Foundations of Mathematics

Lecture notes for the MSc in Cognitive Systems at the University of Potsdam, Germany

Compiled by Shravan Vasishth and Lena Jäger

version of September 14, 2015

Contents

Chapter 1

Preface

These notes are intended to serve as lecture notes for the MSc Cognitive Systems foundational course *Foundations of Mathematics*. The notes are based on the books listed in the references, and on the graduate certificate course taught at the School of Mathematics and Statistics (SOMAS), University of Sheffield, UK.

An important piece of advice we have for students taking this course is that when you get stuck with a problem, don't just give up. Sometimes you will have to keep at it for some hours or days (intense effort, trying different lines of attack) to solve a problem. One trick that works is to put in intense effort, and then take a break (usually involved going to sleep at night—SV often works on math problems last thing at night). When you return to the problem the next time round, you may see the solution right away.

Homework assignments are integral to the course, but these are provided only to students taking the course. Solutions are usually provided a week after the assignment is handed out. The homework assignments will be graded; the student is also expected to evaluate their mistakes on their own by comparing the provided solutions with their own attempt. The solutions to assignments will be discussed in class, so that there is ample opportunity for discussion of any problems that come up.

The formal examination for this course is a 20 minute oral exam that the instructor will conduct at the end of the course, but will also factor in the grades in the homework assignments when determining the final grade.

The offical textbook for the course is [3]; **please get the second edition, published in 2002**. A good, additional text (if you can afford it) is [5]; **if you buy this too, please get the fourth edition, published in 2008**.

SV owes a great debt of gratitude to Fionntan Roukema at SOMAS, Sheffield.

Chapter 2

Course schedule

We have approximately 14 lectures in this course (this varies a bit from year to year).

The approximate schedule is as follows:

1. Precalculus I (HW0 assigned)
2. Discussion of solutions to HW0
3. Precalculus II (HW1 assigned)
4. Solutions HW1
5. Differentiation (HW2 assigned)
6. Solutions HW2
7. Integration (HW3 assigned)
8. Solutions HW3
9. Matrix Algebra I (HW4 assigned)
10. Matrix Algebra II, minima, maxima, partial derivatives (HW 5 assigned)
11. Solutions HW4 and HW5
12. Double integrals, change of variables (HW 6)
13. Solutions HW6
14. Review, and applications in statistics and data mining (time permitting)

Chapter 3

Pre-calculus review

Sources: heavily depended on [8], [10], [3] (the official textbook in the course) and various summaries on the internet on trigonometric functions.

3.1 Counting

The number of ways in which one may select an unordered sample of k subjects from a population that has n distinguishable members is

- $\frac{(n-1+k)!}{[(n-1)!k!]}$ if sampling is done with replacement,
- $\binom{n}{k} = \frac{n!}{[k!(n-k)!]}$ if sampling is done without replacement.

Table 3.1: default

	ordered = TRUE	ordered = FALSE
replace = TRUE	n^k	$(n-1+k)!/[(n-1)!k!]$
replace = FALSE	$n!/(n-k)!$	$\binom{n}{k}$

3.2 Permutations and combinations

3.2.1 Permutations

For n objects, of which $n_1, \ldots, n_r$ are alike, the number of different permutations are

$$\frac{n!}{n_1!n_2!\ldots n_r!} \tag{3.1}$$

3.2.2 Combinations

Choosing k distinct objects from n, when order irrelevant:

$$\binom{n}{k} = \frac{n!}{(n-r)!r!} \tag{3.2}$$

3.2.3 Binomial theorem

$$(x+y)^n = \sum_{n=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{3.3}$$

3.3 Inequalities

To solve an inequality, we need to determine the numbers x that satisfy the inequality. This is called the **solution set** of the inequality.

1. If we multiply or divide an inequality by a negative number, the inequality is reversed.

   Example: $-\frac{1}{2}x < 4$

2. To solve a quadratic inequality, you can always solve it by completing the square. Another way is to factor the quadratic (works sometimes).

   Example: Solve $x^2 - 4x + 3 > 0$.

   Example (use completing the square): Solve $x^2 - 2x + 5 \leq 0$.

3.4 Series

3.4.1 Arithmetic series

General form:

$$a + (a + d) + (a + 2d) + \dots \tag{3.4}$$

k-th partial sum for **arithmetic series**:

$$S_k = \sum_{n=1}^{k} (a + (n-1)d) \tag{3.5}$$

The sum can be found by:

$$S_k = \frac{k}{2}(2a + (k-1)d) \tag{3.6}$$

3.4.2 Geometric series

General form:

$$a + ar + ar^2 \dots \tag{3.7}$$

In summation notation:

$$\sum_{n=1}^{\infty} ar^{n-1} \tag{3.8}$$

k-th partial sum:

$$S_k = \frac{a - (1 - r^k)}{1 - r} \tag{3.9}$$

S_∞ exists just in case $\mid r \mid < 1$.

$$S_\infty = \frac{a}{1 - r} \tag{3.10}$$

3.4.3 Power series

$$\sum_{n=0}^{\infty} a_n (x - a)^n \tag{3.11}$$

3.5 Trigonometry

3.5.1 Basic definitions

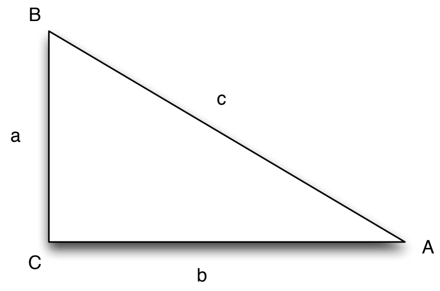


Figure 3.1: Right-angled triangle.

$$\sin A = \frac{opp}{hyp} = \frac{a}{c} \tag{3.12}$$

Cosine is the complement of the sine:

$$\cos A = \sin(90 - A) = \sin B \tag{3.13}$$

$$\cos A = \frac{b}{c} \tag{3.14}$$

3.5.2 Pythagorean identity

$$a^2 + b^2 = c^2 \tag{3.15}$$

$$\begin{gathered} \frac{a^2}{c^2} + \frac{b^2}{c^2} = 1 \\ \sin^2 A + \cos^2 A = 1 \end{gathered} \tag{3.16}$$

3.5.3 Relations between trig functions

$$\tan A = \frac{\sin A}{\cos A} = \frac{a}{c} / \frac{b}{c} = \frac{a}{b} = \frac{opp}{adj} \tag{3.17}$$

$\tan A$ is also the **slope** of a line.

$$\cot A = \frac{1}{\tan A} = \frac{\cos A}{\sin A} \tag{3.18}$$

$$\sec A = \frac{1}{\cos A} \tag{3.19}$$

$$\csc A = \frac{1}{\sin A} \tag{3.20}$$

sin A = a/c (opp/hyp)	csc A = c/a (hyp/opp)
cos A = b/c (adj/hyp)	sec A = c/b (hyp/adj)
tan A = a/b (opp/adj)	cot A = b/a (adj/opp)

Note that cot A = tan B, and csc A = sec B.

3.5.4 Identities expressing trig functions in terms of their complements

cos t = sin(π/2-t)	sin t = cos(π/2-t)
cot t = tan(π/2-t)	tan t = cot(π/2-t)
csc t = sec(π/2-t)	sec t = csc(π/2-t)

3.5.5 Periodicity

$\sin t + 2\pi = \sin t$	$\sin t + \pi = -\sin t$
$\cos t + 2\pi = \cos t$	$\cos t + \pi = -\cos t$
$\tan t + 2\pi = \tan t$	$\tan t + \pi = \tan t$

$\sin 0 = 0$	$\cos 0 = 1$	$\tan 0 = 0$
$\sin \frac{\pi}{2} = 1$	$\cos \frac{\pi}{2} = 0$	$\tan \frac{\pi}{2}$ undefined
$\sin \pi = 0$	$\cos \pi = -1$	$\tan \pi = -1$

3.5.6 Law of cosines

Three ways of writing it:

$$c^2 = a^2 + b^2 - 2ab\cos C \tag{3.21}$$

$$a^2 = b^2 + c^2 - 2bc\cos C \tag{3.22}$$

$$b^2 = c^2 + a^2 - 2ca\cos C \tag{3.23}$$

3.5.7 Law of sines

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c} \tag{3.24}$$

3.5.8 Odd and even functions

A function f is said to be an odd function if for any number x, $f(-x) = -f(x)$ (e.g., $f(y) = x^5$). A function f is said to be an even function if for any number x, $f(-x) = f(x)$ (e.g., $f(y) = x^4$).

Odd functions: sin, tan, cotan, csc.

Even functions: cos, sec.

3.5.9 Sum formulas for sine and cosine

$$\sin(s+t) = \sin s\cos t + \cos s\sin t \tag{3.25}$$

$$\cos(s+t) = \cos s\cos t - \sin s\sin t \tag{3.26}$$

3.5.10 Double angle formulas for sine and cosine

$$\sin 2t = 2\sin t\cos t \tag{3.27}$$

$$\cos 2t = \cos^2 t - \sin^2 t = 2\cos^2 t - 1 = 1 - 2\sin^2 t \tag{3.28}$$

3.5.11 Less important identities

Pythagorean formula for tan and sec:

$$\sec^2 t = 1 + \tan^2 t \tag{3.29}$$

Identities expressing trig functions in terms of their supplements

$$\sin(\pi - t) = \sin t \tag{3.30}$$

$$\cos(\pi - t) = -\cos t \tag{3.31}$$

$$\tan(\pi - t) = -\tan t \tag{3.32}$$

Difference formulas for sine and cosine

$$\sin(s - t) = \sin s \cos t - \cos s \sin t \tag{3.33}$$

$$\cos(s - t) = \cos s \cos t + \sin s \sin t \tag{3.34}$$

Chapter 4

Differentiation and integration

4.1 Differentiation

Given a function $f(x)$, the derivative from first principles is as follows:

If we want to find $\frac{\partial y}{\partial x}$, note that $\partial y = f(x+\partial x) - f(x)$:

$$\begin{aligned} y =& f(x) \\ y+\partial y =& f(x+\partial x) \end{aligned} \tag{4.1}$$

Subtracting y from both sides:

$$\begin{aligned} y+\partial y - y =& f(x+\partial x) - y \\ \partial y =& f(x+\partial x) - f(x) \end{aligned} \tag{4.2}$$

It follows that $\frac{\partial y}{\partial x} = \frac{f(x+\partial x)-f(x)}{\partial x}$. If we write the first derivative $\frac{\partial y}{\partial x}$ as $f^{(1)}(x)$, we have the following identity:

$$f^{(1)}(x) = \frac{f(x+\partial x) - f(x)}{\partial x} \tag{4.3}$$

Other notations we will use interchangeably: Given y = f(x),
$\frac{\partial y}{\partial x} = \frac{dy}{dx} = f^{(1)}(x) = y'$.
Note:

1. The derivative of a function at some point c is the slope of the function at that point c. The slope is the rate of growth: $\frac{dy}{dx}$.

2. The derivative f' is itself a function, and can be further differentiated. So, the second derivative, f'' is the rate of change of the slope. This will soon become a very important fact for us when we try to find maxima and minima of a function.

4.1.1 Deriving a rule for differentiation

Consider the function $y = x^2$. Suppose we increase x by a small amount dx, y will also increase by some small amount dy. We can ask: what is the ratio of the increases: $\frac{dy}{dx}$? We will derive this next:

$$y + dy = (x + dx)^2 \tag{4.4}$$

Expanding out the RHS:

$$y + dy = x^2 + dx^2 + 2xdx \tag{4.5}$$

Observe that squaring a small quantity dx will make it even smaller (e.g., try squaring 1/100000000; it is effectively zero). That leads to the following simplification:

$$y + dy = x^2 + 2xdx \text{ as dx gets infinitesimally small} \tag{4.6}$$

Subtracting y from both sides:

$$dy = 2xdx \Leftrightarrow \frac{dy}{dx} = 2x \tag{4.7}$$

Exercise: Find the derivative of x^3, x^4, x^5 using the above approach. Evaluate each derivative at c=2.

The general rule is that

$$\boxed{\frac{dy}{dx} = nx^{n-1}} \tag{4.8}$$

Verify that this rule works for negative powers and fractional powers: $x^{-2}, x^{1/2}$.

4.1.2 Derivatives of trigonometric functions, exponential, and log

Memorizing these results will simplify our life considerably.

1. $\frac{d(\sin(x))}{dx} = \cos x$ and $\frac{d(\cos(x))}{dx} = -\sin x$.
2. $d(\exp(x))/dx = \exp(x)$
3. $d(\log(x))/dx = \frac{1}{x}$

Proofs will come later.

4.1.3 Derivations of combinations of functions

Let u and v be (differentiable) functions.

1. Sum of functions
$$(u+v)' = u' + v' \tag{4.9}$$
2. Difference of functions
$$(u-v)' = u' - v' \tag{4.10}$$
3. Function multiplied by constant:

   Given a constant c:
$$(cu)' = cu' \tag{4.11}$$
4. Product of functions
$$(uv)' = uv' + vu' \tag{4.12}$$
5. Quotient of functions:
$$(u/v)' = \frac{vu' - uv'}{v^2} \tag{4.13}$$
6. Chain rule:

   If y = g(f(x)), then, letting u=f(x), we get $dy/dx = dy/du \cdot du/dx$

   Example: $y = (x^2+3)^7$ needs the chain rule.

Note: the notation $\left.\frac{dy}{dx}\right|_{x=a}$ means: evaluate the derivative at x=a.

Given one of our standard functions above, which I will call $f(x)$, we can differentiate it repeatedly: f', f'', etc. So:

1. $f'(x) = \frac{d}{dx} f(x)$

2. $f''(x) = \frac{d^2}{dx^2} f(x)$

3. $f'''(x) = \frac{d^3}{dx^3} f(x)$

4.1.4 Maxima and minima

If we have a function like $y = f(x)$, there may be a point (for some x) where the graph of this function turns over. For example, in the normal distribution (see chapterr̃efapps), the graph turns over at the mean.

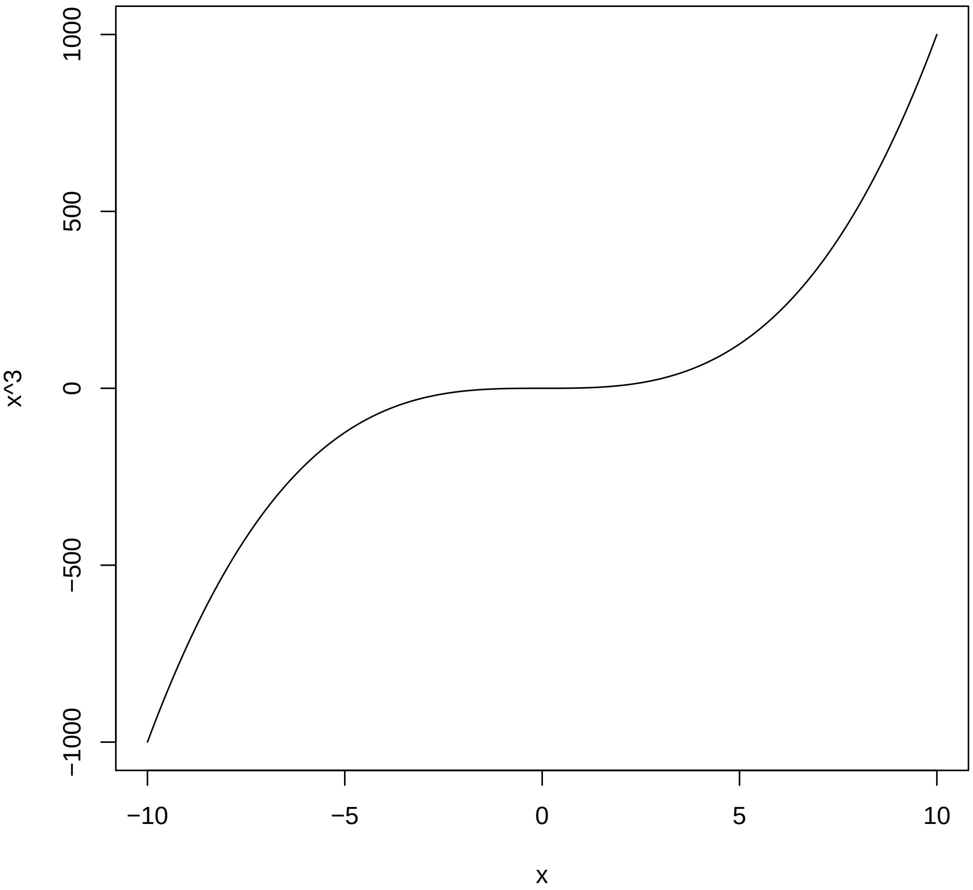
At this turning point, the slope changes from positive to negative, and is 0 at the turning point. Therefore,

$$f'(x) \tag{4.14}$$

is an equation whose solution is the point where the graph turns over.

Note that a graph does not necessarily turn over at a point where $f'(x) = 0$. Example: $y = x^3$.

```r
x<-seq(-10,10,by=0.1)
plot(x,x^3,type="l")
```



If $f'(x) = 0$ at some point x=c, then this point c is called the stationary point of $f(x)$. This point could be a local maximum or a local minimum. To determine whether this point is a local maximum or minimum, take the second derivative: $f''(x)$.

1. If $f'(c) = 0$ and $f''(c) < 0$, then we have a local maximum.
2. If $f'(c) = 0$ and $f''(c) > 0$, then we have a local minimum.
3. If $f'(c) = 0$ and $f''(c) = 0$, then we have either a maximum, mininum, or a stationary point of inflection like in the $y = x^3$ figure above. In this case, examine the sign of $f'(x)$ on both sides of $x = c$.

4.2 Integration

4.2.1 Riemann sums

A simple example:

Given $\phi(x)$, the probability density function of the standard normal distribution:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

We have to find an approximate value of

$$\int_0^1 \phi(x)\, dx$$

We divide the interval $[0,1]$ into 10 intervals of width $1/10$, and approximate the area under the curve by taking the sum of the 10 rectangles under the curve. The width of each rectangle will be $\partial x = 1/10$, and each of the ten x_i are $1/10, 2/10, \dots, 10/10$, i.e., $i/10$, where $i = 1, \dots, 10$.

The area A can be computed by summing up the areas of the ten rectanges. Each rectangle's area is length × width, which is $\phi(x_i) \times \partial x$. Hence,

$$A = \sum_{i=1}^{10} \phi(x_i) \partial x = \sum_{i=1}^{10} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} \times \frac{1}{10}$$

The constant terms $\frac{1}{\sqrt{2\pi}}$ and $\frac{1}{10}$ can be pulled out of the summation:

$$A = \frac{1}{\sqrt{2\pi}} \frac{1}{10} \sum_{i=1}^{10} e^{-x_i^2/2}$$

We use R for the above calculations. First, we define the function for $e^{-x_i^2/2}$:

```
my.fn<-function(x)
    {exp(1)^(-(x^2/2))}
```

Then we define x_i (I made the code very general so that the number of intervals n can be increased arbitrarily) and plug this into the function:

```r
n<-10
x.i<-(1:n)/n
A<- ((1/n) * (1/sqrt(2 * pi)) * sum(my.fn(x.i)))
A<-round(A,digits=5)
```

Compare this to the exact value, computed using R:

```r
fprob2<-function(x){
        (1/sqrt(2 * pi))*exp(1)^(-(x^2/2))
}

integrate(fprob2,lower=0,upper=1)

## 0.3413447 with absolute error < 3.8e-15
```

Answer: The approximate area is: 0.33329. This is a bit lower than the value computed by R, but this is because the ten rectangles fall inside the curve.

```r
## As an aside, note that one can get really close
## to the pnorm value by increasing n,
## say to a high number like 2000:
n<-2000

## 2000 rectangles now:
x.i<-(1:n)/n

(A<- ((1/n) * (1/sqrt(2 * pi)) * sum(my.fn(x.i))))

## [1] 0.3413055
```

With 2000 rectangles, we can get a better estimate of the area than with 10 rectangles: 0.3413. Compare this with the theoretical value:

```r
round(pnorm(1)-pnorm(0),4)

## [1] 0.3413
```

4.2.2 Some common integrals

$$\int \frac{1}{x}\,dx = \log \mid x \mid + c \tag{4.15}$$

$$\int \log x\,dx = \frac{1}{x} + c \tag{4.16}$$

4.2.3 The Fundamental Theorem of Calculus

The Fundamental Theorem states the following:

Let f be a continuous real-valued function defined on a closed interval $[a,b]$. Let F be the function defined, for all x in $[a,b]$, by

$$F(x) = \int_a^x f(u)\,du$$

Then, F is continuous on $[a,b]$, differentiable on the open interval (a,b), and

$$F'(x) = f(x)$$

for all x in (a,b).

4.2.4 The u-substitution

From [8, 306]:

An integral of the form

$$\int f(g(x))g'(x)\,dx \tag{4.17}$$

can be written as

$$\int f(u)\,du \tag{4.18}$$

by setting

$$u = g(x) \tag{4.19}$$

and

$$du = g'(x)\,dx \tag{4.20}$$

If F is an antiderivative for f, then

$$\frac{d}{dx}[F(g(x))] \underset{\substack{\uparrow\\ \text{by the chain rule}}}{=} F'(g(x))g'(x) \underset{\substack{\uparrow\\ F'=f}}{=} f(g(x))g'(x) \tag{4.21}$$

We can obtain the same result by calculating:

$$\int f(u)\,du \tag{4.22}$$

and then substituting g(x) back in for u:

$$\int f(u)\,du = F(u) + C = F(g(x)) + C \tag{4.23}$$

A frequently occurring type of integral is

$$\int \frac{g'(x)}{g(x)}\,dx \tag{4.24}$$

Let $u = g(x)$, giving $\frac{du}{dx} = g'(x)$, i.e., $du = g'(x)\,dx$, so that

$$\int \frac{g'(x)}{g(x)}\,dx = \int \frac{1}{u} du = ln \mid u \mid + C \tag{4.25}$$

Examples:

$$\int \tan x\,dx = \int \frac{1}{\cos x} \sin x\,dx$$

$$\int \frac{2x+b}{x^2+bx+c}\,dx$$

Functions of linear functions: E.g., $\int cos(2x-1)\,dx$. Here, the general form is $\int f(ax+b)\,dx$. We do $u = ax+b$, and then $du = a\,dx$

Using integration by substitution to compute the expectation of a standard normal random variable:
The expectation of the standard normal random variable:

$$E[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-x^2/2}\, dx$$

Let $u = -x^2/2$.
Then, $du/dx = -2x/2 = -x$. I.e., $du = -x\,dx$ or $-du = x\,dx$.
We can rewrite the integral as:

$$E[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{u} x\, dx$$

Replacing $x\,dx$ with $-du$ we get:

$$-\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{u}\, du$$

which yields:

$$-\frac{1}{\sqrt{2\pi}} [e^{u}]_{-\infty}^{\infty}$$

Replacing u with $-x^2/2$ we get:

$$-\frac{1}{\sqrt{2\pi}} [e^{-x^2/2}]_{-\infty}^{\infty} = 0$$

4.2.5 Change of variables (Gamma functions)

We can solve integrals like

$$\int_{0}^{\infty} x^2 e^{-x^5}\, dx \tag{4.26}$$

by restating it as the gamma function:

$$\Gamma(z) = \int_{0}^{\infty} x^{z-1} e^{-x}\, dx \tag{4.27}$$

This can be done by, e.g., letting $y = x^5$, so that $dy/dx = 5x^4$, and therefore $dx = dy/5x^4 = dy/(5 \times y^{4/5})$. This lets us rewrite the above integral in terms of y:

$$\frac{y^{2/5}}{5y^{4/5}}e^{y}\,dy \tag{4.28}$$

This has the form of the gamma function, allowing us to state the integral in terms of the gamma function.

Note that

$$\Gamma(z) = (z-1)\Gamma(z-1) \tag{4.29}$$

R has a function, gamma, that allows us to compute $\Gamma(z)$:

```
gamma(2)

## [1] 1

gamma(10)

## [1] 362880

## this is equal to:
(10-1)*gamma(10-1)

## [1] 362880
```

Chapter 5

Matrix algebra

[Some of this material is based on [5].]

A rectangular array of numbers (real numbers in our case) which obeys certain algebraic rules of operations is a matrix. The main application for us will be in solving systems of linear equations in statistics (and in data mining).

Example of a 2x2 matrix:

```
## a 2x2 matrix:
(m1<-matrix(1:4,2,2))

##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4

## transpose of a matrix:
t(m1)

##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

5.1 Introductory concepts

5.1.1 Basic operations

Matrix addition and subtraction are cell-wise operations:

```
m1+m1

##      [,1] [,2]
## [1,]    2    6
## [2,]    4    8

m1-m1

##      [,1] [,2]
## [1,]    0    0
## [2,]    0    0
```

Matrix multiplication: Think of this simple situation first, where you have a vector of values:

```
(m1<-rnorm(5))

## [1]  0.8259497 -0.9576636  2.0985226 -0.8925887
## [5] -1.2140469
```

If you want to find out the sum of the squares of these values ($\sum\limits_{i=1}^{n} x_i^2$), then you can multiply each value in m1 with itself, and sum up the result:

```
(sum(m1*m1))

## [1] 8.273734
```

Matrix multiplication does exactly the above operation (for arbitrarily large matrices):

```
t(m1)%*%m1

##          [,1]
## [1,] 8.273734
```

Note that two matrices can only be multiplied if they are **conformable**: the number of columns of the first matrix has to be the same as the number of rows of the second matrix.

Scalar matrix multiplication: Multiplying a matrix M with a scalar value just amounts to multiplying the scalar with each cell of the matrix:

```
5*m1

## [1]  4.129749 -4.788318 10.492613 -4.462943
## [5] -6.070235
```

5.1.2 Diagonal matrix and identity matrix

A diagonal matrix is a square matrix that has zeros in its off-diagonals:

```
diag(c(1,2,4))

##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    2    0
## [3,]    0    0    4
```

An identity matrix is a diagonal matrix with 1's along its diagonal:

```
diag(rep(1,3))

##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

For a 3x3 identity matrix, we can write I_3, and for an nxn identity matrix, I_n.

Multiplying an identity matrix with any (conformable) matrix gives that matrix back:

```
(I<-diag(c(1,1)))
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1

(m4<-matrix(c(6,7,8,9),2,2))

##      [,1] [,2]
## [1,]    6    8
## [2,]    7    9

(I%*%m4)

##      [,1] [,2]
## [1,]    6    8
## [2,]    7    9

## the matrix does not have to be square:
(m5<-matrix(c(2,3,6,7,8,9),2,3))

##      [,1] [,2] [,3]
## [1,]    2    6    8
## [2,]    3    7    9

(I%*%m5)

##      [,1] [,2] [,3]
## [1,]    2    6    8
## [2,]    3    7    9
```

```
betterway<-function(n){
  return(diag(rep(1,n)))
}
```

5.1.3 Powers of matrices

If A is a square nxn matrix, then we write AA as A^2, and so on. If A is diagonal, then AA is just the diagonal matrix with the diagonal elements of A squared:

```
m<-diag(c(1,2,3))
##
m%*%m

##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    4    0
## [3,]    0    0    9
```

For all positive integers m, $I_n^m = I_n$.

5.1.4 Inverse of a matrix

If A,B are square matrices, of order nxn, and the following relation is satisfied:

$$AB = BA = I_n \tag{5.1}$$

then, B is the inverse (it is unique) of A. We write $B = A^{-1}$. The inverse is analogous to division and allows us to solve matrix equations: AB=C can be solved by post-multiplying both sides by B^{-1} to get $ABB^{-1} = AI = A = CB^{-1}$.

How to find the inverse? Consider a 2x2 matrix A, and consider the equation:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \tag{5.2}$$

We can write this in compact matrix form: Ax=d. If we multiply A and x, we get a system of linear equations:

$$a_{11}x_1 + a_{12}x_2 = d_1 \tag{5.3}$$

$$a_{21}x_1 + a_{22}x_2 = d_2 \tag{5.4}$$

If we solve these equations, we get (exercise: prove this as homework):

$$x_1 = \frac{a_{22}d_1 - a_{12}d_2}{a_{11}a_{22} - a_{21}a_{12}} \tag{5.5}$$

$$x_2 = \frac{-a_{21}d_1 + a_{11}d_2}{a_{11}a_{22} - a_{21}a_{12}} \tag{5.6}$$

We can now express the solution in matrix form (verify that this is equivalent to the above two equations for x_1 and x_2):

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{a_{11}a_{22} - a_{21}a_{12}} \begin{pmatrix} a_{22}d_1 - a_{12}d_2 \\ -a_{21}d_1 + a_{11}d_2 \end{pmatrix} = Cd \tag{5.7}$$

where

$$C = \frac{1}{detA} \begin{pmatrix} a_{22} - a_{12} \\ -a_{21} + a_{11} \end{pmatrix} \qquad detA = a_{11}a_{22} - a_{21}a_{12} \tag{5.8}$$

Now, if we pre-multiply $Ax = d$ by A^{-1}, we get

$$A^{-1}Ax = Ix = x = A^{-1}d \tag{5.9}$$

So, because $x = A^{-1}d$ and $x = Cd$, it must be the case that $A^{-1} = C$.

> **Note**: Multiplying a matrix by its inverse gives an identity matrix (the R function `solve` computes the inverse of a matrix):
>
> ```
> (m3<-matrix(c(2,3,4,5),2,2))
>
> ##      [,1] [,2]
> ## [1,]    2    4
> ## [2,]    3    5
>
> (round(solve(m3)%*%m3))
>
> ##      [,1] [,2]
> ## [1,]    1    0
> ## [2,]    0    1
> ```

So:
$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{det}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$
Here, det is the determinant. Note that $det(m)$ is going to be the same as $det(m^{-1})$.

5.1.4.1 Inverse of a 3×3 matrix

Note: you will normally only use `solve` in R, you won't be doing this by hand.

Say you are given: $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$

1. Find determinant (the way to do this by hand is explained below). Det(m)=6.

2. Define co-factors of matrix and add alternating +/- signs, then find determinants:

$$\begin{pmatrix} +\begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} & -\begin{vmatrix} 0 & 1 \\ 0 & 2 \end{vmatrix} & +\begin{vmatrix} 0 & 2 \\ 0 & 1 \end{vmatrix} \\ -\begin{vmatrix} 0 & 0 \\ 1 & 2 \end{vmatrix} & +\begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} & -\begin{vmatrix} 2 & 0 \\ 0 & 1 \end{vmatrix} \\ +\begin{vmatrix} 0 & 0 \\ 2 & 1 \end{vmatrix} & -\begin{vmatrix} 2 & 0 \\ 0 & 1 \end{vmatrix} & +\begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} \end{pmatrix} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 4 & -2 \\ 0 & -2 & 4 \end{pmatrix}$$

3. Then take reflection of the above matrix; here, none is needed because it's symmetric. Then multiply by 1/det to get inverse:

$$\frac{1}{6} \times \begin{pmatrix} 3 & 0 & 0 \\ 0 & 4 & -2 \\ 0 & -2 & 4 \end{pmatrix} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 2/3 & -1/3 \\ 0 & -1/3 & 2/3 \end{pmatrix}$$

5.1.4.2 Inverse of a product of non-singular matrices

If A and B are non-singular matrices then $(AB)^{-1} = B^{-1}A^{-1}$.

5.1.5 Linear independence

Consider a 3x3 matrix. The The rows r_1, r_2, r_3 are linearly **dependent** if α, β, γ, not all zero, exist such that $\alpha r_1 + \beta r_2 + \gamma r_3 = (0,0,0)$.

If the rows or columns of A are linearly **dependent**, then det A=0 (the matrix is singular, not invertible).

5.1.6 The rank of a matrix

The column rank of a matrix is the maximum number of **linearly independent** columns in the matrix. The row rank is the maximum number of linearly independent rows. Column rank is always equal to row rank, so we can just call it rank.

5.1.7 More on determinants

The determinant is a value associated with a square (nxn) matrix.

1. If the determinant is non-zero, the system of linear equations expressed by the matrix has a unique solution.

2. If the determinant is zero, there are no solutions or many solutions.

3. We can invert a matrix only if its determinant is non-zero. (Inversion of matrices turns up a lot).

The determinant of a square matrix can be computed using an in-built R function.

```
(m1<-matrix(1:4,2,2))

##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4

det(m1)

## [1] -2
```

5.1.7.1 Computing determinants by hand

Note that you will almost never have to do this by hand, except for exercises. In real work, we just use R.

The det[erminant] of a 2×2 matrix like the one below is ad-bc.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

The det of a $n\times$ n matrix A, $n > 2$, is computed as follows:

$$detA = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} detA_{1j} \tag{5.10}$$

An example makes it easier to understand:

$$A = \begin{pmatrix} 1 & 5 & 0 \\ 2 & 4 & -1 \\ 0 & -2 & 0 \end{pmatrix} \tag{5.11}$$

Then, the determinant is (the vectors in red disappear):

$$detA = 1 \times \begin{pmatrix} 1 & 5 & 0 \\ 2 & 4 & -1 \\ 0 & -2 & 0 \end{pmatrix} - 5 \times \begin{pmatrix} 1 & 5 & 0 \\ 2 & 4 & -1 \\ 0 & -2 & 0 \end{pmatrix} + 0 \times \begin{pmatrix} 1 & 5 & 0 \\ 2 & 4 & -1 \\ 0 & -2 & 0 \end{pmatrix}$$

Another trick, for 3x3 matrices for example, is the following (from Vinod's book (Hands-on matrix algebra)):

1. Rewrite the matrix $\begin{pmatrix} a & d & g \\ b & e & h \\ c & f & i \end{pmatrix}$ by repeating the first two columns and coloring the matrix in two ways: $\begin{pmatrix} \textcolor{red}{a} & \textcolor{green}{d} & \textcolor{orange}{g} & a & d \\ b & \textcolor{red}{e} & \textcolor{green}{h} & \textcolor{orange}{b} & e \\ c & f & \textcolor{red}{i} & \textcolor{green}{c} & \textcolor{orange}{f} \end{pmatrix}$ and $\begin{pmatrix} a & d & \textcolor{red}{g} & \textcolor{green}{a} & \textcolor{orange}{d} \\ b & \textcolor{red}{e} & \textcolor{green}{h} & \textcolor{orange}{b} & e \\ \textcolor{red}{c} & \textcolor{green}{f} & \textcolor{orange}{i} & c & f \end{pmatrix}$

2. Then write out:

   $det = aei + dhc + gbf - ceg - fha - ibd$

5.1.8 Singularity

If the determinant of a square matrix is zero, then it can't be inverted; we say that the matrix is singular.

5.2 Solving simultaneous linear equations

We know how to solve for x and y in these two equations (by eliminating one variable—this is the method of elimination):

$$2x+3y=-1 \quad x-2y=3 \tag{5.12}$$

But what about:

$$x+y=2 \quad 2x+2y=1 \tag{5.13}$$

These two equations are contradictory (check this). There is no solution; we can also say that they are incompatible. Similarly, consider:

$$x+y=2 \quad 2x+2y=4 \tag{5.14}$$

These are both saying the same thing, and so reduce to one equation, with two unknowns. Therefore, there is an infinity of solutions.

If we have two equations, and the right-hand side has 0 in both equations, the equations are called homogeneous. Here, there are two possibilities:

First, if the equations are equivalent, as in:

$$x+y=0 \quad 2x+2y=0 \tag{5.15}$$

we again have an infinity of solutions (x=c and y=-c for any value of c).

Second, if the equations are not equivalent, as in:

$$2x+3y=0 \quad x-2y=0 \tag{5.16}$$

they are not incompatible because they have a single, unique solution, x=0, and y=0. This is called the trivial solution.

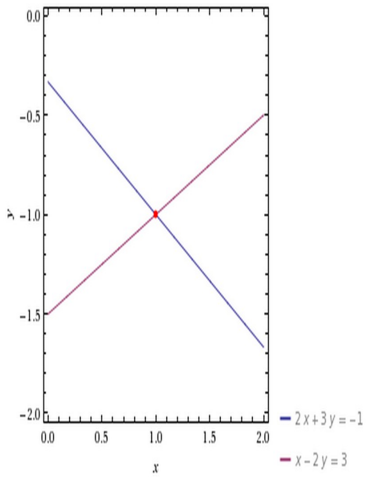
To summarize: Inhomogeneous and homogeneous equations can have

1. a unique solution
2. no solution
3. an infinity of solutions

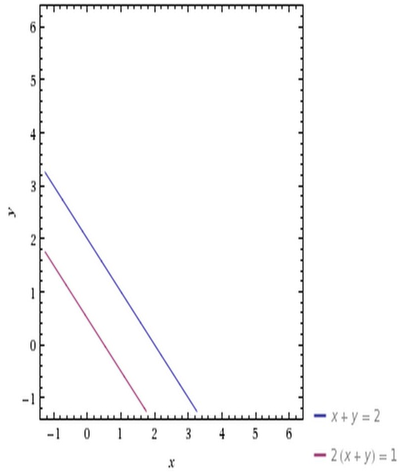
In addition, homogeneous equations that are not equivalent can have a trivial solution.

It is easy to use the elimination method to solve equations with two unknowns, but for equations with many unknowns, we need different methods to avoid tedium and error-prone calculations.

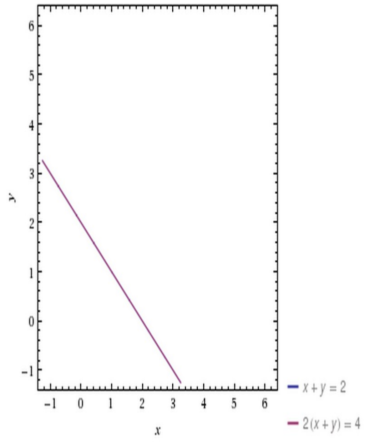
Unique solution:



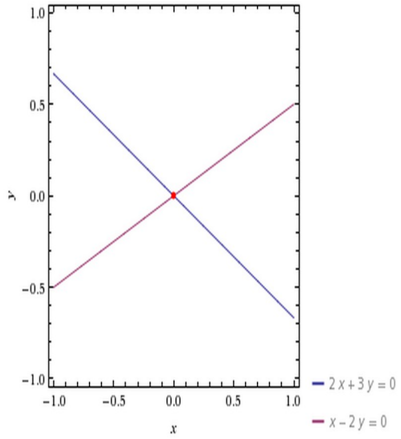
No solution:



Infinity of solutions:



Trivial solution:



5.2.1 Gaussian elimination

Given three (or more) equations like:

$$x_1 + 2x_2 + x_3 = 1 \tag{5.17}$$

$$-2x_1 + 3x_2 - x_3 = -7 \tag{5.18}$$

$$x_1 + 4x_2 - 2x_3 = -7 \tag{5.19}$$

There are three elementary row operations that do not affect the solution:

1. Any equation can be multiplied by a non-zero constant
2. any two equations can be interchanged
3. any equation can be replaced the sum of itself and any multiple of another equation

It's easiest to illustrate how these operations lead to a solution with a concrete example:

Suppose we are given the three equations shown below. We use Gaussian elimination to solve the equations. I will refer to the rows of a matrix by ρ_n, where n refers to the n-th row.

$$\begin{array}{ccccccc} 2x & + & -y & + & 3z & = & 8 \\ -x & + & 6y & + & z & = & 17 \\ -2x & + & 5y & + & 3z & = & 24 \end{array} \tag{5.20}$$

We could write this in matrix form: Ax = d, where

$$A = \begin{pmatrix} 2 & -1 & 3 \\ -1 & 6 & 1 \\ -2 & 5 & 3 \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad d = \begin{pmatrix} 8 \\ 17 \\ 24 \end{pmatrix} \tag{5.21}$$

The general strategy will be to convert the matrix to **echelon form** and then to **reduced echelon form**.

A matrix is in echelon form if it has 0's below the diagonal elements starting from the top left. See equation 5.26. A matrix is in reduced echelon form if it in echelon form and has 1's along the diagonal starting from the top left. See equation 5.27.

First, we convert the coefficients into an **augmented matrix**:

$$\left(\begin{array}{ccc|c} 2 & -1 & 3 & 8 \\ -1 & 6 & 1 & 17 \\ -2 & 5 & 3 & 24 \end{array}\right) \tag{5.22}$$

Next, move the third row (ρ) to second position:

$$\left(\begin{array}{ccc|c} 2 & -1 & 3 & 8 \\ -2 & 5 & 3 & 24 \\ -1 & 6 & 1 & 17 \end{array}\right) \tag{5.23}$$

Add ρ_1 to the new ρ_2 :

$$\left(\begin{array}{ccc|c} 2 & -1 & 3 & 8 \\ 0 & 4 & 6 & 32 \\ -1 & 6 & 1 & 17 \end{array}\right) \tag{5.24}$$

Add ρ_1 to $2 \times \rho_3$ (this is the new ρ_3):

$$\left(\begin{array}{ccc|c} 2 & -1 & 3 & 8 \\ 0 & 4 & 6 & 32 \\ 0 & 11 & 5 & 42 \end{array}\right) \tag{5.25}$$

Subtract $11 \times \rho_2$ from $4 \times \rho_3$. This gives us the **echelon form**:

$$\left(\begin{array}{ccc|c} 2 & -1 & 3 & 8 \\ 0 & 4 & 6 & 32 \\ 0 & 0 & 46 & 184 \end{array}\right) \tag{5.26}$$

Next, we convert the matrix to **reduced echelon form**. Divide ρ_1 by 2, and ρ_2 by 4, and ρ_3 by 46:

$$\left(\begin{array}{ccc|c} 1 & -1/2 & 3/2 & 4 \\ 0 & 1 & 6/4 & 8 \\ 0 & 0 & 1 & 4 \end{array}\right) \tag{5.27}$$

Next, we eliminate the nonzero values to the right of the 1's in each row. Subtract 1.5 times ρ_3 from ρ_2

$$\left(\begin{array}{ccc|c} 1 & -1/2 & 3/2 & 4 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \end{array}\right) \tag{5.28}$$

Add $1.5 \times \rho_2$ to ρ_1:

$$\left(\begin{array}{ccc|c} 1 & 0 & 3/2 & 5 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \end{array}\right) \tag{5.29}$$

Subtract $1.5\rho_3$ from ρ_1:

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \end{array}\right) \tag{5.30}$$

This tells us that $x = -1$, $y = 2$, $z = 4$.
Answer: $x = -1$, $y = 2$, $z = 4$.

How to solve multiple-variable simultaneous linear equations in R:

```
X <- matrix(c(2,-1,3,-1,6,1,-2,5,3), ncol=3,byrow=T)
y<-c(8,17,24)
library(MASS)
## correct:
fractions(solve(crossprod(X))%*%crossprod(X,y))

##      [,1]
## [1,] -1
## [2,]  2
## [3,]  4
```

You can use Gaussian elimination to find the inverse of a matrix:

$$AA^{-1} = I \tag{5.31}$$

Apply a sequence of row operations that transform A into I on the left-hand side, so that the left hand side becomes A^{-1}. Apply the same sequence of row operations to I on the right-hand side, and you will get A^{-1}. The two examples below illustrate this point.

Example 1: Find the inverse of

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 5 & 3 \\ 0 & 3 & 2 \end{pmatrix}$$

The algorithm is: row-reduce the augmented matrix $[M_1 I]$. If M_1 is row equivalent to I, then $[M_1 I]$ is row equivalent to $[IM_1^{-1}]$. Otherwise M_1 doesn't have an inverse.

First, we append the identity matrix to the right side of M_1:

$$[M_1 I] = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ -1 & 5 & 3 & 0 & 1 & 0 \\ 0 & 3 & 2 & 0 & 0 & 1 \end{pmatrix}$$

The next step is to convert the left half of the matrix $[M_1 I]$ into an identity matrix.

Add ρ_1 to ρ_2:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ -1+1 & 5+0 & 3+0 & 0+1 & 1+0 & 0+0 \\ 0 & 3 & 2 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 5 & 3 & 1 & 1 & 0 \\ 0 & 3 & 2 & 0 & 0 & 1 \end{pmatrix}$$

Subtract $5 \times \rho_3$ from $3 \times \rho_2$ and place the result in ρ_3:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 5 & 3 & 1 & 1 & 0 \\ 0 & 0 & -1 & 3 & 3 & -5 \end{pmatrix}$$

Multiply ρ_3 by -1:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 5 & 3 & 1 & 1 & 0 \\ 0 & 0 & 1 & -3 & -3 & 5 \end{pmatrix}$$

Divide ρ_2 by 5:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 3/5 & 1/5 & 1/5 & 0 \\ 0 & 0 & 1 & -3 & -3 & 5 \end{pmatrix}$$

Subtract ρ_2 from $3/5 \times \rho_3$ and place the result in ρ_2:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & -2 & -2 & 3 \\ 0 & 0 & 1 & -3 & -3 & 5 \end{pmatrix}$$

Multiply ρ_2 by -1:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 2 & 2 & -3 \\ 0 & 0 & 1 & -3 & -3 & 5 \end{pmatrix}$$

The inverse of M_1 is therefore:

$$M_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & -3 \\ -3 & -3 & 5 \end{pmatrix}$$

We cross-check this with R:

```
## correct:
M1<-matrix(c(1,0,0,-1,5,3,0,3,2),byrow=T,nrow=3)
## note: ginv is in library MASS, loaded earlier
fractions(ginv(M1))

##      [,1] [,2] [,3]
## [1,]  1    0    0
## [2,]  2    2   -3
## [3,] -3   -3    5
```

We check our result regarding the inverse by evaluating $M_1 M_1^{-1}$. If the product of this matrix multiplication is I, then M_1^{-1} is the correct inverse.

$$M_1 M_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 5 & 3 \\ 0 & 3 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & -3 \\ -3 & -3 & 5 \end{pmatrix}$$

The result is:

$$M_1 M_1^{-1} = \begin{pmatrix} 1\cdot 1 + 2\cdot 0 + -3\cdot 0 & 0\cdot 1 + 2\cdot 0 + -3\cdot 0 & 0\cdot 1 + -3\cdot 0 + 5\cdot 0 \\ 1\cdot -1 + 2\cdot 5 + -3\cdot 3 & 0\cdot -1 + 2\cdot 5 + -3\cdot 3 & 0\cdot -1 + -3\cdot 5 + 5\cdot 3 \\ 1\cdot 0 + 2\cdot 3 + -3\cdot 2 & 0\cdot 0 + 2\cdot 3 + -3\cdot 2 & 0\cdot 0 + -3\cdot 3 + 5\cdot 2 \end{pmatrix}$$

Simplifying:

$$M_1 M_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

This confirms that we have the correct inverse.

We cross-check our calculations using R:

```
M1<-matrix(c(1,0,0,-1,5,3,0,3,2),byrow=T,nrow=3)
## correct:
fractions(M1%*%fractions(ginv(M1)))

##      [,1] [,2] [,3]
## [1,] 1    0    0
## [2,] 0    1    0
## [3,] 0    0    1
```

Answer: The inverse of $\begin{pmatrix} 1 & 0 & 0 \\ -1 & 5 & 3 \\ 0 & 3 & 2 \end{pmatrix}$ is $\begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & -3 \\ -3 & -3 & 5 \end{pmatrix}$.

Example 2: Find the inverse of M_2:

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 6 & 4 \\ 0 & 3 & 2 \end{pmatrix}$$

First, we augment the matrix:

$$M_2 I = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ -1 & 6 & 4 & 0 & 1 & 0 \\ 0 & 3 & 2 & 0 & 0 & 1 \end{pmatrix}$$

We replace ρ_2 with $\rho_1 + \rho_2$:

$$M_2 I = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 6 & 4 & 1 & 1 & 0 \\ 0 & 3 & 2 & 0 & 0 & 1 \end{pmatrix}$$

Divide ρ_2 by 2:

$$M_2 I = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 3 & 2 & 1/2 & 1/2 & 0 \\ 0 & 3 & 2 & 0 & 0 & 1 \end{pmatrix}$$

There is no way to proceed here since the second and third rows can never be changed such that we have an identity matrix on the left hand side. I.e., no inverse exists for this matrix.

We cross-check this result using R by computing the inverse and then multiplying the original matrix with the purported inverse; we should get the identity matrix. As the R code shows, we do not get the identity matrix.

```
M2<-matrix(c(1,0,0,-1,6,4,0,3,2),byrow=T,nrow=3)
## should have been an identity matrix:
fractions(M2%*%fractions(ginv(M2)))

##      [,1] [,2] [,3]
## [1,]  5/6 -1/6  1/3
## [2,] -1/6  5/6  1/3
## [3,]  1/3  1/3  1/3
```

Answer: M_2 does not have an inverse because it is not row equivalent to I.

5.3 Eigenvalues and eigenvectors

Any set of equations Ax=0 is a homogeneous set. Clearly, this will have at least the trivial solution x=0. The more interesting question is: are there any non-trivial solutions? It turns out that such a set of equations will have non-trivial solutions only if det A is 0.

Now consider the nxn set of equations: $Ax = \lambda x$ or $(A - \lambda I)x = 0$. In order for these equations to have non-trivial solutions, it must be true (see previous paragraph) that

$$det(A - \lambda I) = 0 \tag{5.32}$$

The above is called the **characteristic equation** of A, and the λ that satisfy the equation are called **eigenvalues**.

Example:

Find the eigenvalues of

$A = \begin{pmatrix} 1 & 3 \\ 2 & 2 \end{pmatrix}$

Solve $det(A - \lambda I) = 0$:

$$det \begin{pmatrix} 1-\lambda & 3 \\ 2 & 2-\lambda \end{pmatrix} = (1-\lambda)(2-\lambda) - 6 \Rightarrow \lambda^2 - 3\lambda - 4 = 0 \tag{5.33}$$

This gives us the factorization $(\lambda-4)(\lambda+1)$. It follows that the eigenvalues are $\lambda_1=-1, \lambda_2=4$ (we will order them from smallest to largest).

> Recall the rule for figuring out the roots of a quadratic equation: $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$

Associated with each eigenvalue we have a non-trivial solution to the equation $(A-\lambda I)x=0$. Each of the solutions corresponding to each eigenvalue is called the eigenvector. So, once you have found an eigenvalue λ_1, you can find the eigenvector by solving the simultaneous linear equation $(A-\lambda_1 I)x=0$ using Gaussian elimination.

For our above example, we have $\lambda_1=-1, \lambda_2=4$. So, the eigenvectors are:

$$s_1=\begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \quad s_2=\begin{pmatrix} a_2 \\ b_2 \end{pmatrix} \tag{5.34}$$

Recall that $A=\begin{pmatrix} 1 & 3 \\ 2 & 2 \end{pmatrix}$.

We need to find the two solutions: $(A-\lambda_1 I)s_1=0$ and $(A-\lambda_2 I)s_2=0$.

As an example, consider $(A-\lambda_1 I)s_2=0$. Expanding everything out:

$$\begin{pmatrix} 1 & 3 \\ 2 & 2 \end{pmatrix} - \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{5.35}$$

This gives us:

$$\begin{pmatrix} -3 & 3 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{5.36}$$

The solution is $a_2=b_2=\alpha$ for any value α.

You can do a similar calculation to get the eigenvector s_1.

Normally, we will use R for calculating eigenvalues and eigenvectors:

```
m<-matrix(c(1,3,2,2),byrow=T,ncol=2)
eigen(m)

## $values
## [1]  4 -1
##
## $vectors
##             [,1]       [,2]
## [1,] -0.7071068 -0.8320503
## [2,] -0.7071068  0.5547002
```

However, you should know how to do this by hand for 2x2 matrices; for larger matrices, just use R.

Note that a zero eigenvalue implies that A is singular, i.e., det A = 0. If A singular, then A has at least one 0 eigenvalue.

5.3.1 Linear dependence

Consider the set of all mx1 column vectors; this set is the m-dimensional vector space V_m. Thus, if

$$s_1 = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} s_2 = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \tag{5.37}$$

Then s_1 and s_2 belong to V_m and so does

$$\alpha s_1 + \beta s_2 \tag{5.38}$$

for any α and β.

We refer to as the set of **base vectors** in V_m the following:

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \dots e_m = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \tag{5.39}$$

Any vector in V_m can be expressed as a linear combination of these vectors: For any $i = 1, \dots, m$,

$$s_i = a_1 e_1 + \cdots + a_m e_m \tag{5.40}$$

This is why $e_1, \dots, e_m$ forms a **basis** for V_m. Note that none of the e_i can be expressed as a linear combination of the others; i.e., they are **linearly independent**.

The set of n column vectors $s_1, \ldots, s_n$ is said to be linearly dependent if there exists constants $a_1, \ldots, a_n$ (not all zero) such that

$$a_1 s_1 + \cdots + a_n s_n = 0 \tag{5.41}$$

If the above equation holds only when $a_1 = \cdots = a_n = 0$, then the $s_1 \ldots s_n$ are linearly dependent.

Note that any set of m linearly independent vectors form a basis of the vector space V_m.

Example:

The column vectors below for a basis in three dimensions: $a_1 = [1,1,0]^T, a_2 = [1,0,1]^T, a_3 = [0,0,1]^T$.

To show this, we have to verify that

$$x a_1 + y a_2 + z a_3 = 0 \tag{5.42}$$

has non-zero solutions for x, y, z. We are trying to solve the equations:

$$\begin{pmatrix} x & + & y & & & = 0 \\ x & + & & & z & = 0 \\ & & y & + & z & = 0 \end{pmatrix} \tag{5.43}$$

The coefficient matrix is:

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \tag{5.44}$$

Its determinant is -2 (i.e., not 0). The only solution is $x = y = z = 0$. Therefore the vectors are linearly independent, and can form a basis.

> For a square matrix A, if det A = 0, there is an infinite number of non-trivial solutions. If $detA \neq 0$, the only solution is x = 0.

5.3.2 Diagonalization of a matrix

The eigenvalue and eigenvector information contained in a matrix can be used in a useful factorization of the type $D = C^{-1}AC$ where D is a diagonal matrix. Since computing powers of a diagonal matrix D are easy, we can use this factorization to find powers of the matrix A very quickly (see section 5.1.3).

Consider the matrix m and its eigenvalues and eigenvectors:

```
m<-matrix(c(1,2,1,2,1,1,1,1,2),byrow=FALSE,nrow=3)
eigen_values_m<-eigen(m)$values
eigen_vectors_m<-eigen(m)$vectors
```

Note that the eigenvectors are linearly independent, and so the matrix is non-singular:

```
det(eigen_vectors_m)

## [1] -1
```

Let the matrix of eigenvectors (`eigen_vectors_m` above) be the matrix C. Note that $AC = A[s_1\ s_2\ s_3]$, where s_1, s_2, s_3 are the eigenvectors of A. Now, s_i is by definition a non-zero solution to $As_i = \lambda_1 s_i$.

Let D be the diagonal matrix of eigenvalues:

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \tag{5.45}$$

Now,

$$AC = CD \tag{5.46}$$

[Verify this.]

Pre-multiplying both sides with C^{-1}, we get:

$$C^{-1}AC = C^{-1}CD = D \tag{5.47}$$

We say that the operation $C^{-1}AC$ has diagonalized the matrix A.

Summary of steps: To diagonalize a matrix A:

1. find the eigenvalues of A
2. find n linearly independent eigenvectors s_n of A (if they exist)
3. construct the matrix C of eigenvectors
4. calculate the inverse of C
5. compute $C^{-1}AC$

5.3.2.1 Application of diagonalization: finding powers of matrices

For a diagonal matrix D, we can find any power D^n quickly by just raising the diagonal elements $d_i i$ to the n-th power.

To find the power of any matrix A, first note that:

$$AC = CD \tag{5.48}$$

Post-multiplying both sides by C^{-1}, we get:

$$ACC^{-1} = A = CDC^{-1} \tag{5.49}$$

It follows that

$$A^2 = CDC^{-1}CDC^{-1} = CDDC^{-1} = CD^2C^{-1} \tag{5.50}$$

$$A^3 = CDC^{-1}CDC^{-1}CDC^{-1} = CD^3C^{-1} \tag{5.51}$$

and in general, it should be obvious that:

$$A^n = \overbrace{CDC^{-1}CDC^{-1} \ldots CDC^{-1}CDC^{-1}}^{n \text{ times}} = CD^nC^{-1} \tag{5.52}$$

5.3.3 Quadratic forms

Let $x = [x_1, x_2, \ldots, x_n]^T$ be an n-dimensional column vector. Any polynomial function of these elements in which every term is of degree 2 in them is called a **quadratic form**. For n=3, we have (for example):

$$x_1^2 + 8x_1x_2 + 6x_2x_3 + x_3^2 \tag{5.53}$$

Quadratic forms can always be expressed in matrix form:

$$x^T Ax \tag{5.54}$$

The above example in equation 5.53 can be restated as:

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1 & 4 & 0 \\ 4 & 1 & 3 \\ 0 & 3 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{5.55}$$

Note that A is a symmetric matrix. Any quadratic form can be written using a symmetric A (although non-symmetric representations are possible).

Now let's compute the eigenvalues and eigenvectors:

```
m<-matrix(c(1,4,0,4,1,3,0,3,1),byrow=FALSE,ncol=3)
val<-eigen(m)$values
vec<-eigen(m)$vectors
```

Note that the eigenvectors are **orthogonal**:

```
round(t(vec[,1])%*%vec[,2],5)

##      [,1]
## [1,]    0

round(t(vec[,1])%*%vec[,3],5)

##      [,1]
## [1,]    0

round(t(vec[,2])%*%vec[,3],5)

##      [,1]
## [1,]    0
```

[Two vectors a_n and b_n are orthogonal when $a^T b = 0$. We say that they are **mutually perpendicular**.]

This property of orthogonality comes from the fact that the matrix is symmetric:

Theorem 1. *If A is a real symmetric matrix, then the eigenvectors associated with any two distinct eigenvalues are orthogonal.*

5.3.3.1 Positive-definite matrices

A quadratic form is positive definite if $x^T Ax > 0$ for all $x \neq 0$. The matrix A is also called positive definite.

A symmetric matrix A is positive definite if and only if its eigenvalues are positive.

There is a unique decomposition of A

$$A = LL^T \tag{5.56}$$

where L is the lower triangular matrix (this matrix has non-zero values in the lower-triangular part, and 0's in the upper triangular part). The equation 5.56 is called the Cholesky decomposition. How to compute this in R:

```
m <- matrix(c(5,1,1,3),2,2)
L <- t(chol(m))
L%*%t(L)

##      [,1] [,2]
## [1,]    5    1
## [2,]    1    3
```

There is also a unique decomposition of A such that

$$A = VDV^T \quad V^TV = I \tag{5.57}$$

is called the **singular value decomposition**.

Chapter 6

Multivariate calculus

6.1 Partial derivatives

It is easy to imagine that more than one variable may be involved in a function:

$z = f(x,y) = x^3 + y^3$

We will encounter such functions in multivariate statistics (see chapter 7).

We will say that we are taking a partial derivative if we differentiate a function like z above with respect to x (treating y as a constant) or with respect to y (treating x as a constant).

We write these two cases as $\frac{\partial f(x,y)}{\partial x}$ and $\frac{\partial f(x,y)}{\partial y}$. Higher derivatives can be computed in the usual way we have seen.

We can also take mixed derivatives, where we first differentiate f(x,y) with respect to x and then with respect to y; this is written:

$\frac{\partial^2 f(x,y)}{\partial \partial y}$

Note that order of differentiation does not matter for the kinds of functions we will encounter: $\frac{\partial^2 f(x,y)}{\partial x \partial y} = \frac{\partial^2 f(x,y)}{\partial y \partial x}$.

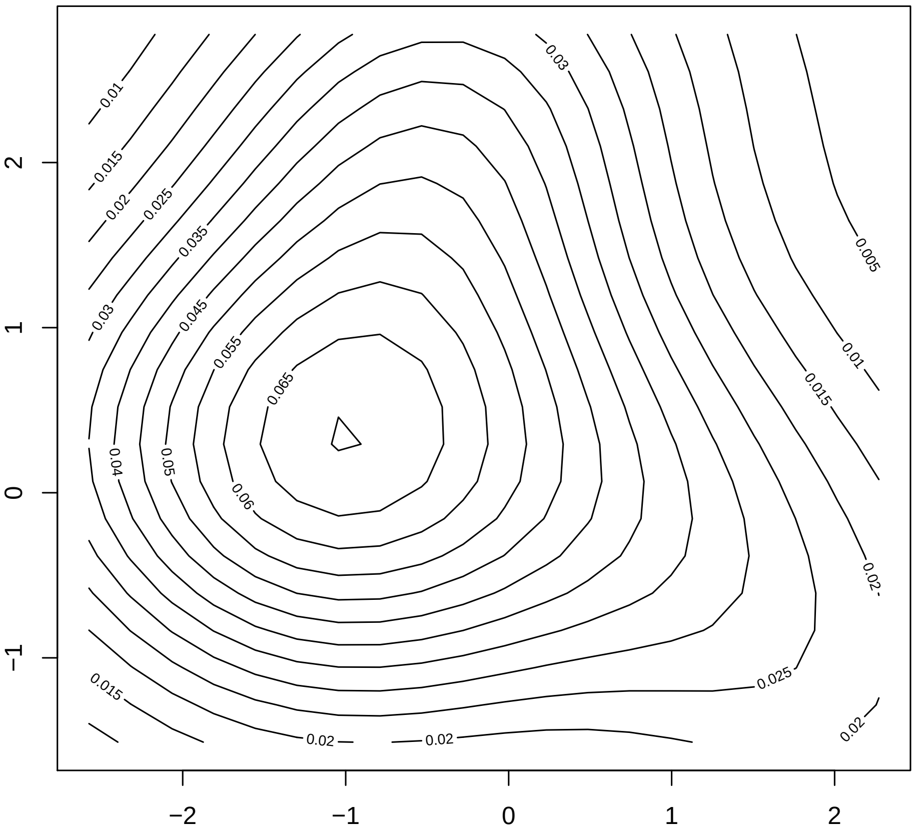
6.2 Drawing level curves

You can draw contour plots of a bivariate function, also called level curves, given x and y values stored in a matrix m (x in the first column, and y in the second).

```
## source:
#https://stat.ethz.ch/pipermail/r-help/
```

```
##2007-October/142470.html
dens2d<-function(x, nx = 20, ny = 20,
                 margin = 0.05, h = 1)
{
 xrange <- max(x[, 1]) - min(x[, 1])
 yrange <- max(x[, 2]) - min(x[, 2])
 xmin <- min(x[, 1]) - xrange * margin
 xmax <- max(x[, 1]) + xrange * margin
 ymin <- min(x[, 2]) - yrange * margin
 ymax <- max(x[, 2]) + yrange * margin
 xstep <- (xmax - xmin)/(nx - 1)
 ystep <- (ymax - ymin)/(ny - 1)
 xx <- xmin + (0:(nx - 1)) * xstep
 yy <- ymin + (0:(ny - 1)) * ystep
 g <- matrix(0, ncol = nx, nrow = ny)
 n <- dim(x)[[1]]
 for(i in 1:n) {
  coefx <- dnorm(xx - x[i, 1], mean = 0, sd = h)
  coefy <- dnorm(yy - x[i, 2], mean = 0, sd = h)
  g <- g + coefx %*% t(coefy)/n
 }
 return(list(x = xx, y = yy, z = g))
}

m<-matrix(cbind(x=rnorm(10),y=rnorm(10)),ncol=2)
```

```
contour(dens2d(m))
```



6.3 Double integrals

6.3.1 How to solve double integrals

An example may help in seeing how to solve double integrals. Let

$$I = \int_0^1 \int_0^2 (xy + y^2 - 1)\, dxdy \tag{6.1}$$

1. Put brackets around the inner integral, and relabel the limits so that it is clear which integral is for which variable:

$$I = \int_{y=0}^{1} \left(\int_{x=0}^{2} (xy + y^2 - 1)\,dx \right) dy \tag{6.2}$$

2. Treat y as a constant and evaluate the inner integral with respect to x:

$$\left(\int_{x=0}^{2} (xy + y^2 - 1)\,dx \right) = \left[\frac{1}{2}x^2y + xy^2 - x \right]_{x=0}^{2} = 2y + 2y^2 - 2 \tag{6.3}$$

   Note that x has been eliminated.

3. Use the result of the inner integration for the second integral:

$$\int_{y=0}^{1} 2y + 2y^2 - 2dy = \left[y^2 + \frac{2}{3}y^3 - 2y \right]_{y=0}^{1} = 1 - \frac{2}{3} - 2 = -\frac{1}{3} \tag{6.4}$$

See the section on jointly distributed random variables (page 95, section 7.4) for some applications of double integrals in statistics.

6.3.2 Double integrals using polar coordinates

Any (x,y) point in a cartesian plane can be stated in terms of the angle θ of the line from the origin to the point (x,y), and the length r of the line from the origin to the point (x,y). Recall that

$$\cos\theta = \frac{x}{r} \quad \sin\theta = \frac{y}{r} \tag{6.5}$$

Therefore, $x = r\cos\theta, y = r\sin\theta$.

It is easier to solve the double integral $\iint (1 - x^2 - y^2)\,dxdy$, (where $x^2 + y^2 < 1; x, y < 1$) if we convert to polar coordinates.

Note that $\iint (1 - x^2 - y^2)\,dxdy$ is just a sum of all the areas of small near-rectangles; we can call the areas dA:

$$\iint f(x,y)\,dA \tag{6.6}$$

```r
library(plotrix)
plot(0:2,0:2,type="n",xlab="",ylab="")
draw.circle(0,0,1,border="black",lty=1,lwd=1)
arrows(x0=0,y0=0,x1=.7,y1=.7,code=2,length=0)
arrows(x0=0,y0=0,x1=.45,y1=.9,code=2,length=0)
draw.circle(0,0,.7,border="black",lty=2,lwd=1)
draw.circle(0,0,.9,border="black",lty=2,lwd=1)
text(.48,.75,"dA",cex=2)
text(.4,.6,expression(paste("rd",theta)),cex=1)
text(.6,.55,"dr",cex=1)
```

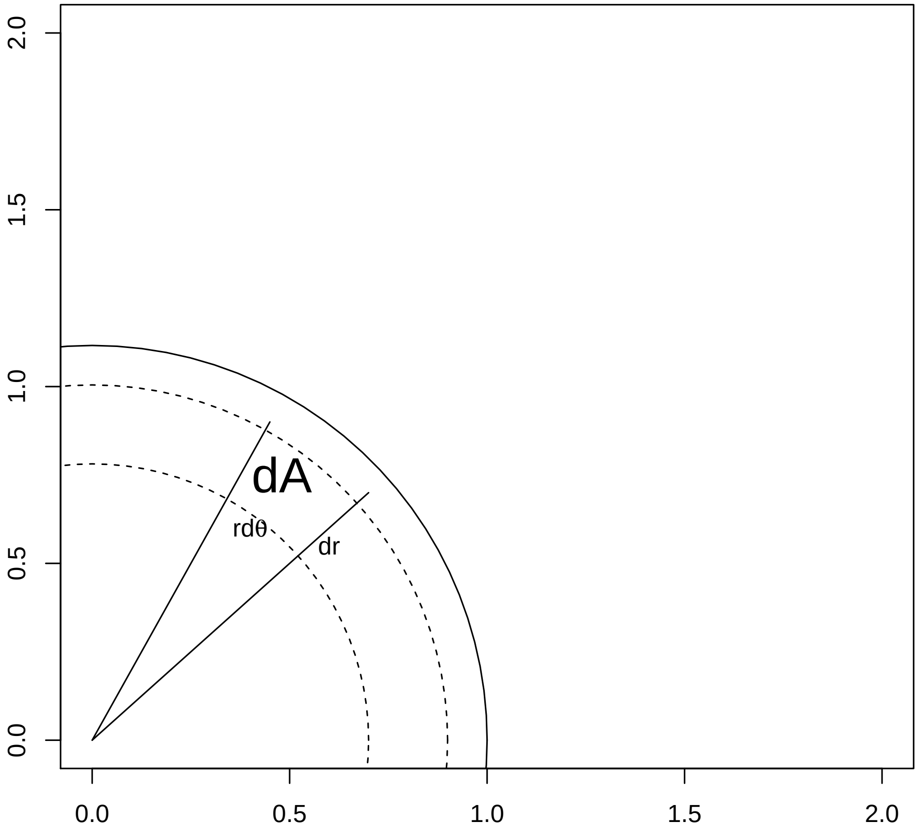


Figure 6.1: The function f(x,y). (The circle has not been drawn correctly. Need to fix this.)

Now, given a small rectangle with side dr, and a small angle $d\theta$, its area will be $rdrd\theta$, because the arc's length will be (by the definition of radian) $rd\theta$, the width of the rectangle is dr. See Figure 6.1.

So, in polar coordinate terms, our integral will look like:

$$\iint f(x,y)\,dA = \iint (1-x^2-y^2)\,rdrd\theta = \iint (1-(x^2+y^2))\,rdrd\theta = \iint (1-r^2)\,rdrd\theta \tag{6.7}$$

The last line above holds because $r^2 = x^2+y^2$ by Pythagoras' theorem.

Next, we define the upper and lower bounds of the integrals; we can do this because we are given that $x^2+y^2 < 1; x,y < 1$. This describes a quarter circle on the first quadrant of the cartesian plane, with radius 1.

$$\begin{aligned}\iint (1-r^2)\,rdrd\theta &= \int_0^{\pi/2}\int_0^1 (1-r^2)\,rdrd\theta \\ &= \int_0^{\pi/2} [\frac{r^2}{2}-\frac{r^4}{4}]_0^1 d\theta \\ &= \int_0^{\pi/2} \frac{1}{4} d\theta \\ &= \frac{\pi}{8}\end{aligned} \tag{6.8}$$

More generally, if the region R is the region $a \le r \le b, \alpha \le \theta \le \beta$, then

$$\iint_R f(x,y)\,dxdy = \int_\alpha^\beta \int_a^b f(r,\theta) rdrd\theta \tag{6.9}$$

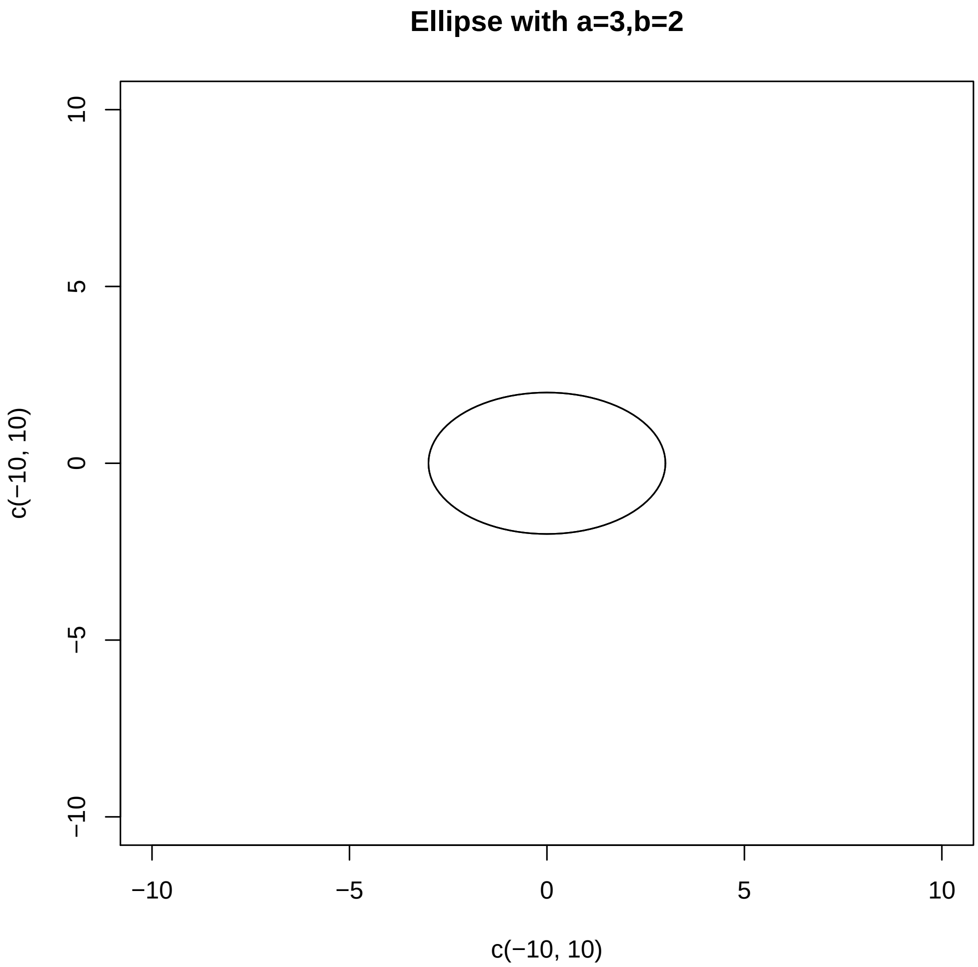
This change of variables is related to the next topic.

6.3.3 The Jacobian in a change of variables transformation

We just saw that it is sometimes convenient to transform a function that is in terms of x,y, into another function in terms of u,v (above, we transformed to r, θ).

Suppose we need to solve $\iint (\frac{x}{a})^2 + (\frac{y}{b})^2\,dxdy$, given that $(\frac{x}{a})^2 + (\frac{y}{b})^2 < 1$. We are basically looking to find the area inside an ellipse.

```
plot(c(-10,10), c(-10,10), type="n",
     main="Ellipse with a=3,b=2")
draw.ellipse(c(0,0),c(0,0),a=3,b=2,angle=0)
```



To make this integral simpler to solve, we can change variables. Say $x/a = u, y/b = v$. Now our region R is $u^2 + v^2 < 1$. This looks much easier to solve because it is a circle now, with radius 1. Remember that its area is going to π.

$x/a = u$ implies $du = \frac{1}{a}dx$

$y/b = v$ implies $dv = \frac{1}{b}dy$

So, $dudv = \frac{1}{ab}dxdy$. Or, $dxdy = abdudv$. So, in the double integral, I can replace dxdy with *abdudv*:

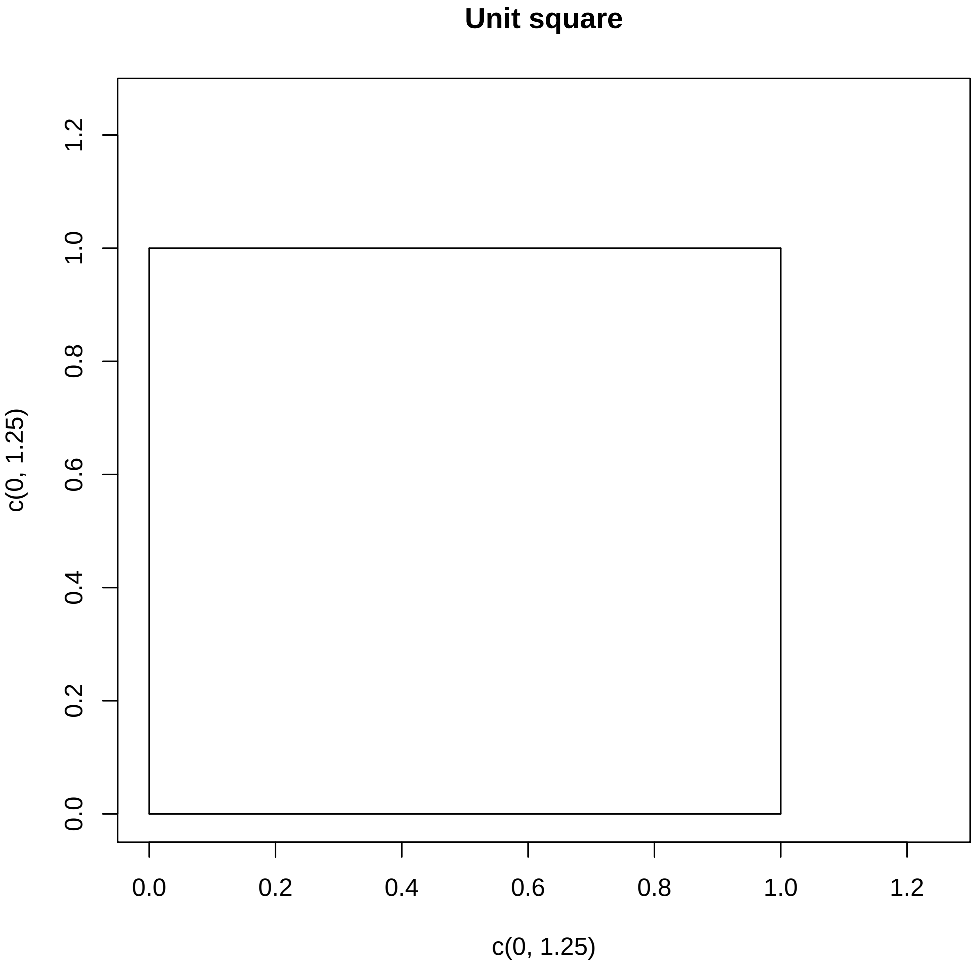
$$\iint_R (u^2+v^2)abdudv = ab\iint_R u^2+v^2dudv = ab\pi \tag{6.10}$$

(because the area of a unit circle is π).

When we want to find the area in a double integral with variables x and y, the area will be $dA = dxdy$. When we transform the variables to u,v, the area in the u,v dimensions will be $dA' = dudv$. To take a concrete example, suppose $u = 3x-2y$ and $v = x+y$. Then dA will be the area of a rectangle, and the dA' will be the area of a parallelogram (the transformation just twists the rectangle in the x,y space to a parallelogram in u,v space). The area dA will not be the same as the are dA'; i.e., there is some scaling factor k such that $dA = kdA'$.

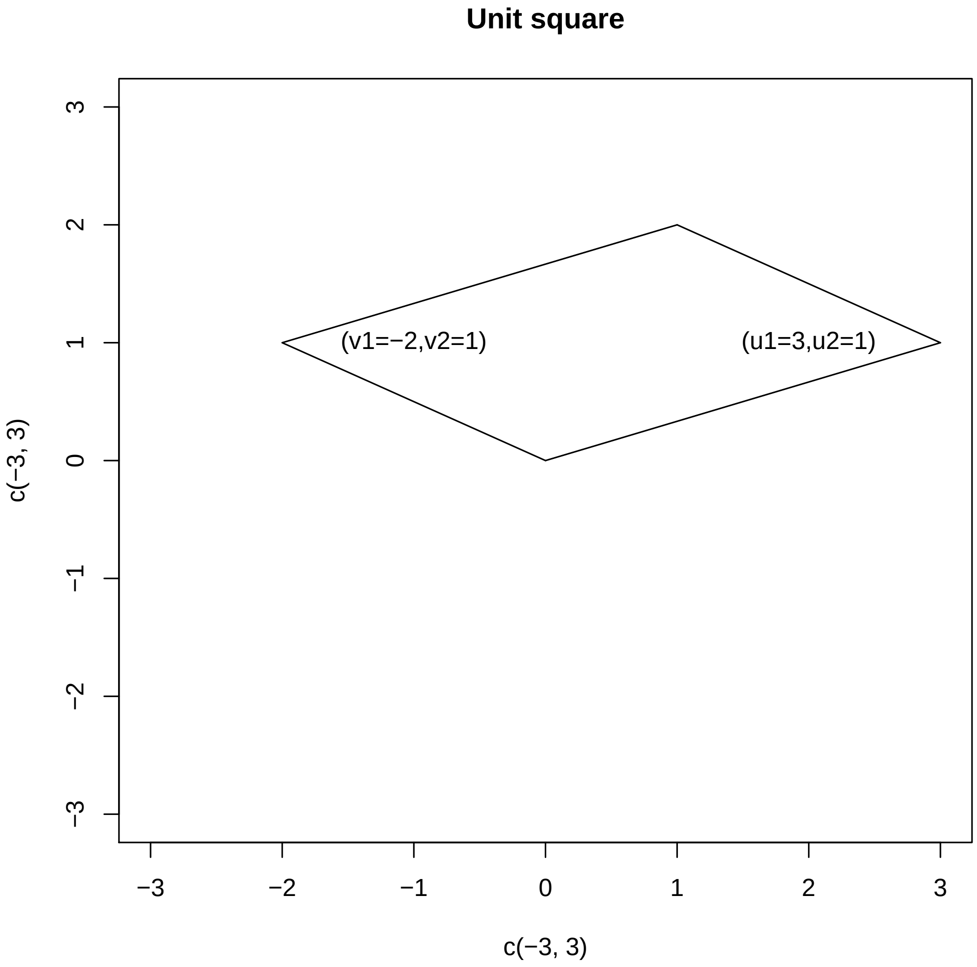
As a concrete example, consider the unit square:

```
plot(c(0,1.25), c(0,1.25), type="n", main="Unit square")
arrows(x0=1,y0=0,x1=1,y1=1,code=2,length=0)
arrows(x0=1,y0=1,x1=0,y1=1,code=2,length=0)
arrows(x0=0,y0=0,x1=0,y1=1,code=2,length=0)
arrows(x0=0,y0=0,x1=1,y1=0,code=2,length=0)
```



In u,v terms, we have a paralleogram:

```
plot(c(-3,3), c(-3,3), type="n", main="Unit square")
arrows(x0=0,y0=0,x1=3,y1=1,code=2,length=0)
arrows(x0=3,y0=1,x1=1,y1=2,code=2,length=0)
arrows(x0=1,y0=2,x1=-2,y1=1,code=2,length=0)
arrows(x0=-2,y0=1,x1=0,y1=0,code=2,length=0)
text(2,1,"(u1=3,u2=1)",cex=1)
text(-1,1,"(v1=-2,v2=1)",cex=1)
```



Now, it can be shown that the area (dA') of a parallegram is the length of the cross-product of the two vectors that make up two vectors **u** and **v**. That can be shown to be the absolute value of the determinant of the matrix A:

$$A = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 \\ \mathbf{v}_1 & \mathbf{v}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 = 3 & \mathbf{u}_2 = 1 \\ \mathbf{v}_1 = -2 & \mathbf{v}_2 = 1 \end{pmatrix} \tag{6.11}$$

Now, $\mid detA \mid = u_1 v_2 - u_2 v_1 = 5$. That means that the parallelogram has five times the size than the unit square. So there is always a scaling factor needed when we do a change of variables.

So, in the above example, $dA' = 5dA$, i.e., $dudv = 5dxdy$. If we are integrating in terms of u,v, we have to correct for the fact that the area in u,v coordinates is 5 times larger:

$$\iint \dots dxdy = \iint \dots \frac{1}{5} dudv \tag{6.12}$$

The scaling factor $\frac{1}{5}$ is called the Jacobian. We could have written

$$\iint \dots dxdy = \iint \dots \frac{1}{\mid detA \mid} dudv \tag{6.13}$$

We say that the Jacobian J=det A, where A is the relevant matrix. I discuss this matrix next by considering the general case.

Let $u = f(x,y)$, and $v = g(x,y)$, i.e., u and v are some functions of x and y. Let f and g be continuously differentiable functions over some region Γ. The point (u,v) generates a region Ω in the u-v plane: (f(x,y),g(x,y)). The Jacobian in this case can be computed by:

$$J(x,y) = \frac{\partial(u,v)}{\partial(x,y)} = det \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{pmatrix} \tag{6.14}$$

The main thing to understand here is that the Jacobian is the scaling factor when you change variables.

> To express a double integral in new coordinates
> If $x = g(u,v), y = h(u,v)$, then:
>
> $$\iint_E f(x,y)\, dxdy = \iint_S f(g(u,v),h(u,v)) \mid \frac{\partial(x,y)}{\partial(u,v)} \mid dudv \tag{6.15}$$
>
> S is the region R transformed to the cartesian u-v plane.

Chapter 7

Applications of mathematical methods in Statistics

The tools we've acquired in this course have applications in many areas of science, but in the MSc in Cognitive Systems we are mainly interested in their applications for statistics and (by extension) data mining.

We begin by considering some facts about random variables. Then we look at how expectation and variance etc. are computed. Several typical probability distributions and their properties are discussed. Two major applications of the mathematics we covered are in maximum likelihood estimation and in the matrix formulation of linear models.

7.1 Discrete random variables; Expectation

A random variable X is a function $X : S \rightarrow \mathbb{R}$ that associates to each outcome $\omega \in S$ exactly one number $X(\omega) = x$.

S_X is all the x's (all the possible values of X, the support of X). I.e., $x \in S_X$.

Good example: number of coin tosses till H

- $X : \omega \rightarrow x$
- ω: H, TH, TTH,… (infinite)
- $x = 0, 1, 2, \dots ; x \in S_X$

Every discrete random variable X has associated with it a **probability mass/distribution function (PDF)**, also called **distribution function**.

$$p_X : S_X \to [0,1] \tag{7.1}$$

defined by

$$p_X(x) = P(X(\omega) = x), x \in S_X \tag{7.2}$$

[**Note**: Books sometimes abuse notation by overloading the meaning of X. They usually have: $p_X(x) = P(X = x), x \in S_X$]

The **cumulative distribution function** is

$$F(a) = \sum_{\text{all } x \leq a} p(x) \tag{7.3}$$

Basic results:

$$E[X] = \sum_{i=1}^{n} x_i p(x_i) \tag{7.4}$$

$$E[g(X)] = \sum_{i=1}^{n} g(x_i) p(x_i) \tag{7.5}$$

$$Var(X) = E[(X - \mu)^2] \tag{7.6}$$

$$Var(X) = E[X^2] - (E[X])^2 \tag{7.7}$$

$$Var(aX + b) = a^2 Var(X) \tag{7.8}$$

$$SD(X) = \sqrt{Var(X)} \tag{7.9}$$

For two independent random variables X and Y,

$$E[XY] = E[X]E[Y] \tag{7.10}$$

Covariance of two random variables:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] \tag{7.11}$$

Note that Cov(X,Y)=0 if X and Y are independent.

Corollary in 4.1 of [7]:

$$E[aX+b] = aE[X]+b \tag{7.12}$$

A related result is about **linear combinations of RVs**:

Theorem. Given two **not necessarily independent** random variables X and Y:

$$E[aX+bY] = aE[X]+bE[Y] \tag{7.13}$$

If X and Y are independent,

$$Var(X+Y) = Var[X]+Var[Y] \tag{7.14}$$

and

$$Var(aX+bY) = a^2Var(X)+b^2Var(Y) \tag{7.15}$$

If $a=1, b=-1$, then

$$Var(X-Y) = Var(X)+Var(Y) \tag{7.16}$$

If X and Y are not independent, then

$$Var(X-Y) = Var(X)+Var(Y)-2Cov(X,Y) \tag{7.17}$$

7.1.1 Examples of discrete probability distributions

7.1.1.1 Geometric

Suppose that independent trials are performed, each with probability p, where $0 < p < 1$, until a success occurs. Let X equal the number of trials required. Then,

$$P(X=n) = (1-p)^{n-1}p \quad n=1,2,\ldots \tag{7.18}$$

Note that:

$$\sum_{x=0}^{\infty} p(1-p)^x = p\sum_{x=0}^{\infty} q^x = p\frac{1}{1-q} = 1.$$

The mean and variance are

$$\mu = \frac{1-p}{p} = \frac{q}{p} \text{ and } \sigma^2 = \frac{q}{p^2}. \tag{7.19}$$

7.1.1.2 Negative binomial

[Taken nearly verbatim from [6].]

Consider the case where we wait for more than one success. Suppose that we conduct Bernoulli trials repeatedly, noting the respective successes and failures. Let X count the number of failures before r successes. If $\mathbb{P}(S) = p$ then X has PMF

$$f_X(x) = \binom{r+x-1}{r-1} p^r (1-p)^x, \quad x = 0, 1, 2, \ldots \tag{7.20}$$

We say that X has a **Negative Binomial distribution** and write $X \sim \mathtt{nbinom}(\mathtt{size} = r, \mathtt{prob} = p)$.

Note that $f_X(x) \geq 0$ and the fact that $\sum f_X(x) = 1$ follows from a generalization of the geometric series by means of a Maclaurin's series expansion:

$$\frac{1}{1-t} = \sum_{k=0}^{\infty} t^k, \quad \text{for } -1 < t < 1, \text{ and} \tag{7.21}$$

$$\frac{1}{(1-t)^r} = \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} t^k, \quad \text{for } -1 < t < 1. \tag{7.22}$$

Therefore

$$\sum_{x=0}^{\infty} f_X(x) = p^r \sum_{x=0}^{\infty} \binom{r+x-1}{r-1} q^x = p^r (1-q)^{-r} = 1, \tag{7.23}$$

since $\mid q \mid = \mid 1-p \mid < 1$.

7.2 Continuous random variables

Recall from the discrete random variables section that: A random variable X is a function $X : S \to \mathbb{R}$ that associates to each outcome $\omega \in S$ exactly one number $X(\omega) = x$. S_X is all the x's (all the possible values of X, the support of X). I.e., $x \in S_X$.

X is a continuous random variable if there is a non-negative function f defined for all real $x \in (-\infty, \infty)$ having the property that for any set B of real numbers,

$$P\{X \in B\} = \int_B f(x)\,dx \tag{7.24}$$

Kerns has the following to add about the above:

> Continuous random variables have supports that look like
>
> $$S_X = [a,b] \text{ or } (a,b), \tag{7.25}$$
>
> or unions of intervals of the above form. Examples of random variables that are often taken to be continuous are:
>
> - the height or weight of an individual,
> - other physical measurements such as the length or size of an object, and
> - durations of time (usually).
>
> Every continuous random variable X has a probability density function (PDF) denoted f_X associated with it that satisfies three basic properties:
>
> 1. $f_X(x) > 0$ for $x \in S_X$,
> 2. $\int_{x \in S_X} f_X(x)\,\mathrm{d}x = 1$, and
> 3. $\mathbb{P}(X \in A) = \int_{x \in A} f_X(x)\,\mathrm{d}x$, for an event $A \subset S_X$.
>
> We can say the following about continuous random variables:
>
> - Usually, the set A in condition 3 above takes the form of an interval, for example, $A = [c,d]$, in which case
>
> $$\mathbb{P}(X \in A) = \int_c^d f_X(x)\,\mathrm{d}x. \tag{7.26}$$
>
> - It follows that the probability that X falls in a given interval is simply the area under the curve of f_X over the interval.

- Since the area of a line $x = c$ in the plane is zero, $\mathbb{P}(X = c) = 0$ for any value c. In other words, the chance that X equals a particular value c is zero, and this is true for any number c. Moreover, when $a < b$ all of the following probabilities are the same:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b). \tag{7.27}$$

- The PDF f_X can sometimes be greater than 1. This is in contrast to the discrete case; every nonzero value of a PMF is a probability which is restricted to lie in the interval $[0,1]$.

$f(x)$ is the probability density function of the random variable X.
Since X must assume some value, f must satisfy

$$1 = P\{X \in (-\infty,\infty)\} = \int_{-\infty}^{\infty} f(x)\,dx \tag{7.28}$$

If $B = [a,b]$, then

$$P\{a \leq X \leq b\} = \int_{a}^{b} f(x)\,dx \tag{7.29}$$

If $a = b$, we get

$$P\{X = a\} = \int_{a}^{a} f(x)\,dx = 0 \tag{7.30}$$

Hence, for any continuous random variable,

$$P\{X < a\} = P\{X \leq a\} = F(a) = \int_{-\infty}^{a} f(x)\,dx \tag{7.31}$$

F is the **cumulative distribution function**. Differentiating both sides in the above equation:

$$\frac{dF(a)}{da} = f(a) \tag{7.32}$$

The density (PDF) is the derivative of the CDF. In the discrete case [6, 128]:

$$f_X(x) = F_X(x) - \lim_{t \to x^-} F_X(t) \tag{7.33}$$

Ross [7] says that it is more intuitive to think about it as follows:

$$P\{a-\frac{\varepsilon}{2} \leq X \leq a+\frac{\varepsilon}{2}\} = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)\,dx \approx \varepsilon f(a) \tag{7.34}$$

when ε is small and when $f(\cdot)$ is continuous. I.e., $\varepsilon f(a)$ is the approximate probability that X will be contained in an interval of length ε around the point a.

Basic results (proofs omitted):

1.
$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx \tag{7.35}$$

2.
$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,dx \tag{7.36}$$

3.
$$E[aX+b] = aE[X]+b \tag{7.37}$$

4.
$$Var[X] = E[(X-\mu)^2] = E[X^2]-(E[X])^2 \tag{7.38}$$

5.
$$Var(aX+b) = a^2 Var(X) \tag{7.39}$$

7.3 Important classes of continuous random variables

7.3.1 Uniform random variable

A random variable (X) with the continuous uniform distribution on the interval (α, β) has PDF

$$f_X(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha < x < \beta, \\ 0, & \text{otherwise} \end{cases} \tag{7.40}$$

The associated R function is $\mathtt{dunif}(\mathtt{min} = a, \mathtt{max} = b)$. We write $X \sim \mathtt{unif}(\mathtt{min} = a, \mathtt{max} = b)$. Due to the particularly simple form of this PDF we can also write down explicitly a formula for the CDF F_X:

$$F_X(a) = \begin{cases} 0, & a < 0, \\ \frac{a-\alpha}{\beta-\alpha}, & \alpha \leq t < \beta, \\ 1, & a \geq \beta. \end{cases} \tag{7.41}$$

$$E[X] = \frac{\beta+\alpha}{2} \tag{7.42}$$

$$Var(X) = \frac{(\beta-\alpha)^2}{12} \tag{7.43}$$

7.3.2 Normal random variable

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty. \tag{7.44}$$

We write $X \sim \mathsf{norm}(\mathtt{mean} = \mu, \mathtt{sd} = \sigma)$, and the associated R function is `dnorm(x, mean = 0, sd = 1)`.

If X is normally distributed with parameters μ and σ^2, then $Y = aX + b$ is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$.

Computing areas under the curve with R:

```
integrate(function(x) dnorm(x, mean = 0, sd = 1),
lower=-Inf,upper=Inf)

## 1 with absolute error < 9.4e-05

## alternatively:
pnorm(Inf)-pnorm(-Inf)

## [1] 1

integrate(function(x) dnorm(x, mean = 0, sd = 1),
          lower=-2,upper=2)

## 0.9544997 with absolute error < 1.8e-11

## alternatively:
pnorm(2)-pnorm(-2)
```

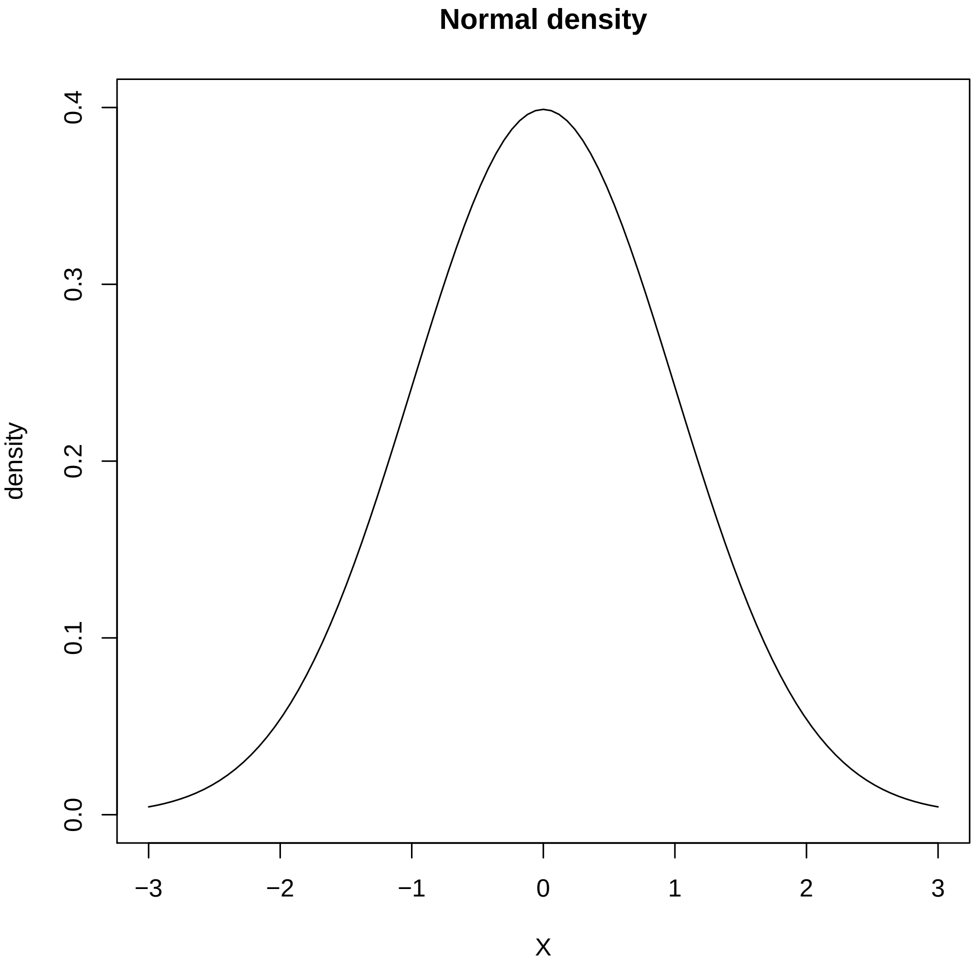


Figure 7.1: Normal distribution.

```
## [1] 0.9544997

integrate(function(x) dnorm(x, mean = 0, sd = 1),
          lower=-1,upper=1)

## 0.6826895 with absolute error < 7.6e-15

## alternatively:
pnorm(1)-pnorm(-1)

## [1] 0.6826895
```

7.3.2.1 Standard or unit normal random variable

If X is normally distributed with parameters μ and σ^2, then $Z = (X - \mu)/\sigma$ is normally distributed with parameters $0, 1$.

We conventionally write $\Phi(x)$ for the CDF:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}}\, dy \quad \text{where} y = (x - \mu)/\sigma \tag{7.45}$$

Neave's tables give the values for positive x; for negative x we do:

$$\Phi(-x) = 1 - \Phi(x), \quad -\infty < x < \infty \tag{7.46}$$

If Z is a standard normal random variable (SNRV) then

$$p\{Z \leq -x\} = P\{Z > x\}, \quad -\infty < x < \infty \tag{7.47}$$

Since $Z = ((X - \mu)/\sigma)$ is an SNRV whenever X is normally distributed with parameters μ and σ^2, then the CDF of X can be expressed as:

$$F_X(a) = P\{X \leq a\} = P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{7.48}$$

The standardized version of a normal random variable X is used to compute specific probabilities relating to X (it's also easier to compute probabilities from different CDFs so that the two computations are comparable).

The expectation of the standard normal random variable:

$$E[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-x^2/2}\,dx$$

Let $u = -x^2/2$.
Then, $du/dx = -2x/2 = -x$. I.e., $du = -x\,dx$ or $-du = x\,dx$.
We can rewrite the integral as:

$$E[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^u x\,dx$$

Replacing $x\,dx$ with $-du$ we get:

$$-\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^u\,du$$

which yields:

$$-\frac{1}{\sqrt{2\pi}}[e^u]_{-\infty}^{\infty}$$

Replacing u with $-x^2/2$ we get:

$$-\frac{1}{\sqrt{2\pi}}[e^{-x^2/2}]_{-\infty}^{\infty} = 0$$

The variance of the standard normal distribution:
We know that

$$\mathrm{Var}(Z) = E[Z^2] - (E[Z])^2$$

Since $(E[Z])^2 = 0$ (see immediately above), we have

$$\mathrm{Var}(Z) = E[Z^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underset{\underset{\text{This is } Z^2.}{\uparrow}}{x^2} e^{-x^2/2}\,dx$$

Write x^2 as $x \times x$ and use integration by parts:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underset{\underset{u}{\uparrow}}{x} \underset{\underset{dv/dx}{\uparrow}}{xe^{-x^2/2}}\,dx = \frac{1}{\sqrt{2\pi}} \underset{\underset{u}{\uparrow}}{x} \underset{\underset{v}{\uparrow}}{-e^{-x^2/2}} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underset{\underset{v}{\uparrow}}{-e^{-x^2/2}} \underset{\underset{du/dx}{\uparrow}}{1}\,dx = 1$$

[Explained on p. 274 of [4]; it wasn't obvious to me, and [7, 200] is pretty terse]: "The first summand above can be shown to equal 0, since as $x \to \pm\infty$,

$e^{-x^2/2}$ gets small more quickly than x gets large. The second summand is just the standard normal density integrated over its domain, so the value of this summand is 1. Therefore, the variance of the standard normal density equals 1."

Example: Given N(10,16), write distribution of $\bar{X}$, where $n = 4$. Since $SE = sd/sqrt(n)$, the distribution of $\bar{X}$ is $N(10, 4/\sqrt{4})$.

7.3.3 Exponential random variables

For some $\lambda > 0$,

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

A continuous random variable with the above PDF is an exponential random variable (or is said to be exponentially distributed).

The CDF:

$$\begin{aligned} F(a) =& P(X \leq a) \\ =& \int_0^a \lambda e^{-\lambda x}\, dx \\ =& \left[-e^{-\lambda x}\right]_0^a \\ =& 1 - e^{-\lambda a} \quad a \geq 0 \end{aligned}$$

[Note: the integration requires the u-substitution: $u = -\lambda x$, and then $du/dx = -\lambda$, and then use $-du = \lambda dx$ to solve.]

7.3.3.1 Expectation and variance of an exponential random variable

For some $\lambda > 0$ (called the rate), if we are given the PDF of a random variable X:

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

Find E[X].

[This proof seems very strange and arbitrary—one starts really generally and then scales down, so to speak. The standard method can equally well be used, but this is more general, it allows for easy calculation of the second moment, for example. Also, it's an example of how reduction formulae are used in integration.]

$$E[X^n] = \int_0^\infty x^n \lambda e^{-\lambda x}\, dx$$

Use integration by parts:

Let $u = x^n$, which gives $du/dx = nx^{n-1}$. Let $dv/dx = \lambda e^{-\lambda x}$, which gives $v = -e^{-\lambda x}$. Therefore:

$$\begin{aligned} E[X^n] &= \int_0^\infty x^n \lambda e^{-\lambda x}\, dx \\ &= \Big[-x^n e^{-\lambda x}\Big]_0^\infty + \int_0^\infty e^{\lambda x} n x^{n-1}\, dx \\ &=0 + \frac{n}{\lambda}\int_0^\infty \lambda e^{-\lambda x} n^{n-1}\, dx \end{aligned}$$

Thus,

$$E[X^n] = \frac{n}{\lambda} E[X^{n-1}]$$

If we let $n = 1$, we get $E[X]$:

$$E[X] = \frac{1}{\lambda}$$

Note that when $n = 2$, we have

$$E[X^2] = \frac{2}{\lambda} E[X] = \frac{2}{\lambda^2}$$

Variance is, as usual,

$$var(X) = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}$$

7.3.4 Weibull distribution

$$f(x \mid \alpha, \beta) = \alpha\beta(\beta x)^{\alpha-1} \exp(-(\beta x)^\alpha) \tag{7.49}$$

When $\alpha = 1$, we have the exponential distribution.

7.3.5 Gamma distribution

[The text is an amalgam of [6] and [7, 215]. I don't put it in double-quotes as a citation because it would look ugly.]

This is a generalization of the exponential distribution. We say that X has a gamma distribution and write $X \sim \mathtt{gamma}(\mathtt{shape} = \alpha, \mathtt{rate} = \lambda)$, where $\alpha > 0$ (called shape) and $\lambda > 0$ (called rate). It has PDF

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

$\Gamma(\alpha)$ is called the gamma function:

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1}\, dy \underset{\substack{\uparrow \\ \text{integration by parts}}}{=} (\alpha-1)\Gamma(\alpha-1)$$

Note that for integral values of n, $\Gamma(n) = (n-1)!$ (follows from above equation).

The associated R functions are `gamma(x, shape, rate = 1)`,`pgamma`, `qgamma`, and `rgamma`, which give the PDF, CDF, quantile function, and simulate random variates, respectively. If $\alpha = 1$ then $X \sim \mathtt{exp}(\mathtt{rate} = \lambda)$. The mean is $\mu = \alpha/\lambda$ and the variance is $\sigma^2 = \alpha/\lambda^2$.

To motivate the gamma distribution recall that if X measures the length of time until the first event occurs in a Poisson process with rate λ then $X \sim \mathtt{exp}(\mathtt{rate} = \lambda)$. If we let Y measure the length of time until the α^{th} event occurs then $Y \sim \mathtt{gamma}(\mathtt{shape} = \alpha, \mathtt{rate} = \lambda)$. When α is an integer this distribution is also known as the **Erlang** distribution.

The Chi-squared distribution is the gamma distribution with $\lambda = 1/2$ and $\alpha = n/2$, where n is an integer:

7.3.5.1 Mean and variance of gamma distribution

Let X be a gamma random variable with parameters α and λ.

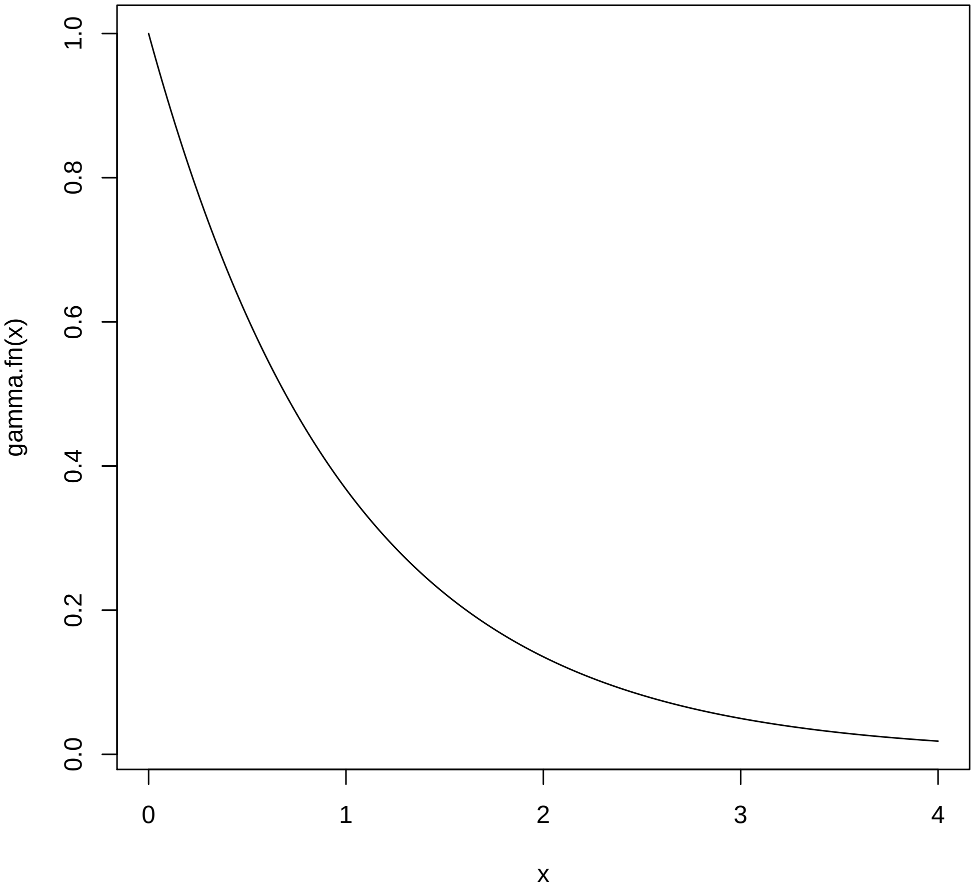


Figure 7.2: The gamma distribution.

.

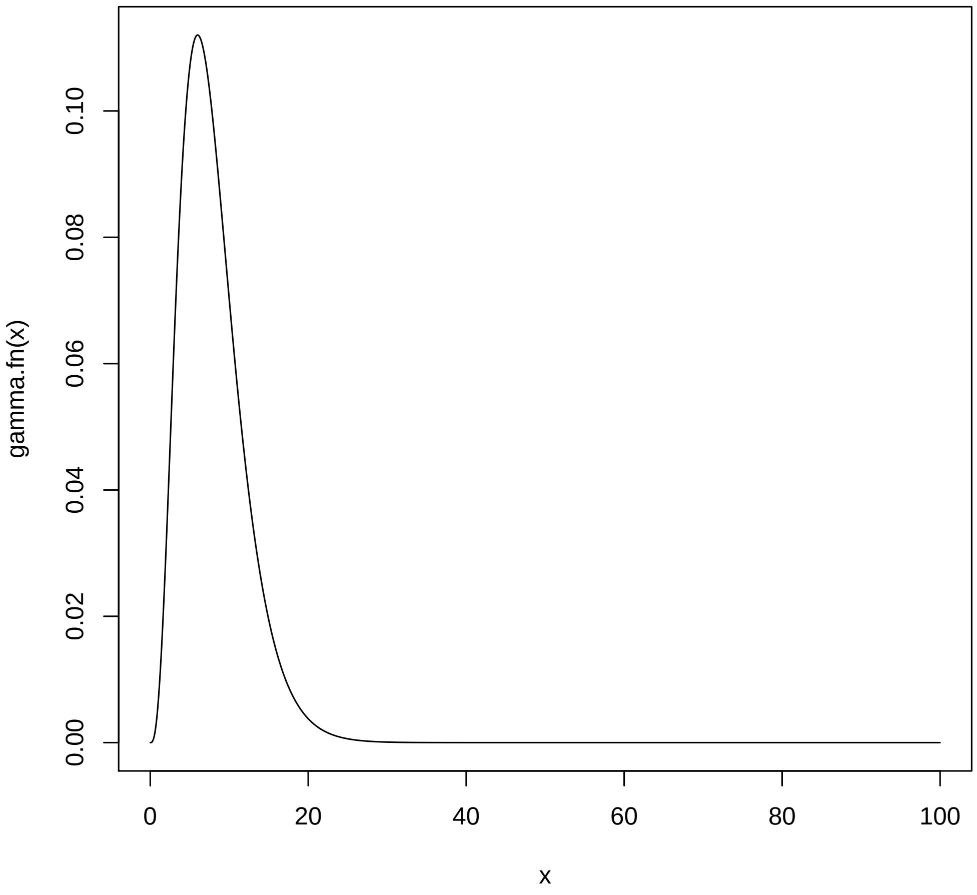


Figure 7.3: The chi-squared distribution.

$$\begin{aligned}
E[X] =& \frac{1}{\Gamma(\alpha)}\int_0^\infty x\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}\,dx \\
=& \frac{1}{\lambda\Gamma(\alpha)}\int_0^\infty e^{-\lambda x}(\lambda x)^{\alpha}\,dx \\
=& \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} \\
=& \frac{\alpha}{\lambda} \quad \text{see derivation of } \Gamma(\alpha), p.\ 215 \text{ of } [7]
\end{aligned}$$

It is easy to show (exercise) that

$$Var(X) = \frac{\alpha}{\lambda^2}$$

7.3.6 Memoryless property (Poisson, Exponential, Geometric)

A nonnegative random variable is memoryless if

$$P(X > s+t) \mid X > t) = P(X > s) \quad \text{for all } s,t \geq 0$$

Two equivalent ways of stating this:

$$\frac{P(X > s+t, X > t)}{P(X > t)} = P(X > s)$$

[just using the definition of conditional probability]
or

$$P(X > s+t) = P(X > s)P(X > t)$$

[not clear yet why the above holds]
Recall definition of conditional probability:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad \text{if } \mathbb{P}(A) > 0.$$

What memorylessness means is: let $s = 10$ and $t = 30$. Then

$$\frac{P(X > 10+30, X \geq 30)}{P(X \geq 30)} = P(X > 10)$$

or

$$P(X > 10+30) = P(X > 10)P(X \geq 30)$$

It does **not** mean:

$$P(X > 10+30 \mid X \geq 30) = P(X > 40)$$

It's easier to see graphically what this means:

7.3.6.1 Examples of memorylessness

Suppose we are given that a discrete random variable X has probability function $\theta^{x-1}(1-\theta)$, where $x = 1, 2, \dots$. Show that

$$P(X > t+a \mid X > a) = \frac{P(X > t+a)}{P(X > a)} \tag{7.50}$$

hence establishing the 'absence of memory' property:

$$P(X > t+a \mid X > a) = P(X > t) \tag{7.51}$$

Proof:

First, restate the pdf given so that it satisfies the definition of a geometric distribution. Let $\theta = 1-p$; then the pdf is

$$(1-p)^{x-1}p \tag{7.52}$$

This is clearly a geometric random variable (see p. 155 of Ross [7]). On p. 156, Ross points out that

$$P(X > a) = (1-p)^a \tag{7.53}$$

[Actually Ross points out that $P(X \geq k) = (1-p)^{k-1}$, from which it follows that $P(X \geq k+1) = (1-p)^k$; and since $P(X \geq k+1) = P(X > k)$, we have $P(X > k) = (1-p)^k$.]

Similarly,

$$P(X > t) = (1-p)^t \tag{7.54}$$

and

$$P(X > t+a) = (1-p)^{t+a} \tag{7.55}$$

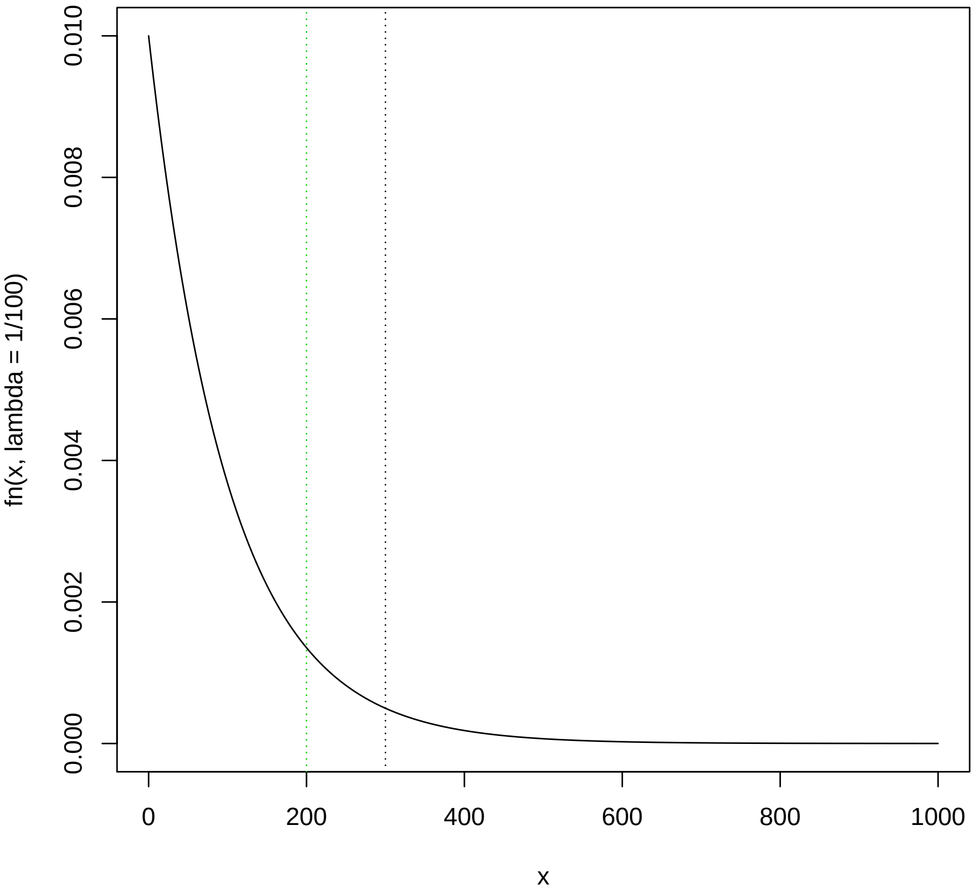


Figure 7.4: The memoryless property of the exponential distribution. The graph after point 300 is an exact copy of the original graph (this is not obvious from the graph, but redoing the graph starting from 300 makes this clear, see figure 7.5 below).

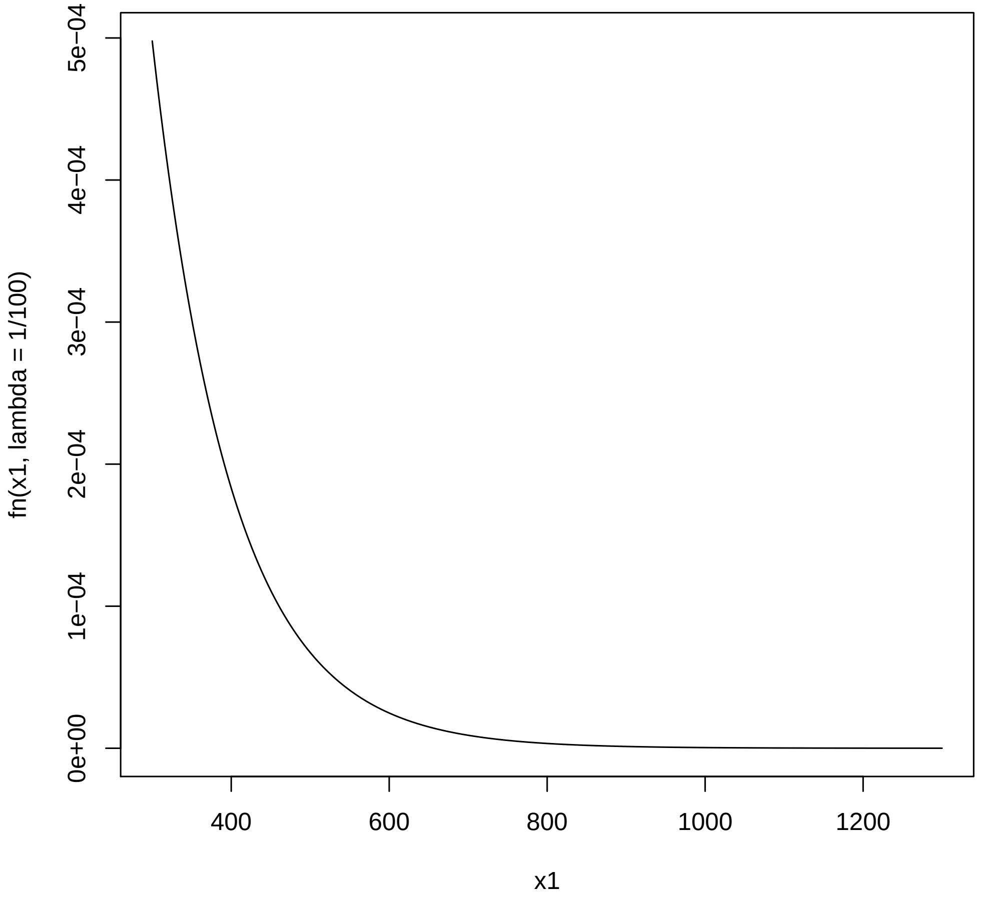


Figure 7.5: Replotting the distribution starting from 300 instead of 0, and extending the x-axis to 1300 instead of 1000 (the number in figure 7.4) gives us an exact copy of original. This is the meaning of the memoryless property of the distribution.

Now, we plug in the values for the right-hand side in equation 7.50, repeated below:

$$P(X>t+a \mid X>a)=\frac{P(X>t+a)}{P(X>a)}=\frac{(1-p)^{t+a}}{(1-p)^{a}}=(1-p)^{t} \tag{7.56}$$

Thus, since $P(X>t)=(1-p)^{t}$ (see above), we have proved that

$$P(X>t+a \mid X>a)=P(X>t) \tag{7.57}$$

This is the definition of memorylessness (equation 5.1 in Ross [7], p. 210). Therefore, we have proved the memorylessness property.

■

7.3.6.2 Prove the memorylessness property for Gamma and Exponential distributions

Exponential:

The CDF is:

$$P(a)=1-e^{-\lambda a} \tag{7.58}$$

Therefore:

$$P(X>s+t)=1-P(s+t)=1-(1-e^{-\lambda(s+t)})=e^{-\lambda(s+t)}=e^{-\lambda s}e^{-\lambda t}=P(X>s)P(X>t) \tag{7.59}$$

The above is the definition of memorylessness.

■

Gamma distribution:

The CDF (not sure how this comes about, see Ross [7]) is

$$F(x;\alpha,\beta)=1-\sum_{i=0}^{\alpha-1}\frac{1}{i!}(\beta x)^{i}e^{-\beta x} \tag{7.60}$$

Therefore,

$$P(X>s+t)=1-P(X<s+t)=1-(1-\sum_{i=0}^{\alpha-1}\frac{1}{i!}(\beta(s+t))^{i}e^{-\beta(s+t)})=\sum_{i=0}^{\alpha-1}\frac{1}{i!}(\beta(s+t))^{i}e^{-\beta(s+t)} \tag{7.61}$$

7.3.7 Beta distribution

This is a generalization of the continuous uniform distribution.

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

There is a connection between the beta and the gamma:

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

which allows us to rewrite the beta PDF as

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1. \tag{7.62}$$

The mean and variance are

$$E[X] = \frac{a}{a+b} \text{ and } Var(X) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{7.63}$$

7.3.8 Distribution of a function of a random variable (transformations of random variables)

A nice and intuitive description:

Consider a continuous RV Y which is a continuous differentiable increasing function of X:

$$Y = g(X) \tag{7.64}$$

Because g is differentiable and increasing, g' and g^{-1} are guaranteed to exist. Because g maps all $x \leq s \leq x + \Delta x$ to $y \leq s \leq y + \Delta y$, we can say:

$$\int_x^{x+\Delta x} f_X(s)\,ds = \int_y^{y+\Delta y} f_Y(t)\,dt \tag{7.65}$$

Therefore, for small Δx:

$$f_Y(y)\Delta y \approx f_X(x)\Delta x \tag{7.66}$$

Dividing by Δy we get:

$$f_Y(y) \approx f_X(x)\frac{\Delta x}{\Delta y} \tag{7.67}$$

Theorem 2 (Theorem 7.1 in Ross [7]). *Let X be a continuous random variable having probability density function f_X. Suppose that $g(x)$ is a strict monotone (increasing or decreasing) function, differentiable and (thus continuous) function of x. Then the random variable Y defined by $Y = g(X)$ has a probability density function defined by*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \mid \frac{d}{dx}g^{-1}(y) \mid & \text{if } y = g(x) \text{ for some } x \\ 0 & \text{if } y \neq g(x) \text{ for all } x. \end{cases}$$

where $g^{-1}(y)$ is defined to be equal to the value of x such that $g(y-y)$.

Proof:
Suppose $y = g(x)$ for some x. Then, with $Y = g(X)$,

$$\begin{aligned} F_Y(y) =& P(g(X) \leq y) \\ =& P(X \leq g^{-1}(y)) \\ =& F_X(g^{-1}(y)) \end{aligned} \tag{7.68}$$

Differentiation gives

$$f_Y(y) = f_X(g^{-1}(y))\frac{d(g^{-1}(y))}{dy} \tag{7.69}$$

Detailed explanation for the above equation: Since

$$\begin{aligned} F_Y(y) =& F_X(g^{-1}(y)) \\ =& \int f_X(g^{-1}(y))\, dy \end{aligned} \tag{7.70}$$

Differentiating:

$$\frac{d(F_Y(y))}{dy} = \frac{d}{dy}(F_X(g^{-1}(y))) \tag{7.71}$$

We use the chain rule. To simplify things, rewrite $w(y) = g^{-1}(y)$ (otherwise typesetting things gets harder). Then, let

$$u = w(y)$$

which gives

$$\frac{du}{dy} = w'(y)$$

and let

$$x = F_X(u)$$

This gives us

$$\frac{dx}{du} = F_X'(u) = f_X(u)$$

By the chain rule:

$$\frac{du}{dy} \times \frac{dx}{du} = w'(y) f_X(u) \underset{\substack{\uparrow \\ \text{plugging in the variables}}}{=} \frac{d}{dy}(g^{-1}(y)) f_X(g^{-1}(y))$$

■

Exercises:

1. $Y = X^2$
2. $Y = \sqrt{X}$
3. $Y =\mid X \mid$
4. $Y = aX + b$

7.3.9 The Poisson distribution

As Kerns [6] puts it (I quote him nearly exactly, up to the definition):

> This is a distribution associated with “rare events”, for reasons which will become clear in a moment. The events might be:
>
> - traffic accidents,
> - typing errors, or
> - customers arriving in a bank.
>
> Let λ be the average number of events in the time interval $[0,1]$. Let the random variable X count the number of events occurring in the interval. Then under certain reasonable conditions it can be shown that
>
> $$f_X(x) = \mathbb{P}(X = x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0,1,2,\ldots \tag{7.72}$$

7.3.9.1 Poisson conditional probability and binomial

If $X_1 \sim \mathit{Pois}(\lambda_1)$ and $X_2 \sim \mathit{Pois}(\lambda_2)$ are independent and $Y = X_1 + X_2$, then the distribution of X_1 conditional on $Y = y$ is a binomial. Specifically, $X_1 \mid Y = y \sim \mathit{Binom}(y,\lambda_1)/(\lambda_1,\lambda_2)$. More generally, if $X_1, X_2, \ldots, X_n$ are independent Poisson random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ then

$$X_i \mid \sum_{j=1}^{n} X_j \sim \mathit{Binom}(\sum_{j=1}^{n} X_j, \frac{\lambda_i}{\underset{j=1}{\overset{n}{\lambda_j}}}) \tag{7.73}$$

[Source for above: wikipedia]. Relevant for q3 in P-Ass 5.

To see why this is true, see p. 173 of Dekking et al. Also see the stochastic processes book.

7.3.10 Geometric distribution [discrete]

From Ross [7, 155]:

> Let independent trials, each with probability p, $0 < p < 1$ of success, be performed until a success occurs. If X is the number of trials required till success occurs, then
>
> $$P(X = n) = (1-p)^{n-1} p \quad n = 1, 2, \ldots$$
>
> I.e., for X to equal n, it is necessary and sufficient that the first $n-1$ are failures, and the nth trial is a success. The above equation comes about because the successive trials are independent.

X is a geometric random variable with parameter p.
Note that a success will occur, with probability 1:

$$\sum_{i=1}^{\infty} P(X = n) = p \sum_{i=1}^{\infty} (1-p)^{n-1} \underset{\substack{\uparrow \\ \text{see geometric series section.}}}{=} \frac{p}{1-(1-p)} = 1$$

7.3.10.1 Mean and variance of the geometric distribution

$$E[X] = \frac{1}{p}$$

$$Var(X) = \frac{1-p}{p^2}$$

For proofs, see Ross [7, 156-157].

7.3.11 Normal approximation of the binomial and poisson

Excellent explanation available at:

`http://www.johndcook.com/normal_approx_to_poisson.html`

If $P(X = n)$ use $P(n - 0.5 < X < n + 0.5)$
If $P(X > n)$ use $P(X > n + 0.5)$
If $P(X \leq n)$ use $P(X < n + 0.5)$
If $P(X < n)$ use $P(X < n - 0.5)$
If $P(X \geq n)$ use $P(X > n - 0.5)$

7.4 Jointly distributed random variables

7.4.1 Joint distribution functions

7.4.1.1 Discrete case

[This section is an extract from [6].]

Consider two discrete random variables X and Y with PMFs f_X and f_Y that are supported on the sample spaces S_X and S_Y, respectively. Let $S_{X,Y}$ denote the set of all possible observed **pairs** (x,y), called the **joint support set** of X and Y. Then the **joint probability mass function** of X and Y is the function $f_{X,Y}$ defined by

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y), \quad \text{for } (x,y) \in S_{X,Y}. \tag{7.74}$$

Every joint PMF satisfies

$$f_{X,Y}(x,y) > 0 \text{ for all } (x,y) \in S_{X,Y}, \tag{7.75}$$

and

$$\sum_{(x,y)\in S_{X,Y}} f_{X,Y}(x,y) = 1. \tag{7.76}$$

It is customary to extend the function $f_{X,Y}$ to be defined on all of $\mathbb{R}^2$ by setting $f_{X,Y}(x,y) = 0$ for $(x,y) \notin S_{X,Y}$.

In the context of this chapter, the PMFs f_X and f_Y are called the **marginal PMFs** of X and Y, respectively. If we are given only the joint PMF then we may recover each of the marginal PMFs by using the Theorem of Total Probability: observe

$$\begin{aligned} f_X(x) &= \mathbb{P}(X = x), && (7.77)\\ &= \sum_{y\in S_Y} \mathbb{P}(X = x, Y = y), && (7.78)\\ &= \sum_{y\in S_Y} f_{X,Y}(x,y). && (7.79) \end{aligned}$$

By interchanging the roles of X and Y it is clear that

$$f_Y(y) = \sum_{x\in S_X} f_{X,Y}(x,y). \tag{7.80}$$

Given the joint PMF we may recover the marginal PMFs, but the converse is not true. Even if we have **both** marginal distributions they are not sufficient to determine the joint PMF; more information is needed.

Associated with the joint PMF is the **joint cumulative distribution function** $F_{X,Y}$ defined by

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y), \quad \text{for } (x,y) \in \mathbb{R}^2.$$

The bivariate joint CDF is not quite as tractable as the univariate CDFs, but in principle we could calculate it by adding up quantities of the form in Equation 7.74. The joint CDF is typically not used in practice due to its inconvenient form; one can usually get by with the joint PMF alone.

Examples from [6]: **Example 1**:
Roll a fair die twice. Let X be the face shown on the first roll, and let Y be the face shown on the second roll. For this example, it suffices to define

$$f_{X,Y}(x,y) = \frac{1}{36}, \quad x = 1,\ldots,6,\ y = 1,\ldots,6.$$

The marginal PMFs are given by $f_X(x) = 1/6, x = 1,2,\ldots,6$, and $f_Y(y) = 1/6$, $y = 1,2,\ldots,6$, since

$$f_X(x) = \sum_{y=1}^{6} \frac{1}{36} = \frac{1}{6}, \quad x = 1,\ldots,6,$$

and the same computation with the letters switched works for Y.
Here, and in many other ones, the joint support can be written as a product set of the support of X "times" the support of Y, that is, it may be represented as a cartesian product set, or rectangle, $S_{X,Y} = S_X \times S_Y$, where $S_X \times S_Y = \{(x,y) : x \in S_X, y \in S_Y\}$. This form is a necessary condition for X and Y to be **independent** (or alternatively **exchangeable** when $S_X = S_Y$). But please note that in general it is not required for $S_{X,Y}$ to be of rectangle form.
Example 2: very involved example in [6], worth study.

7.4.1.2 Continuous case

For random variables X and y, the **joint cumulative pdf** is

$$F(a,b) = P(X \leq a, Y \leq b) \quad -\infty < a,b < \infty \tag{7.81}$$

The **marginal distributions** of F_X and F_Y are the CDFs of each of the associated RVs:

1. The CDF of X:

$$F_X(a) = P(X \leq a) = F_X(a, \infty) \tag{7.82}$$

2. The CDF of Y:

$$F_Y(a) = P(Y \leq b) = F_Y(\infty, b) \tag{7.83}$$

Definition 1. ***Jointly continuous****: Two RVs X and Y are jointly continuous if there exists a function $f(x,y)$ defined for all real x and y, such that for every set C:*

$$P((X,Y) \in C) = \iint\limits_{(x,y)\in C} f(x,y)\,dx\,dy \tag{7.84}$$

$f(x,y)$ is the ***joint PDF*** *of X and Y.*
Every joint PDF satisfies

$$f(x,y) \geq 0 \textit{ for all } (x,y) \in S_{X,Y}, \tag{7.85}$$

and

$$\iint\limits_{S_{X,Y}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1. \tag{7.86}$$

For any sets of real numbers A and B, and if $C = \{(x,y) : x \in A, y \in B\}$, it follows from equation 7.84 that

$$P((X \in A, Y \in B) \in C) = \int_B \int_A f(x,y)\,dx\,dy \tag{7.87}$$

Note that

$$F(a,b) = P(X \in (-\infty, a]), Y \in (-\infty, b])) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)\,dx\,dy \tag{7.88}$$

Differentiating, we get the joint pdf:

$$f(a,b) = \frac{\partial^2}{\partial a \partial b} F(a,b) \tag{7.89}$$

One way to understand the joint PDF:

$$P(a<X<a+da, b<Y<b+db) = \int_b^{d+db} \int_a^{a+da} f(x,y)\,dx\,dy \approx f(a,b)dadb \tag{7.90}$$

Hence, $f(x,y)$ is a measure of how probable it is that the random vector (X,Y) will be near (a,b).

7.4.1.3 Marginal probability distribution functions

If X and Y are jointly continuous, they are individually continuous, and their PDFs are:

$$\begin{aligned} P(X\in A) =& P(X\in A, Y\in(-\infty,\infty)) \\ =& \int_A \int_{-\infty}^{\infty} f(x,y)\,dy\,dx \\ =& \int_A f_X(x)\,dx \end{aligned} \tag{7.91}$$

where

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \tag{7.92}$$

Similarly:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx \tag{7.93}$$

7.4.1.4 Independent random variables

Random variables X and Y are independent iff, for any two sets of real numbers A and B:

$$P(X\in A, Y\in B) = P(X\in A)P(Y\in B) \tag{7.94}$$

In the jointly continuous case:

$$f(x,y) = f_X(x)f_Y(y) \quad \text{for all } x,y \tag{7.95}$$

A necessary and sufficient condition for the random variables X and Y to be independent is for their joint probability density function (or joint probability mass function in the discrete case) $f(x,y)$ to factor into two terms, one depending only on x and the other depending only on y. This can be stated as a proposition:

Proposition 1.

Easy-to-understand example from [6]: Let the joint PDF of (X,Y) be given by

$$f_{X,Y}(x,y) = \frac{6}{5}\left(x+y^2\right), \quad 0 < x < 1,\ 0 < y < 1.$$

The marginal PDF of X is

$$\begin{aligned} f_X(x) &= \int_0^1 \frac{6}{5}\left(x+y^2\right)\,\mathrm{d}y, \\ &= \frac{6}{5}\left(xy+\frac{y^3}{3}\right)\Bigg|_{y=0}^{1}, \\ &= \frac{6}{5}\left(x+\frac{1}{3}\right), \end{aligned}$$

for $0 < x < 1$, and the marginal PDF of Y is

$$\begin{aligned} f_Y(y) &= \int_0^1 \frac{6}{5}\left(x+y^2\right)\,\mathrm{d}x, \\ &= \frac{6}{5}\left(\frac{x^2}{2}+xy^2\right)\Bigg|_{x=0}^{1}, \\ &= \frac{6}{5}\left(\frac{1}{2}+y^2\right), \end{aligned}$$

for $0 < y < 1$.
In this example the joint support set was a rectangle $[0,1]\times[0,1]$, but it turns out that X and Y are not independent. This is because $\frac{6}{5}\left(x+y^2\right)$ cannot be stated as a product of two terms ($f_X(x)f_Y(y)$).

7.4.1.5 Sums of independent random variables

[Taken nearly verbatim from Ross.]

Suppose that X and Y are independent, continuous random variables having probability density functions f_X and f_Y. The cumulative distribution function of $X+Y$ is obtained as follows:

$$\begin{aligned}
F_{X+Y}(a) =& P(X+Y \leq a) \\
=& \iint\limits_{x+y\leq a} f_{XY}(x,y)\,dx\,dy \\
=& \iint\limits_{x+y\leq a} f_X(x) f_Y(y)\,dx\,dy \\
=& \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x) f_Y(y)\,dx\,dy \\
=& \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)\,dx f_Y(y)\,dy \\
=& \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\,dy
\end{aligned} \tag{7.96}$$

The CDF F_{X+Y} is the **convolution** of the distributions F_X and F_Y.
If we differentiate the above equation, we get the pdf f_{X+Y}:

$$\begin{aligned}
f_{X+Y} =& \frac{d}{dx}\int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\,dy \\
=& \int_{-\infty}^{\infty} \frac{d}{dx} F_X(a-y) f_Y(y)\,dy \\
=& \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\,dy
\end{aligned} \tag{7.97}$$

7.4.2 Conditional distributions

7.4.2.1 Discrete case

Recall that the conditional probability of B given A, denoted $\mathbb{P}(B \mid A)$, is defined by

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad \text{if } \mathbb{P}(A) > 0. \tag{7.98}$$

If X and Y are discrete random variables, then we can define the conditional PMF of X given that $Y = y$ as follows:

$$\begin{aligned} p_{X|Y}(x \mid y) =& P(X = x \mid Y = y) \\ =& \frac{P(X = x, Y = y)}{P(Y = y)} \\ =& \frac{p(x,y)}{p_Y(y)} \end{aligned} \tag{7.99}$$

for all values of y where $p_Y(y) = P(Y = y) > 0$.

The **conditional cumulative distribution function** of X given $Y = y$ is defined, for all y such that $p_Y(y) > 0$, as follows:

$$\begin{aligned} F_{X|Y} =& P(X \leq x \mid Y = y) \\ =& \sum_{a \leq x} p_{X|Y}(a \mid y) \end{aligned} \tag{7.100}$$

If X and Y are independent then

$$p_{X|Y}(x \mid y) = P(X = x) = p_X(x) \tag{7.101}$$

See the examples starting p. 264 of Ross.

An important thing to understand is the phrasing of the question (e.g., in P-Ass3): “Find the conditional distribution of X given all the possible values of Y”.

7.4.2.2 Continuous case

[Taken almost verbatim from Ross.]

If X and Y have a joint probability density function $f(x,y)$, then the conditional probability density function of X given that $Y = y$ is defined, for all values of y such that $f_Y(y) > 0$,by

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)} \tag{7.102}$$

We can understand this definition by considering what $f_{X|Y}(x \mid y)\,dx$ amounts to:

$$
\begin{aligned}
f_{X|Y}(x \mid y)\,dx =& \frac{f(x,y)}{f_Y(y)}\frac{dxdy}{dy} \\
=& \frac{f(x,y)dxdy}{f_Y(y)dy} \\
=& \frac{P(x < X < d + dx, y < Y < y + dy)}{y < P < y + dy}
\end{aligned}
\tag{7.103}
$$

7.4.3 Joint and marginal expectation

[Taken nearly verbatim from [6].]

Given a function g with arguments (x,y) we would like to know the long-run average behavior of $g(X,Y)$ and how to mathematically calculate it. Expectation in this context is computed by integrating (summing) with respect to the joint probability density (mass) function.

Discrete case:

$$\mathbb{E}\,g(X,Y) = \sum\sum_{(x,y)\in S_{X,Y}} g(x,y)\,f_{X,Y}(x,y). \tag{7.104}$$

Continuous case:

$$\mathbb{E}\,g(X,Y) = \iint\limits_{S_{X,Y}} g(x,y)\,f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y, \tag{7.105}$$

7.4.3.1 Covariance and correlation

There are two very special cases of joint expectation: the **covariance** and the **correlation**. These are measures which help us quantify the dependence between X and Y.

Definition 2. *The* ***covariance*** *of X and Y is*

$$Cov(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y). \tag{7.106}$$

Shortcut formula for covariance:

$$\operatorname{Cov}(X,Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y). \tag{7.107}$$

The **Pearson product moment correlation** between X and Y is the covariance between X and Y rescaled to fall in the interval $[-1,1]$. It is formally defined by

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{7.108}$$

The correlation is usually denoted by $\rho_{X,Y}$ or simply ρ if the random variables are clear from context. There are some important facts about the correlation coefficient:

1. The range of correlation is $-1 \leq \rho_{X,Y} \leq 1$.

2. Equality holds above ($\rho_{X,Y} = \pm 1$) if and only if Y is a linear function of X with probability one.

Discrete example: to-do

Continuous example from [6]: Let us find the covariance of the variables (X,Y) from an example numbered 7.2 in Kerns. The expected value of X is

$$\mathbb{E}X = \int_0^1 x \cdot \frac{6}{5}\left(x + \frac{1}{3}\right) \mathrm{d}x = \frac{2}{5}x^3 + \frac{1}{5}x^2 \bigg|_{x=0}^{1} = \frac{3}{5},$$

and the expected value of Y is

$$\mathbb{E}Y = \int_0^1 y \cdot \frac{6}{5}\left(\frac{1}{2} + y^2\right) \mathrm{d}x = \frac{3}{10}y^2 + \frac{3}{20}y^4 \bigg|_{y=0}^{1} = \frac{9}{20}.$$

Finally, the expected value of XY is

$$\begin{aligned}
\mathbb{E}XY &= \int_0^1 \int_0^1 xy\frac{6}{5}\left(x+y^2\right)\mathrm{d}x\,\mathrm{d}y, \\
&= \int_0^1 \left(\frac{2}{5}x^3y + \frac{3}{10}xy^4\right)\bigg|_{x=0}^{1} \mathrm{d}y, \\
&= \int_0^1 \left(\frac{2}{5}y + \frac{3}{10}y^4\right)\mathrm{d}y, \\
&= \frac{1}{5} + \frac{3}{50},
\end{aligned}$$

which is 13/50. Therefore the covariance of (X,Y) is

$$\mathrm{Cov}(X,Y) = \frac{13}{50} - \left(\frac{3}{5}\right)\left(\frac{9}{20}\right) = -\frac{1}{100}.$$

7.4.4 Conditional expectation

Recall that

$$f_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \tag{7.109}$$

for all y such that $P(Y = y) > 0$.

It follows that

$$\begin{aligned} E[X \mid Y = y] =& \sum_x xP(X = x \mid Y = y) \\ =& \sum_x x p_{X|Y}(x \mid y) \end{aligned} \tag{7.110}$$

$E[X \mid Y]$ is that **function** of the random variable Y whose value at $Y = y$ is $E[X \mid Y = y]$. $E[X \mid Y]$ is a random variable.

7.4.4.1 Relationship to 'regular' expectation

Conditional expectation given that $Y = y$ can be thought of as being an ordinary expectation on a reduced sample space consisting only of outcomes for which $Y = y$. All properties of expectations hold. Two examples:

Example 1:

$$E[g(X) \mid Y = y] = \begin{cases} \sum_x g(x) p_{X|Y}(x,y) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\, dx & \text{in the continuous case} \end{cases}$$

Example 2:

$$E\left[\sum_{i=1}^{n} X_i \mid Y = y\right] = \sum_{i=1}^{n} E[X_i \mid Y = y] \tag{7.111}$$

Proposition 2. ***Expectation of the conditional expectation***

$$E[X] = E[E[X \mid Y]] \tag{7.112}$$

If Y is a discrete random variable, then the above proposition states that

$$E[X] = \sum_y E[X \mid Y = y] P(Y = y) \tag{7.113}$$

7.4.5 Multinomial coefficients and multinomial distributions

[Taken almost verbatim from [6], with some additional stuff from Ross.]

We sample n times, with replacement, from an urn that contains balls of k different types. Let X_1 denote the number of balls in our sample of type 1, let X_2 denote the number of balls of type 2, ..., and let X_k denote the number of balls of type k. Suppose the urn has proportion p_1 of balls of type 1, proportion p_2 of balls of type 2, ..., and proportion p_k of balls of type k. Then the joint PMF of $(X_1,\ldots,X_k)$ is

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) \quad = \quad \binom{n}{x_1\,x_2\,\cdots\,x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \tag{7.114}$$

for $(x_1,\ldots,x_k)$ in the joint support $S_{X_1,\ldots X_K}$. We write

$$(X_1,\ldots,X_k) \sim \mathsf{multinom}(\mathtt{size} = n, \mathtt{prob} = \mathbf{p}_{\mathrm{k}\times 1}). \tag{7.115}$$

Note:

First, the joint support set $S_{X_1,\ldots X_K}$ contains all nonnegative integer k-tuples $(x_1,\ldots,x_k)$ such that $x_1 + x_2 + \cdots + x_k = n$. A support set like this is called a *simplex*. Second, the proportions p_1, p_2, ..., p_k satisfy $p_i \geq 0$ for all i and $p_1 + p_2 + \cdots + p_k = 1$. Finally, the symbol

$$\binom{n}{x_1\,x_2\,\cdots\,x_k} = \frac{n!}{x_1!\,x_2!\,\cdots x_k!} \tag{7.116}$$

is called a *multinomial coefficient* which generalizes the notion of a binomial coefficient.

Example from Ross:

Suppose a fair die is rolled nine times. The probability that 1 appears three times, 2 and 3 each appear twice, 4 and 5 each appear once, and 6 not at all, can be computed using the multinomial distribution formula. Here, for $i = 1, \dots, 6$, it is clear that $p_i == \frac{1}{6}$. And it is clear that $n = 9$, and $x_1 = 3$, $x_2 = 2$, $x_3 = 2$, $x_4 = 1$, $x_5 = 1$, and $x_6 = 0$. We plug in the values into the formula:

$$f_{X_1,\dots,X_k}(x_1,\dots,x_k) = \binom{n}{x_1 x_2 \cdots x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \qquad (7.117)$$

Plugging in the values:

$$f_{X_1,\dots,X_k}(x_1,\dots,x_k) = \binom{9}{3\,2\,2\,1\,1\,0} \frac{1}{6}^3 \frac{1}{6}^2 \frac{1}{6}^2 \frac{1}{6}^1 \frac{1}{6}^1 \frac{1}{6}^0 \qquad (7.118)$$

Answer: $\frac{9!}{3!2!2!}\left(\frac{1}{6}\right)^9$

7.4.6 Multivariate normal distributions

Recall that in the univariate case:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2} e\{-\frac{(\frac{(x-\mu)}{\sigma})^2}{2}\}} \quad -\infty < x < \infty \qquad (7.119)$$

We can write the power of the exponential as:

$$(\frac{(x-\mu)}{\sigma})^2 = (x-\mu)(x-\mu)(\sigma^2)^{-1} = (x-\mu)(\sigma^2)^{-1}(x-\mu) = Q \qquad (7.120)$$

Generalizing this to the multivariate case:

$$Q = (x-\mu)'\Sigma^{-1}(x-\mu) \qquad (7.121)$$

So, for multivariate case:

$$f(x) = \frac{1}{\sqrt{2\pi det\Sigma} e\{-Q/2\}} \quad -\infty < x_i < \infty, i = 1, \dots, n \qquad (7.122)$$

Properties of normal MVN X:

- Linear combinations of X are normal distributions.
- All subset's of X's components have a normal distribution.
- Zero covariance implies independent distributions.
- Conditional distributions are normal.

7.5 Application of differentiation: Method of maximum likelihood estimation

For this section, note that if we write the equation $f'(x) = 0$, and solve for x, we are basically trying to find out what the value of x is when the slope is 0. That is exactly the situation when the function f "turns", i.e., is at a maximum or minimum.

Next, we look at an application of differentation that will come up again and again. In statistics, we look at the sample values and then choose as our estimates of the unknown parameters the values for which the probability or probability density of getting the sample values is a maximum.

Discrete case: Suppose the observed sample values are $x_1, x_2, \dots, x_n$. The probability of getting them is

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = f(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta) \quad (7.123)$$

i.e., the function f is the value of the joint probability **distribution** of the random variables $X_1, \dots, X_n$ at $X_1 = x_1, \dots, X_n = x_n$. Since the sample values have been observed and are fixed, $f(x_1, \dots, x_n; \theta)$ is a function of θ. The function f is called a **likelihood function**.

Continuous case

Here, f is the joint probability **density**, the rest is the same as above.

Definition 3. *If $x_1, x_2, \dots, x_n$ are the values of a random sample from a population with parameter θ, the* ***likelihood function*** *of the sample is given by*

$$L(\theta) = f(x_1, x_2, \dots, x_n; \theta) \quad (7.124)$$

for values of θ within a given domain. Here, $f(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta)$ is the joint probability distribution or density of the random variables $X_1, \ldots, X_n$ at $X_1 = x_1, \ldots, X_n = x_n$.

So, the method of maximum likelihood consists of maximizing the likelihood function with respect to θ. The value of θ that maximizes the likelihood function is the **MLE** (maximum likelihood estimate) of θ.

Example: The likelihood function in the binomial case:

$$L(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{7.125}$$

Log likelihood:

$$\ell(\theta) = \log \binom{n}{x} + x \log \theta + (n-x)\log(1-\theta) \tag{7.126}$$

Differentiating and equating to zero to get the maximum:

$$\ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \tag{7.127}$$

How to get the second term: let $u = 1-\theta$.

Then, $du/d\theta = -1$. Now, $y = (n-x)\log(1-\theta)$ can be rewritten in terms of u: $y = (n-x)\log(u)$. So, $dy/du = \frac{n-x}{u}$.

Now, by the chain rule, $dy/d\theta = dy/du \times du/d\theta = \frac{n-x}{u} \times (-1) = -\frac{n-x}{1-\theta}$.

Rearranging terms, we get:

$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \Leftrightarrow \frac{x}{\theta} = \frac{n-x}{1-\theta} \Leftrightarrow \hat{\theta} = \frac{x}{n}$

7.5.0.1 Finding maximum likelihood estimates for different distributions

7.5.0.1.1 Example 1 Let X_i, $i = 1, \ldots, n$ be a random variable with PDF $f(x;\sigma) = \frac{1}{2\sigma}\exp(-\frac{|x|}{\sigma})$. Find $\hat{\sigma}$, the MLE of σ.

$$L(\sigma) = \prod f(x_i;\sigma) = \frac{1}{(2\sigma)^n}\exp(-\sum \frac{|x_i|}{\sigma}) \tag{7.128}$$

$$\ell(x;\sigma) = \sum \left[-\log 2 - \log \sigma - \frac{|x_i|}{\sigma} \right] \tag{7.129}$$

Differentiating and equating to zero to find maximum:

$$\ell'(\sigma) = \sum \left[-\frac{1}{\sigma} + \frac{|x_i|}{\sigma^2} \right] = -\frac{n}{\sigma} + \frac{|x_i|}{\sigma^2} = 0 \tag{7.130}$$

Rearranging the above, the MLE for σ is:

$$\hat{\sigma} = \frac{\sum \mid x_i \mid}{n} \tag{7.131}$$

7.5.0.1.2 Exponential

$$f(x;\lambda) = \lambda \exp(-\lambda x) \tag{7.132}$$

Log likelihood:

$$\ell = n \log \lambda - \sum \lambda x_i \tag{7.133}$$

Differentiating and equating to zero:

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum x_i = 0 \tag{7.134}$$

$$\frac{n}{\lambda} = \sum x_i \tag{7.135}$$

I.e.,

$$\frac{1}{\hat{\lambda}} = \frac{\sum x_i}{n} \tag{7.136}$$

7.5.0.1.3 Poisson

$$L(\mu;x) \qquad = \prod \tfrac{\exp^{-\mu} \mu^{x_i}}{x_i!} \tag{7.137}$$

$$= \exp^{-\mu} \mu^{\sum x_i} \tfrac{1}{\prod x_i!} \tag{7.138}$$

Log likelihood:

$$\ell(\mu;x) = -n\mu + \sum x_i \log \mu - \sum \log y! \tag{7.139}$$

Differentiating and equating to zero:

$$\ell'(\mu) = -n + \frac{\sum x_i}{\mu} = 0 \tag{7.140}$$

Therefore:

$$\hat{\lambda} = \frac{\sum x_i}{n} \tag{7.141}$$

7.5.0.1.4 Geometric

$$f(x;p) = (1-p)^{x-1}p \tag{7.142}$$

$$L(p) = p^n(1-p)^{\sum x - n} \tag{7.143}$$

Log likelihood:

$$\ell(p) = n\log p + (\sum x - n)\log(1-p) \tag{7.144}$$

Differentiating and equating to zero:

$$\ell'(p)\frac{n}{p} - \frac{\sum x - n}{1-p} = 0 \tag{7.145}$$

$$\hat{p} = \frac{1}{\bar{x}} \tag{7.146}$$

7.5.0.1.5 Normal Let $X_1, \ldots, X_n$ constitute a random variable of size n from a normal population with mean μ and variance σ^2, find joint maximum likelihood estimates of these two parameters.

$$\begin{aligned} L(\mu;\sigma^2) &= \prod N(x_i;\mu,\sigma) && (7.147)\\ &= (\tfrac{1}{\sigma\sqrt{2\pi}})^n \exp(-\tfrac{1}{2\sigma^2}\sigma(x_i-\mu)^2) && (7.148)\\ & && (7.149)\end{aligned}$$

Taking logs and differentiating with respect to μ and σ, we get:

$$\hat{\mu} = \frac{1}{n}\sum x_i = \bar{x} \tag{7.150}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 \tag{7.151}$$

7.6 Application of matrices: Least squares estimation

[Note: I draw heavily from [1], chapters 1 and 4; [9]; and [2].]

Consider the goal of estimating the intercept and slope β_0 and β_1 (which are unknown population parameters) from a sample:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{7.152}$$

Let's call the estimates of these parameters $\hat{\beta}_0$ and $\hat{\beta}_1$; $\hat{Y}$ is the predicted value of the dependent variable.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{7.153}$$

We want to find those beta-hats where the sum of squares of deviation are minimized. I.e., squaring and rearranging (7.152) we get:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{7.154}$$

If we differentiate with respect to β_0 and then β_1, we get:

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \tag{7.155}$$

and

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) \tag{7.156}$$

this gives us the beta-hats through the two equations:

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{7.157}$$

$$\sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{7.158}$$

Rearranging equations (7.157) and (7.158) we get the **normal equations**:

$$\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \tag{7.159}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} {X_i}^2 = \sum_{i=1}^{n} X_i Y_i \tag{7.160}$$

The above equations (7.159) and (7.160) are used to solve for the beta-hats.

Next, we show how to state the normal equations in matrix form.

Our linear model equation is a system of i equations. For each i, we can write the single equation:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{7.161}$$

can be expanded to:

$$\begin{array}{ccccccc} Y_1 & = & \beta_0 & + & X_1\beta_1 & + & \varepsilon_1 \\ Y_2 & = & \beta_0 & + & X_2\beta_1 & + & \varepsilon_2 \\ Y_3 & = & \beta_0 & + & X_3\beta_1 & + & \varepsilon_3 \\ Y_4 & = & \beta_0 & + & X_4\beta_1 & + & \varepsilon_4 \\ \vdots & & \vdots & & \vdots & & \vdots \\ Y_n & = & \beta_0 & + & X_n\beta_1 & + & \varepsilon_n \end{array} \tag{7.162}$$

And this system of linear equations can be restated in matrix form:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{7.163}$$

where

Vector of responses:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_n \end{pmatrix} \tag{7.164}$$

Design Matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ 1 & X_4 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \tag{7.165}$$

Vector of parameters:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \tag{7.166}$$

and
Vector of error terms:

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{7.167}$$

We could write the whole equation as:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ 1 & X_4 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{7.168}$$

Now, we want to minimize the square of the error, as we did earlier. We can do this compactly using the matrix notation. Note that if we want to sum the squared values of a vector or numbers, we can either do it like this:

```
(epsilon<-rnorm(10))
```

```
##  [1] -1.1096067  0.7127422  1.1692349 -0.7438639
##  [5] -1.0989905 -0.9997620  2.2934026  1.3629571
##  [9]  0.2898685  1.4069600

(epsilon.squared<-epsilon^2)

##  [1] 1.23122701 0.50800145 1.36711033 0.55333347
##  [5] 1.20778012 0.99952400 5.25969534 1.85765193
##  [9] 0.08402375 1.97953636

sum(epsilon.squared)

## [1] 15.04788
```

or we can use matrix multiplication:

```
(t(epsilon)%*%epsilon)

##          [,1]
## [1,] 15.04788
```

We can apply this fact (that we can do all the computations we did above by hand using matrix multiplication) to state our least squares estimation problem by multiplying each side of the equation with its transpose (this is equivalent to taking the square):

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \tag{7.169}$$

Now, as we did earlier, we take the derivative with respect to the *vector* β, and we get:

$$\frac{dS}{d\beta} = -2X'(Y - X\beta) \tag{7.170}$$

Set this to 0 and solve for β:

$$-2X'(Y - X\beta) = 0 \tag{7.171}$$

Rearranging this equation gives us the normal equations we saw earlier, except that it's in matrix form:

$$X'Y = X'X\beta \tag{7.172}$$

Here, we use the fact that multiplying a matrix by its inverse gives the identity matrix. This fact about inverses allows us to solve the equation.We just premultiply by the inverse of $X'X$ (written $(X'X)^{-1}$) on both sides:

$$(X'X)^{-1}X'Y = (X'X)^{-1}X'X\beta \tag{7.173}$$

which gives us:

$$\beta = (X'X)^{-1}X'Y \tag{7.174}$$

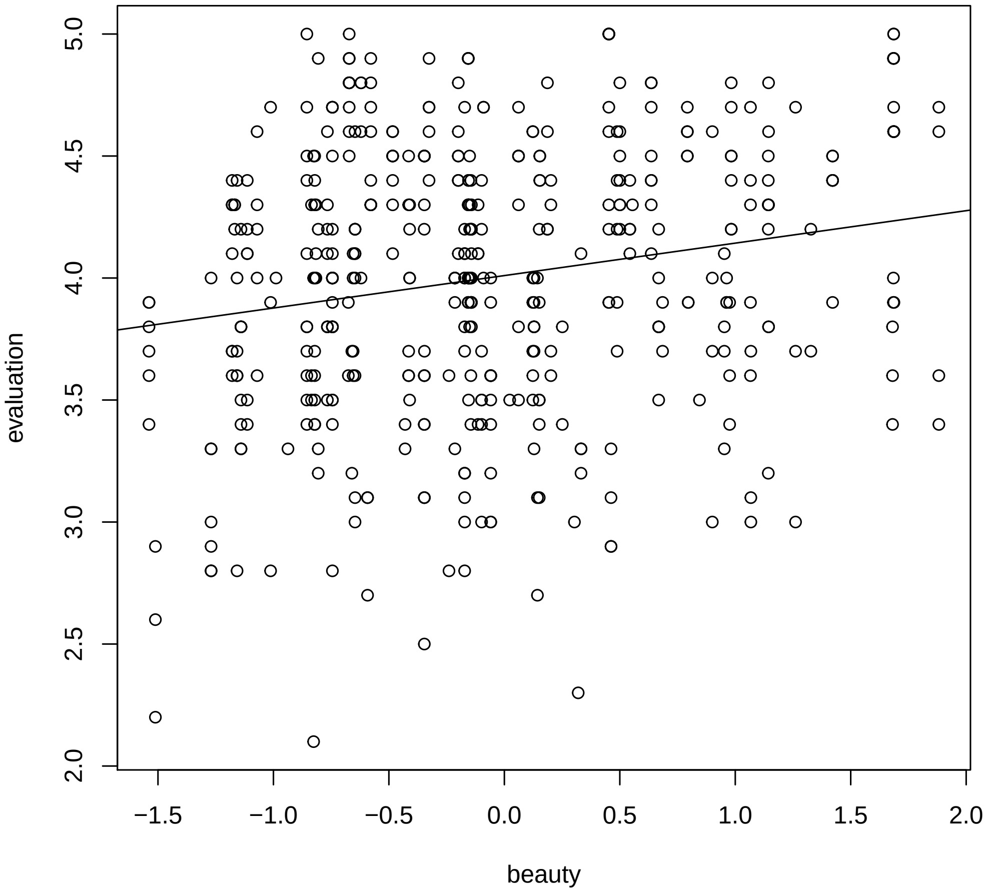
Example:

```r
beauty<-read.table("data/beauty.txt",header=TRUE)

head(beauty)

##       beauty evaluation
## 1  0.2015666        4.3
## 2 -0.8260813        4.5
## 3 -0.6603327        3.7
## 4 -0.7663125        4.3
## 5  1.4214450        4.4
## 6  0.5002196        4.2
```

Fit a linear model:

```r
fm<-lm(evaluation~beauty,beauty)
round(summary(fm)$coefficients,5)

##             Estimate Std. Error   t value
## (Intercept)  4.01002    0.02551 157.20522
## beauty       0.13300    0.03218   4.13337
##             Pr(>|t|)
## (Intercept)    0e+00
## beauty         4e-05
```

```
with(beauty,plot(evaluation~beauty))
abline(fm)
```



```
X<-model.matrix(fm)
head(X)

##   (Intercept)     beauty
## 1           1  0.2015666
## 2           1 -0.8260813
## 3           1 -0.6603327
## 4           1 -0.7663125
## 5           1  1.4214450
## 6           1  0.5002196

Y<-beauty$evaluation
```

Use equation 7.174 to find the intercept and slope:

```
solve(t(X)%*%X)%*%t(X)%*%Y

##                  [,1]
## (Intercept) 4.0100227
## beauty      0.1330014
```

Compare with R output:

```
coef(fm)

## (Intercept)      beauty
##   4.0100227   0.1330014
```

Bibliography

[1] N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, 1998.

[2] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, UK, 2007.

[3] J. Gilbert and C.R. Jordan. *Guide to mathematical methods*. Macmillan, 2002.

[4] C.M. Grinstead and J.L. Snell. *Introduction to probability*. American Mathematical Society, 1997.

[5] D.W. Jordan and P. Smith. *Mathematical Techniques: An Introduction for the Engineering, Physical, and Mathematical Sciences*. Oxford University Press, 4 edition, 2008.

[6] G. Jay Kerns. *Introduction to Probability and Statistics Using R*. 2010.

[7] Sheldon Ross. *A first course in probability*. Pearson Education, 2002.

[8] S.L. Salas, E. Hille, and G.J. Etgen. *Calculus: One and several variables*. Wiley, 2003.

[9] A. Sen and M. Srivastava. *Regression Analysis: Theory, Methods and Applications*. Springer, New York, 1990.

[10] M. Spivak. *Calculus*. Cambridge, 2010.