although a popular model, it is important to keep in mind that this is a *linear* model and assumes *Gaussian Statistics.*

8 Markov Chains

A *Markov Process* is one where the result of a trial depends at most on the result of the previous trial. A *Markov Chain* is a sequence of Markov trials.

Example There are 8 blue and 4 green balls in a bag and one blue one in my hand. A trial consists of taking a random ball from the bag and replacing it with the one in my hand. This is repeated indefinitely. Let $\boldsymbol{X} \in \{B, G\}$ be the result of the drawn ball, and let $H \in \{B, G\}$ be the ball in hand. Then

$$\begin{aligned} Pr(\boldsymbol{X} = B|H = B) &= \frac{2}{3} \quad & Pr(\boldsymbol{X} = G|H = B) &= \frac{1}{3} \\ Pr(\boldsymbol{X} = B|H = G) &= \frac{3}{4} \quad & Pr(\boldsymbol{X} = G|H = G) &= \frac{1}{4} \end{aligned}$$

8.1 Transition Matrices

Starting with the initial case $H = B$ the first result is

$$Pr(\boldsymbol{X}_1 = B) = \frac{2}{3}, \quad Pr(\boldsymbol{X}_1 = G) = \frac{1}{3}$$

The second result is

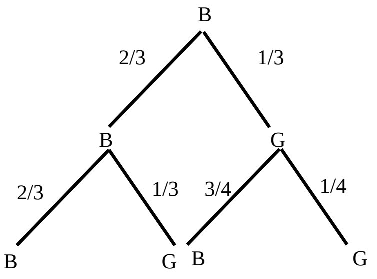
$$Pr(\boldsymbol{X}_2 = B) = \frac{2}{3} \cdot \frac{2}{3} + \frac{3}{4} \cdot \frac{1}{3}, \quad Pr(\boldsymbol{X}_2 = G) = \frac{1}{3} \cdot \frac{2}{3} + \frac{1}{4} \cdot \frac{1}{3}$$

Define b_n as $Pr(n^{th} \text{ result } = B)$ and g_n as $Pr(n^{th} \text{ result } = G)$ then we can write

$$\begin{pmatrix} b_{n+1} \\ g_{n+1} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & \frac{3}{4} \\ \frac{1}{3} & \frac{1}{4} \end{pmatrix} \begin{pmatrix} b_n \\ g_n \end{pmatrix} \rightarrow \quad \mathbf{u}_{n+1} = \mathsf{P}\mathbf{u}_n$$

Vector $\mathbf{u}_n$ is called a *probability vector* whose components are non-negative and sum to unity

Matrix P is called a *transition matrix* each of whose columns have components that are non-negative and sum to unity. A matrix with such properties is called a *probability matrix* or *stochastic matrix.* The term *transition matrix* arises if we think of the outcomes of the Markov process as one of a possible set of *states* (in this case B or G) I.e



From:		B	G
To :	B	$\frac{2}{3}$	$\frac{3}{4}$
	G	$\frac{1}{3}$	$\frac{1}{4}$

Example A factory produces computer chips, some of which are faulty. When it has produced a satisfactory chip there is a 90% chance the next one is satisfactory, but if it produces a faulty chip, there is a 20% chance that the next one is faulty.

In this case the transition matrix is

From:		S	F
To :	S	0.9	0.8
	F	0.1	0.2

8.2 Properties of Stochastic Matrices

Let P be a stochastic matrix and let $\mathbf{c} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Since the columns of P sum to 1 we have

$$\mathbf{c}^{\mathrm{T}}\mathsf{P} = \mathbf{c}^{\mathrm{T}}$$

conversely, a 2 by 2 matrix with non-negative entries, whose columns sum to 1, and for which $\mathbf{c}^{\mathrm{T}}\mathsf{P} = \mathbf{c}^{\mathrm{T}}$ is a stochastic matrix. It follows that if P and Q are stochastic

$$\mathbf{c}^{\mathrm{T}}\mathsf{PQ} = \mathbf{c}^{\mathrm{T}}\mathsf{Q} = \mathbf{c}^{\mathrm{T}}$$

so that the product PQ is also stochastic.

Taking $\mathsf{Q} = \mathsf{P}$ and using induction, it follows that P^n is stochastic, and we have

$$\mathbf{u}_n = \mathsf{P}^n \mathbf{u}_0$$

We can write a general 2 by 2 stochastic matrix

$$\mathsf{P} = \begin{pmatrix} a & 1-b \\ 1-a & b \end{pmatrix}$$

where $0 \leq a \leq 1, 0 \leq b \leq 1$. We find the following results for the eigendecomposition of P

1. One eigenvector of P has eigenvalue 1, and has non-negative components :

$$\begin{pmatrix} a & 1-b \\ 1-a & b \end{pmatrix} \begin{pmatrix} (1-b)x \\ (1-a)x \end{pmatrix} = \begin{pmatrix} (1-b)x \\ (1-a)x \end{pmatrix}$$

2. The other eigenvector of P has eigenvalue $-1 \leq |\lambda| \leq 1$ and is in the direction $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$:

$$\begin{pmatrix} a & 1-b \\ 1-a & b \end{pmatrix} \begin{pmatrix} -x \\ x \end{pmatrix} = (a+b-1) \begin{pmatrix} -x \\ x \end{pmatrix}$$

8.3 Limiting Behaviour and Steady State Vectors

Consider the limiting behaviour of the matrices P^n. Let E by the matrix of normalised eigenvectors of P. Since E is full rank it posseses an inverse E^{-1}

$$\mathsf{E} = \begin{pmatrix} \alpha & -1 \\ \beta & 1 \end{pmatrix}, \quad \mathsf{E}^{-1} = K \begin{pmatrix} 1 & 1 \\ -\beta & \alpha \end{pmatrix}$$

where $\alpha = \frac{1-b}{2-(a+b)}, \beta = \frac{1-a}{2-(a+b)}$ and K is a constant. We have

$$\begin{aligned} \mathsf{P} &= \mathsf{E}\Lambda\mathsf{E}^{-1} = \mathsf{E}\begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix}\mathsf{E}^{-1} \quad \rightarrow \mathsf{PE} = \mathsf{E}\Lambda \\ \Rightarrow \quad \mathsf{P}^n &= \mathsf{E}\Lambda^n\mathsf{E}^{-1} = \mathsf{E}\begin{pmatrix} 1 & 0 \\ 0 & \lambda^n \end{pmatrix}\mathsf{E}^{-1} \\ \Rightarrow \quad \mathsf{P}^\infty &= \lim_{n\rightarrow\infty} \mathsf{P}^n = \mathsf{E}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\mathsf{E}^{-1} = K\begin{pmatrix} \alpha \\ \beta \end{pmatrix}\begin{pmatrix} 1 & 1 \end{pmatrix} \end{aligned}$$

This implies that *regardless of the initial probability vector* $\mathbf{u}_0$, the limit of a Markov chain is

$$\mathbf{u}_\infty = \mathsf{P}^\infty\mathbf{u}_0 = \frac{1}{2-(a+b)}\begin{pmatrix} 1-b \\ 1-a \end{pmatrix}$$

Returning to the example of blue and green balls, we get

$$\mathbf{u}_\infty = \begin{pmatrix} \frac{9}{13} \\ \frac{4}{13} \end{pmatrix}$$

Notice that these are the probabilities of picking out of a mixture of 9 blue and 4 green balls. I.e. the limiting state has the probability of the total number of balls in the system, irrespective of whether we began with a blue or green ball in hand

8.4 Markov chains with more than two states

Whatever the number of states of the system, any stochastic matrix has at least one eigenvalue with value 1.

If there is only one such eigenvalue, then the Markov chain corresponding to this system has a limiting vector which is the corresponding eigenvector.

If there is more than one eigenvalue equal to 1, then the system can have more than one limiting vector.

Example

$$\mathsf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & 0 & 0 \\ \frac{1}{2} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & \frac{2}{5} & \frac{1}{3} \\ 0 & 0 & \frac{3}{5} & \frac{2}{3} \end{pmatrix} \rightarrow \mathbf{u}_\infty = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ 0 \\ 0 \end{pmatrix} \; or \; \begin{pmatrix} 0 \\ 0 \\ \frac{5}{14} \\ \frac{9}{14} \end{pmatrix}$$

8.5 Transient and Closed Classes

A four state system $\{A, B, C, D\}$ is defined by

$$\mathsf{P} = \begin{pmatrix} \frac{1}{2} & \frac{2}{3} & 0 & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{3}{4} & 0 \end{pmatrix} \rightarrow \mathbf{u}_{\infty} = \begin{pmatrix} \frac{4}{7} \\ \frac{3}{7} \\ 0 \\ 0 \end{pmatrix}$$

In the limiting behaviour, the last two states, have zero probability. This can be seen to be the case if the transition matrix is considered as a directed graph

The set $L = \{A, B\}$ is called a *closed class* and $M = \{C, D\}$ is a *transient class.* The state can migrate from $M \rightarrow L$ but not vice-versa.

8.6 Random Walks in One Dimension

Suppose there are n points on a line and states $A_1, A_2, \ldots A_n$ correspond to a body being found at the point $1, 2, \ldots n$. At a transition, the body can travel forward, backward, or stay still with probabilities f_i, b_i, s_i except at the endpoints. The transition matrix will look like

$$\mathsf{P} = \begin{pmatrix} s_1 & f_1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ b_2 & s_2 & f_2 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 0 & b_n & s_n \end{pmatrix}$$

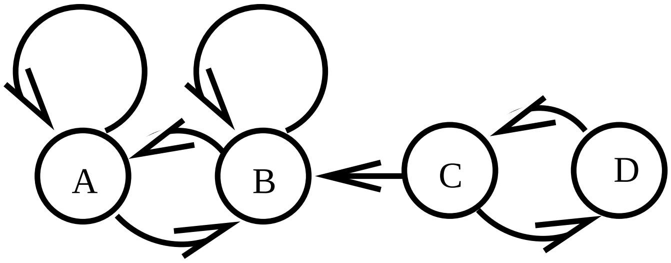
The system is called a *random walk*

States A_1 and A_n are closed states, and the others are transient. The limiting vectors are

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad or \quad \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

8.7 Expectation for Markov Processes

Example The weather over two years at one weather station was recorded and classified as



"Fine", "Dull", or "Wet". The relative frequencies, as well as the transitions were

$$\mathbf{m} = \begin{pmatrix} 0.354 \\ 0.331 \\ 0.315 \end{pmatrix}, \quad \mathsf{P} = \begin{pmatrix} 0.529 & 0.293 & 0.222 \\ 0.274 & 0.467 & 0.252 \\ 0.197 & 0.240 & 0.526 \end{pmatrix}$$

Note that $\mathbf{m}$ is close to the eigenvector of P with eigenvalue 1.

Lengths of various runs of the same type of day :

Days	1	2	3	4	5	6	7 or more	
Fine	69	30	17	7	4	3	2	(9,12)
Dull	77	29	14	7	1	2	3	(7,11,16)
Wet	61	18	10	11	5	2	2	(7,10)

Bernouilli Model : $p = Pr(\text{day being fine})$ is independent of previous days, so

$$Pr(\text{run of length } k \text{ days}) = (1-p)p^k(1-p)$$

which gives the expectation table

Days	1	2	3	4	5	6	7 or more
Fine	108	38	14	5	2	1	0
Dull	108	36	12	4	1	0	0
Wet	108	34	11	3	1	0	0

Markov model : $a = Pr(\text{fine day follows a fine day})$

$$Pr(\text{run of length } k \text{ days}) = p(1-p)a^{k-1}(1-p)$$

which gives

Days	1	2	3	4	5	6	7 or more
Fine	58	30	16	9	4	2	1
Dull	69	32	15	7	3	2	1
Wet	52	27	14	8	4	2	1

Which is a better approximation to the actual figures

8.8 Exercises

1. A roulette wheel has spaces that are alternately red and black. The ball starts in one of the spaces and when the wheel is spun it moves round before coming to rest again in one of the spaces. The wheel is biased. It is observed that when it starts in a red space it ends in a red space 40% of the time, but when it starts in a black space it ends in a black space 70% of the time.

   a) write down the probability matrix with "red" in the first row and column.
   b) If the initial probability vector is $\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$, find the probability vector for the outcome of the second spin
   c) If the probability vector at a certain stage is $\begin{pmatrix} \frac{3}{8} \\ \frac{5}{8} \end{pmatrix}$, find the probability vector for the previous spin.

2. A general probability matrix P and a probability vector $\boldsymbol{u}$ can be written

$$\mathsf{P} = \begin{pmatrix} 1-a & b \\ a & 1-b \end{pmatrix}, \quad \boldsymbol{u} = \begin{pmatrix} u \\ v \end{pmatrix}$$

   Show that the vector $\mathsf{P}\boldsymbol{u}$ has components whose sum is unity.

3. P_1, P_2, P_3, are three probability matrices :

$$\mathsf{P}_1 = \begin{pmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 1 \end{pmatrix} \quad \mathsf{P}_2 = \begin{pmatrix} \frac{1}{6} & \frac{3}{4} \\ \frac{5}{6} & \frac{1}{4} \end{pmatrix} \quad \mathsf{P}_3 = \begin{pmatrix} \frac{1}{5} & \frac{3}{5} \\ \frac{4}{5} & \frac{2}{5} \end{pmatrix}$$

   a) Find the products of any pair of matrices and show that the result is a probability matrix.
   b) Show that P_1^3 and P_3^3 are probability matrices.
   c) Find the eigenvalues and eigenvectors of P_1, P_2, P_3
   d) Show that the product of the eigenvalues is equal to the determinant of the matrix in each case.

4. If $\mathsf{P} = \begin{pmatrix} 1-a & b \\ a & 1-b \end{pmatrix}$, and $\mathsf{Q} = \begin{pmatrix} 1-c & d \\ c & 1-d \end{pmatrix}$ are stochastic (with $0 \leq a \leq 1, 0 \leq b \leq 1, 0 \leq c \leq 1, 0 \leq d \leq 1$) show by direct multiplication that PQ is stochastic.

5. if the two eigenvalues of a 2×2 stochastic matrix are equal, what is the form of the matrix ? What are its eigenvectors ?

6. For the three matrices in question 3, verify the eigendecomposition $\mathsf{P} = \mathsf{E}\Lambda\mathsf{E}^{-1}$, where E has the eigenvectors as columns, and Λ is diagonal with the eigenvalues as the diagonal entries. Thus find the form of P^{∞}.

7. Find the eigenvalues and eigenvectors of the matrix $\mathsf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and show that P^n does not tend to a limit as $n \to \infty$.

8. Two bags contain two black and two white balls respectively. A transition consists of interchanging a ball drawn at random from each bag.
   a) Identify the three possible states of the system, and write down the 3×3 transition matrix for the system.
   b) What is the probability that the first bag contains one ball of each colour after three transitions?
   c) Find the eigenvalues and eigenvectors of this matrix and determine the limiting probabilty vector for the system.

9. The probability of a team winning a match is 0.6 and of drawing it is 0.3 if the previous match was won. if the previous match was drawn the corresponding probabilities are 0.2 and 0.6 and if it was lsot 0.2 and 0.4. Find the transition matrix and hence the probability of winning or drawing any particular match in the distant future if the probabilities remain the same.

8.9 Answers

1. a) $\begin{pmatrix} 0.4 & 0.3 \\ 0.6 & 0.7 \end{pmatrix}$

   b) $\begin{pmatrix} 0.35 \\ 0.65 \end{pmatrix}$

   c) $\begin{pmatrix} \frac{3}{4} \\ \frac{1}{4} \end{pmatrix}$

2. $\mathsf{P}\boldsymbol{u} = \begin{pmatrix} (1-a)u + bv \\ au + (1-b)v \end{pmatrix}$.
   Sum of column is $u + v = 1$ (by initial definition of $\boldsymbol{u}$ as a probability vector).

3. a) $\mathsf{P}_1\mathsf{P}_2 = \frac{1}{24}\begin{pmatrix} 2 & 9 \\ 22 & 15 \end{pmatrix}$, $\mathsf{P}_1\mathsf{P}_3 = \frac{1}{10}\begin{pmatrix} 1 & 3 \\ 9 & 7 \end{pmatrix}$, etc ... .
   b) $\mathsf{P}_1^3 = \frac{1}{8}\begin{pmatrix} 1 & 0 \\ 7 & 8 \end{pmatrix}$, $\mathsf{P}_1^3 = \begin{pmatrix} 0.392 & 0.456 \\ 0.608 & 0.544 \end{pmatrix}$
   c) $\mathsf{P}_1 : 1, \frac{1}{2}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$; $\mathsf{P}_2 : 1, -\frac{7}{12}, \frac{1}{19}\begin{pmatrix} 9 \\ 10 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$; $\mathsf{P}_3 : 1, -\frac{2}{5}, \frac{1}{7}\begin{pmatrix} 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
   d) three determinants are $\frac{1}{2}$, $-\frac{7}{12}$, $-\frac{2}{5}$.

4. $\mathsf{PQ} = \begin{pmatrix} (1-a)(1-c) + bc & (1-a)d + b(1-d) \\ a(1-c) + (1-b)c & ad + (1-b)(1-d) \end{pmatrix}$
   All entries are between 0 and 1, and columns sum to 1.

5. the matrix is the identity matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Any vector is an eigenvector (the eigenvalues are both 1).

6. $\mathsf{P}_1 : \mathsf{E} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}$, $\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$, $\mathsf{E}^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, $\mathsf{P}_1^\infty = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$;
   $\mathsf{P}_2 : \mathsf{E} = \frac{1}{19}\begin{pmatrix} 9 & 1 \\ 10 & -1 \end{pmatrix}$, $\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & -\frac{7}{12} \end{pmatrix}$, $\mathsf{E}^{-1} = \begin{pmatrix} 1 & 1 \\ 10 & -9 \end{pmatrix}$, $\mathsf{P}_2^\infty = \frac{1}{19}\begin{pmatrix} 9 & 9 \\ 10 & 10 \end{pmatrix}$;
   $\mathsf{P}_3 : \mathsf{E} = \frac{1}{7}\begin{pmatrix} 3 & 1 \\ 4 & -1 \end{pmatrix}$, $\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & -\frac{2}{5} \end{pmatrix}$, $\mathsf{E}^{-1} = \begin{pmatrix} 1 & 1 \\ 4 & -3 \end{pmatrix}$, $\mathsf{P}_3^\infty = \frac{1}{7}\begin{pmatrix} 3 & 3 \\ 4 & 4 \end{pmatrix}$

7. $1, -1, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. There is no limit since $\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and the multiples of this alternate between 1 and -1 on the lower right.

8. a) Three states are

   $$1 : (B, B) + (W, W)\,; \quad 2 : (B, W) + (B, W)\,; \quad 3 : (W, W) + (B, B)\,.$$

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

   b) Initial state is $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. The transistions are $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \end{pmatrix} \rightarrow \begin{pmatrix} \frac{1}{8} \\ \frac{3}{4} \\ \frac{1}{8} \end{pmatrix}$. So the

answer is $\frac{3}{4}$ (i.e. probability of being in state 2 after three transitions).
c) Eigensystem is

$$\lambda_1 = 1, \boldsymbol{u}_1 = \frac{1}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} \qquad \lambda_2 = -0.5, \boldsymbol{u}_2 = \frac{1}{5}\begin{pmatrix}2\\-1\\2\end{pmatrix} \qquad \lambda_3 = 0, \boldsymbol{u}_3 = \frac{1}{2}\begin{pmatrix}1\\0\\-1\end{pmatrix}.$$

Limiting vector is the one corresponding to λ_1, i.e. $\boldsymbol{u}_1 = \frac{1}{3}\begin{pmatrix}1\\1\\1\end{pmatrix}$.

9. $\mathsf{P} = \begin{pmatrix}0.6 & 0.2 & 0.2\\ 0.3 & 0.6 & 0.4\\ 0.1 & 0.2 & 0.4\end{pmatrix}$. Limit $\boldsymbol{u}_\infty = \frac{1}{24}\begin{pmatrix}8\\11\\5\end{pmatrix}$