Notes to self

on mathematics, probability theory, and statistics

Compiled by Shravan Vasishth

version of May 4, 2014

Contents

Preface

It is worth noting that there is **absolutely nothing original** in terms of content in this document. As the title suggests, these are just notes for myself, for reviewing material I learnt in the Graduate Certificate in Statistics at Sheffield University (2011-12), and in the MSc at Sheffield (2012-2015). I try to remember to cite sources at the beginning of a section, but I might not do that consistently (sometimes I'm short on time). The references at the end of the document are an incomplete listing of the sources I consulted.

Chapter 1

Helpful advice on studying math etc.

Chapter 2

Mathematics

Sources: heavily depended on [8], [9], [3] (the official textbook in the course), [2] and various summaries on the internet on trigonometric functions.

2.1 Homework assignments

These assignments cover pretty much all the course material, so they are worth reviewing.

- Ass 0:
  - Sets
  - Functions
  - Inequalities
- Ass 1:
  - Series
  - Partial Fractions and telescoping series
  - **Taylor, Maclaurin series**
- Ass 2:
  - **Integration methods**
  - Riemann integrals

- Ass 3:
  - Change of variables ($\int_0^\infty x^2 e^{-x^5}\,dx$)
  - Gamma function
  - Partial differentiation (level curves, lagrangian multipliers)
  - Vectors
- Ass 4:
  - Double integrals
  - Gaussian elimination
  - Inverse
  - Determinants
- Ass 5:
  - Polar coordinates
  - Linear independence
  - Diagonalization
  - Quadratic forms

2.2 Trigonometry

2.2.1 Basic definitions

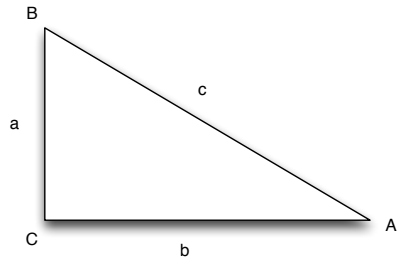

Figure 2.1: Right triangle.

$$\sin A = \frac{opp}{hyp} = \frac{a}{c} \tag{2.1}$$

Cosine is the complement of the sine:

$$\cos A = \sin(90 - A) = \sin B \tag{2.2}$$

$$\cos A = \frac{b}{c} \tag{2.3}$$

2.2.2 Pythagorean identity

$$a^2 + b^2 = c^2 \tag{2.4}$$

$$\begin{aligned} \frac{a^2}{c^2} + \frac{b^2}{c^2} &= 1 \\ \sin^2 A + \cos^2 A &= 1 \end{aligned} \tag{2.5}$$

2.2.3 Relations between trig functions

$$\tan A = \frac{\sin A}{\cos A} = \frac{a}{c} / \frac{b}{c} = \frac{a}{b} = \frac{opp}{adj} \tag{2.6}$$

$$\cot A = \frac{1}{\tan A} = \frac{\cos A}{\sin A} \tag{2.7}$$

$$\sec A = \frac{1}{\cos A} \tag{2.8}$$

$$\csc A = \frac{1}{\sin A} \tag{2.9}$$

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.

2.2.4 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)

2.2.5 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
$\sin \pi = 0$	$\cos \pi = -1$	$\tan \pi = -1$

2.2.6 Law of cosines

Three ways of writing it:

$$c^2 = a^2 + b^2 - 2ab\cos C \tag{2.10}$$

$$a^2 = b^2 + c^2 - 2bc\cos C \tag{2.11}$$

$$b^2 = c^2 + a^2 - 2ca\cos C \tag{2.12}$$

2.2.7 Law of sines

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c} \tag{2.13}$$

2.2.8 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.

2.2.9 Sum formulas for sine and cosine

$$\sin(s+t) = \sin s \cos t + \cos s \sin t \tag{2.14}$$

$$\cos(s+t) = \cos s \cos t - \sin s \sin t \tag{2.15}$$

2.2.10 Double angle formulas for sine and cosine

$$\sin 2t = 2 \sin t \cos t \tag{2.16}$$

$$\cos 2t = \cos^2 t - \sin^2 t = 2\cos^2 t - 1 = 1 - 2\sin^2 t \tag{2.17}$$

2.2.11 Less important identities

Pythagorean formula for tan and sec:

$$\sec^2 t = 1 + \tan^2 t \tag{2.18}$$

Identities expressing trig functions in terms of their supplements

$$\sin(\pi - t) = \sin t \tag{2.19}$$

$$\cos(\pi - t) = -\cos t \tag{2.20}$$

$$\tan(\pi - t) = -\tan t \tag{2.21}$$

Difference formulas for sine and cosine

$$\sin(s-t) = \sin s \cos t - \cos s \sin t \tag{2.22}$$

$$\cos(s-t) = \cos s \cos t + \sin s \sin t \tag{2.23}$$

2.3 Basic differentiation

2.3.1 Diff. from first principles

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

$$f^{(1)}(x) = \frac{f(x+\delta x) - f(x)}{\delta x} \tag{2.24}$$

2.3.2 Derivations of combinations of functions

$$(uv)' = uv' + vu' \tag{2.25}$$

$$(u/v)' = \frac{vu' - uv'}{v^2} \tag{2.26}$$

derivations of trig functions

2.3.3 Leibniz' rule

This is about successive differentiation of products of functions uv:

$$uv^{(n)} = u^{(n)}v + \binom{n}{1} u^{(n-1)}v^{(1)} + \cdots + \binom{n}{r} u^{(n-r)}v^{(r)} + uv^{(n)} \tag{2.27}$$

2.4 Series

2.4.1 Arithmetic series

General form:

$$a + (a+d) + (a+2d) + \ldots \tag{2.28}$$

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

$$S_k = \sum_{n=1}^{k} (a + (n-1)d) \tag{2.29}$$

The sum can be found by:

$$S_k = \frac{k}{2}(2a + (k-1)d) \tag{2.30}$$

2.4.2 Geometric series

General form:

$$a + ar + ar^2 \ldots \tag{2.31}$$

In summation notation:

$$\sum_{n=1}^{\infty} ar^{n-1} \tag{2.32}$$

k-th partial sum:

$$S_k = \frac{a - (1 - r^k)}{1 - r} \tag{2.33}$$

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

$$S_\infty = \frac{a}{1 - r} \tag{2.34}$$

2.4.3 Clever trick for computing partial sums of geometric series

to-do (see my P-Ass1 solution)

2.4.4 Power series

$$\sum_{n=0}^{\infty} a_n (x - a)^n \tag{2.35}$$

radius of convergence: to-do

2.4.5 Taylor's theorem (Taylor series)

We can represent a function as a power series (SV: not sure if we can do this for any function):

$$f(x) = a_0 + a_1(x-a) + a_2(x-a)^2 + \cdots + a_n(x-a)^n + R_n(x) \tag{2.36}$$

where $R_n(x)$ is the remainder term (a power series beginning with $a_{n+1}(x-a)^{n+1}$).

Taylor's theorem: Let f be a function that is $n+1$ times differentiable on an open interval containing points a and x. Then

$$f(x) = f(a) + f^{(1)}\frac{(x-a)^1}{1!} + f^{(2)}\frac{(x-a)^2}{2!} + \cdots + f^{(n)}\frac{(x-a)^n}{n!} + R_n(x) \tag{2.37}$$

where $R_n(x) = f^{(n+1)}(c)\frac{(x-a)^n+1}{n+1!}$, c is some point between a and x.

Taylor series: If $R_n(x)$ tends to zero as $n \to \infty$, then the series in the above theorem converges and is called the Taylor series:

$$\sum_{n=0}^{\infty} f^{(n)}\frac{(x-a)^n}{n!} \tag{2.38}$$

basic taylor series: p. 91
combination rules for power series: p. 75

2.4.6 Maximizing and minimizing functions

2.4.7 Partial derivatives

2.4.7.1 Chain rule for partial derivatives

2.4.8 Maxima and minima in higher dimensions

2.4.9 Lagrangian multipliers

2.5 Integration

2.5.1 Riemann sums

A simple example comes from M-Ass-2-problem-2:

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 $\delta x = 1/10$, and each of the ten x_i are $1/10, 2/10, \ldots, 10/10$, i.e., $i/10$, where $i = 1, \ldots, 10$.

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

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

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:

```
> n<-10
> (x.i<-(1:n)/n)
 [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> A<- ((1/n) * (1/sqrt(2 * pi)) * sum(my.fn(x.i)))
> (A<-round(A,digits=5))
[1] 0.33329
```

Compare this to the exact value, computed using R:

```
> fprob2<-function(x){
          (1/sqrt(2 * pi))*exp(1)^(-(x^2/2))
   }
> integrate(fprob2,lower=0,upper=1)
0.34134 with absolute error < 3.8e-15
```

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

```
> ## As an aside, note that one can get really close
> ## to Neave's 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] 0.34131
```

With 2000 rectangles, we can get a better estimate of the area than with 10 rectangles: 0.341305498138516.

2.5.2 Some common integrals

$$\int \frac{1}{x}\,dx = \log|x| + c \tag{2.39}$$

$$\int \log x\,dx = \frac{1}{x} + c \tag{2.40}$$

2.5.3 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).

2.5.4 Rules of integration

2.5.5 Standard integrals

2.5.6 The u-substitution

From [8, 306]:

An integral of the form

$$\int f(g(x))g'(x)\,dx \tag{2.41}$$

can be written as

$$\int f(u)\,du \tag{2.42}$$

by setting

$$u = g(x) \tag{2.43}$$

and

$$du = g'(x)\,dx \tag{2.44}$$

If F is an antiderivative for f, then

$$\frac{d}{dx}[F(g(x))] \underset{\text{by the chain rule}}{\overset{\uparrow}{=}} F'(g(x))g'(x) \underset{F'=f}{\overset{\uparrow}{=}} f(g(x))g'(x) \tag{2.45}$$

We can obtain the same result by calculating:

$$\int f(u)\,du \tag{2.46}$$

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

$$\int f(u)\,du = F(u) + C = F(g(x)) + C \tag{2.47}$$

A frequently occurring type of integral is

$$\int \frac{g'(x)}{g(x)}\,dx \tag{2.48}$$

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

$$\int \frac{g'(x)}{g(x)}\,dx = \int \frac{1}{u}du = ln \mid u \mid + C \tag{2.49}$$

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:

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

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:

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

Examples:

- From M-Ass2-problem-4a: Use a substitution to find:

$$\int_1^4 x^2 e^{-x^3}\,dx$$

- From M-Ass2-problem-4b: Use a substitution to find

$$\int_1^4 \frac{1+2\log_e x}{x}\,dx$$

- From M-Ass2-problem-5: We have to use an appropriate substitution to find the indefinite integral:

$$\int xe^{-x^2}\,dx$$

Then, use the answer to the above to calculate:

$$\int_0^\infty xe^{-x^2}\,dx$$

Then, we have to use integration by parts (using the first answer above) to find a reduction formula for:

$$I_n = \int_0^\infty x^n e^{-x^2}\,dx \tag{2.50}$$

[Note: this last problem needs a certain amount of concentration, so make yourself comfortable before you start.]

2.5.7 Integration by parts

$$\frac{d(uv)}{dx} = u\frac{dv}{dx} + \int v\frac{du}{dx} \tag{2.51}$$

$$uv = \int u\frac{dv}{dx}dx + \int v\frac{du}{dx}dx \tag{2.52}$$

$$\int u\frac{dv}{dx}dx = uv - \int v\frac{du}{dx}dx \tag{2.53}$$

Examples:
From M-Ass2-problem-4c: Use integration by parts to find

$$\int_0^{\pi/2} \sin(2x)e^{-x}\,dx$$

The trick here: get the same expression on the right-hand side (RHS) as you have on the LHS, and then solve for the LHS.

2.5.8 Change of variables (gamma functions)

We can solve integrals like

$$\int_0^{\infty} x^2 e^{-x^5}\,dx \tag{2.54}$$

by restating it as the gamma function:

$$\Gamma(z) = \int_0^{\infty} x^{z-1}e^{-x}\,dx \tag{2.55}$$

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

$$\frac{y^{2/5}}{5y^{4/5}}e^y\,dy \tag{2.56}$$

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

Note that

$$\Gamma(z) = (z-1)\Gamma(z-1) \tag{2.57}$$

2.5.9 Double integrals

2.5.10 Change of variables in multiple integration

2.5.11 Vectors

A vector is an ordered triple. It has a geometric interpretation only when the coordinates are established.

$$\vec{a} = (x_1, y_1, z_1) \tag{2.58}$$

Norm

$$\|\vec{a}\| = \sqrt{x_1^2 + y_1^2 + z_1^2} \tag{2.59}$$

Vectors of norm 1 are called unit vectors. For each non-zero vector there is a unit vector going in the same direction.

$$u_{\vec{a}} = \frac{1}{\|\vec{a}\|} \tag{2.60}$$

Also, every vector can be expressed as a linear combination of these three unit vectors:

$$\vec{i} = (1,0,0) \quad \vec{j} = (0,1,0) \quad \vec{k} = (0,0,1) \tag{2.61}$$

I.e., if

$$\vec{a} = (x_1, y_1, z_1) \tag{2.62}$$

then

$$\vec{a} = x_1\vec{i} + y_1\vec{j} + z_1\vec{k} \tag{2.63}$$

Dot product:

$$\vec{a} \cdot \vec{b} = a_1b_1 + a_2b_2 + a_3b_3 \tag{2.64}$$

Note that

$$\vec{a}\vec{a} = || \vec{a} ||^2 \tag{2.65}$$

and

$$\vec{a} \cdot \vec{b} = || \vec{ab} || \cos\theta \tag{2.66}$$

Also,

$$\cos\theta = u_{\vec{a}} \cdot u_{\vec{b}} \tag{2.67}$$

2.6 Linear Algebra

This section depends heavily on [1], and sometimes I quote exactly from this book.

2.6.1 Gaussian elimination for solving systems of linear equations

Elementary row operations (note that these are reversible, and the effect of any sequence of row operations on a system of equations is to produce an equivalent system of equations):

1. **Replacement**: Replace one equation by the sum of itself and a multiple of another equation
2. **Interchange**: Interchange two equations
3. **Scaling**: Multiply all the terms of an equation by a nonzero constant

Definition 1. ***Row equivalent****: Two matrices are row equivalent if a sequence of elementary row operations can transform one matrix to the other.*

Definition 2. ***Echelon form****: example:*

$$\begin{array}{c} \\ \textit{pivot position} \\ \\ \textit{pivot column} \end{array} \begin{pmatrix} 3 & 1 & 1 \\ \uparrow & & \\ 0 & 4 & -1 \\ 0 & 0 & 7 \\ \uparrow & & \end{pmatrix} \tag{2.68}$$

Definition 3. ***Reduced echelon form****: example:*

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.69}$$

Theorem 1. ***Consistency****: A linear system is consistent iff the rightmost column of an augmented matrix is not a pivot column, i.e.,*

$[0 \cdots 0b]$ *with b nonzero.*

[A leading non-zero entry of a row, when used in this way, is called a *pivot*.]

Identity matrix

Multiplying a matrix by its inverse (see below) 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
>
```

And multiplying an identity matrix with any (conformable) matrix gives that matrix:

```
> (I<-matrix(c(1,0,0,1),2,2))
     [,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 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.

Determinant

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
>
```

Given a 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, det = $ad - bc$.

For a 3×3 matrix M_1, the determinant is:

$$\det M_1 = a_{11} \det M_{1_{11}} - a_{12} \det M_{1_{12}} + a_{13} \det M_{1_{13}}$$

Inverse of a matrix

For a matrix A, A^{-1} is its inverse. Think of it as the reciprocal of a number: the reciprocal of 10 is 1/10, and if we multiple 10 and 1/10, we get 1. Similarly, if we multiply the matrix by its inverse we get the identity matrix.

A matrix has an inverse iff its determinant is not equal to zero.

Note that $AA^1 = I$ and $A^1A = I$.

Singular matrix=non-invertible matrix

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

2.6.2 Gauss-Jordan reduction

To solve an equation of the form

$$AX = B \tag{2.70}$$

Take $[A \mid B]$ and reduce A in this augmented matrix to the identity matrix:

$$\begin{pmatrix} 1 & 0 & 0 & b_1 \\ 0 & 1 & 0 & b_2 \\ 0 & 0 & 1 & b_3 \end{pmatrix} \tag{2.71}$$

$B' = [b_1 b_2 b_3]$ is the solution. Basically, we have

$$IX = B' \tag{2.72}$$

or

$$X = B' \tag{2.73}$$

This works because elementary operations are reversible and the effect of any sequence of row operations on a system of equations is to produce an equivalent system of equations: the system $AX = B$ is equivalent to the system $X = IX = B'$, which is to say $X = B'$ is a solution of $AX = B$.

Note that if a solution exists, it is unique, i.e., there is only one solution.

If the $n \times n$ matrix A is non-singular, then every equation of the form $AX = B$ (where both X and B are $n \times p$ matrices) does have a solution and also that the solution $X = B'$ is unique. On the other hand, if A is singular, an equation of the form $AX = B$ may have a solution, but there will certainly be matrices B for which $AX = B$ has no solutions.

When A is a singular $n \times n$ matrix, if $AX = B$ has a solution X for a particular B, then it has infinitely many solutions. **See general case below of Gauss-Jordan reduction**.

If A is non-singular then $AX = B$ has a solution for every B, while if A is singular, there are many B for which $AX = B$ has no solution.

2.6.2.1 The general case of Gauss-Jordan reduction

Taken verbatim from [1].

Gauss-Jordan reduction works just as well if the coefficient matrix A is singular or even if it is not a square matrix. Consider the system

$$Ax = b$$

where the coefficient matrix A is an $m \times n$ matrix. The method is to apply elementary row operations to the augmented matrix

$$[A \,|\, b] \to \cdots \to [A' \,|\, b']$$

making the best of it with the coefficient matrix A.

Example:

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 0 & 0 & 3 & 6 \\ 0 & 0 & 0 & 0 \end{array}\right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \\ 3 \end{pmatrix} \tag{2.74}$$

Using Gauss-Jordan reduction we get:

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & -3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right) \tag{2.75}$$

This resulting augmented matrix corresponds to the system

$$\begin{aligned} x_1 + x_2 &= -3 \\ x_3 &= 2 \\ 0 &= 0 \end{aligned}$$

If the last equation were $0 = 6$ or some such, it would be inconsistent—it would have no solution.

We can rewrite the above system as:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -3 - x_2 \\ x_2 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 2 \end{pmatrix} + x_2 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

x_2 can have any value—it is a free variable. The fact that it is arbitrary means that there are infinitely many solutions. x_1 and x_2 are bound variables (and they are in pivot positions).

Geometrically this means that the solution of the above system is a vector equation: A line passing through the end-point of the vector $(-3,0,2)$ and parallel to the vector $(-1,1,0)$.

Summary of the procedure for Gauss-Jordan reduction in the general case

Gauss–Jordan reduction of the coefficient matrix is always possible, but *the pivots don't always end up on the diagonal*. In any case, the Jordan part of the reduction will yield a 1 in each pivot position with zeroes elsewhere in the column containing the pivot. The position of a pivot in a row will be on the diagonal or to its right, and all entries in that row *to the left of the pivot* will be zero. Some of the entries to the right of the pivot may be non-zero.

If the number of pivots is smaller than the number of rows (which will always be the case for a singular square matrix), then some rows of the reduced coefficient matrix will consist entirely of zeroes. If there are non-zero entries in those rows to the right of the divider *in the augmented matrix*, the system is inconsistent and has no solutions.

Otherwise, the system does have solutions. Such solutions are obtained by writing out the corresponding system, and transposing all terms *not associated with the pivot position* to the right side of the equation. Each unknown in a pivot position is then expressed in terms of the non-pivot unknowns (if any). The pivot unknowns are said to be *bound*. The non-pivot unknowns may be assigned any value and are said to be *free*.

2.6.2.2 Geometric visualization (vector space)

In R^3, the graph of a single linear equation

$$a_1x_1 + a_2x_2 + a_3x_3 = b$$

is a plane. Hence, by analogy, we call the 'graph' in R^4 of

$$a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 = b$$

a *hyperplane*.

Example:

Consider the system

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 0 \\ x_1 + 2x_2 + x_3 + 3x_4 &= 0 \\ 2x_1 + 4x_2 + 3x_4 &= 0 \end{aligned}$$

If we solve this (exercise), we get the system corresponding to the reduced augmented matrix:

$$\begin{aligned} x_1 + 2x_2 \qquad + (3/2)x_4 &= 0 \\ x_3 + (3/2)x_4 &= 0 \\ 0 &= 0 \end{aligned}$$

Thus,

$$\begin{aligned} x_1 &= -2x_2 - (3/2)x_4 \\ x_3 &= \qquad -3(/2)x_4 \end{aligned}$$

with x_1 and x_3 *bound* and x_2 and x_4 *free*.

[skipping a few steps] A general solution is:

$$x = x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -3/2 \\ 0 \\ -3/2 \\ 1 \end{pmatrix}$$

where the free variables x_2 and x_4 can assume any value. The bound variables x_1 and x_3 are then determined.

As [1, 40] puts it "This solution may also be interpreted geometrically in R^4. The original set of equations may be thought of as determining a 'graph' which is the intersection of three hyperplanes (each defined by one of the equations.) Note also that each of these hyperplanes passes through the origin since the zero vector is certainly a solution."

2.6.2.3 The LU decomposition

not needed for course, to-do. Notes from Leonard Evens (Northwestern math).

2.6.3 Vector spaces, subspaces, and linear combinations

A homogeneous system is

$$Ax = 0 \tag{2.76}$$

and if the RHS vector is non-zero, inhomogeneous:

$$Ax = b \tag{2.77}$$

Every inhomogeneous system has an associated homogeneous system, and the solutions of the two systems are closely related.

Consider the examples above:

Example 1:

$$\begin{pmatrix} 1 & 1 & 2 \\ -1 & -1 & 1 \\ 1 & 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \\ 3 \end{pmatrix}$$

This was shown above to have the solution

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -3 - x_2 \\ x_2 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 2 \end{pmatrix} + x_2 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

Example 2:

Consider

$$\begin{pmatrix} 1 & 1 & 2 \\ -1 & -1 & 1 \\ 1 & 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This has the general solution

$$x = x_2 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \tag{2.78}$$

where x_2 is also free.

Note that if we set $x_2 = 0$ in (**??**), we obtain the specific solution

$$\begin{pmatrix} -3 \\ 0 \\ 2 \end{pmatrix}$$

and then the remaining part of the solution is a general solution of the homogeneous equation.

The general principle is: *You can always find a general solution of an inhomogeneous linear system by adding one particular solution to a general solution of the corresponding homogeneous system.*

to-do algebraic explanation.

2.6.3.1 Null spaces

The null space is the set of all solutions x to

$$Ax = 0 \tag{2.79}$$

These are interesting because of the relationship described above between homogeneous and non-homogeneous systems.
Notice that the null space of an $m \times n$ matrix is a subset of R^n.

Definition 4. ***Vector subspace****: A non-empty subset V of R^n is called a vector subspace if it has the property that any linear combination of vectors in V is also in V . In symbols, if u and v are vectors in V , and a and b are scalars, then au + bv is also a vector in V .*

Aside: The entire set R^n is considered a subset of itself; i.e., a vector subspace of itself.

Why is the zero vector always in the subspace?

"The zero vector must be in every vector subspace W. Indeed, just pick any two vectors u and v in W—v could even be a multiple of u. Then 0 = (0)u + (0)v, the linear combination with both scalars a = b = 0, must also be in W. The upshot is that any set which does not contain the zero vector cannot be a vector subspace.

The set consisting only of the zero vector 0 has the desired property—any linear combination of zero with itself is also zero. Hence, that set is also a vector subspace, called the **zero subspace**."

Why are we interested in vector subspaces?

At least two reasons:

First, they arise in null spaces, and these are interesting because they constitute solutions to systems of homogeneous linear equations.

To see why a null space satisfies the definition, suppose u and v are both solutions of Ax = 0. That is, Au = 0 and Av = 0. Then

$$A(au+bv) = A(au)+A(bv) = aAu+bAv = a0+b0 = 0 \tag{2.80}$$

Second (related to the above point), talking about vector subspaces as the solution set of a homogeneous system of linear equations gives us a compact way of talking about the solution set (or null space). For example, in

$$\begin{pmatrix} 1 & 2 & -1 & 0 \\ 1 & 2 & 1 & 3 \\ 2 & 4 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

As shown earlier, its null space consists of all vectors of the form

$$x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -3/2 \\ 0 \\ -3/2 \\ 0 \end{pmatrix}$$

as the free scalars x_2 and x_4 range over all possible values. Let

$$v_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} \qquad v_2 = \begin{pmatrix} -3/2 \\ 0 \\ -3/2 \\ 0 \end{pmatrix}$$

Then, what we have discovered is that the solution set or null space consists of all linear combinations of the set (v_1, v_2) of vectors. **This is a much more useful way of presenting the answer, since we specify it in terms of a small number of objects—in this case just two. Since the null space itself is infinite, this simplifies things considerably.**

2.6.3.2 Spanning set

In general, suppose W is a vector subspace of R^n and $\{v_1, v_2, \ldots, v_k\}$ is a finite subset of W. We say that $\{v_1, v_2, \ldots, v_n\}$ is a *spanning set* for W (or more simply that it *spans* W) if each vector v in W can be expressed as a linear combination

$$v = s_1 v_1 + s_2 v_2 + \cdots + s_k v_k,$$

for appropriate scalars $s_1, s_2, \ldots, s_k$. The simplest case of this is when $k = 1$, i.e., the spanning set consists of a single vector v. Then the subspace spanned by this vector is just the set of all sv with s an arbitrary scalar. If $v \neq \mathbf{0}$, this set is just the line through the origin containing v.

Example:

Consider the set of solutions x in R^4 of the single homogeneous equation

$$x_1 - x_2 + x_3 - 2x_4 = 0.$$

This is the null space of the 1×4 matrix

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

The matrix is already reduced with pivot 1 in the $1,1$-position. The general solution is

$$x_1 = x_2 - x_3 + 2x_4 \qquad x_2, x_3, x_4 \quad \text{free},$$

and the general solution vector is

$$x = \begin{pmatrix} x_2 - x_3 + 2x_4 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = x_2 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 2 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

It follows that the null space is spanned by

$$\left\{ v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad v_3 = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

This is a special case of a more general principle: **Gauss-Jordan reduction for a homogeneous system always results in a description of the null space as the vector subspace spanned by a finite set of basic solution vectors**.

Chapter 3

Probability

Sources: Kerns [5].

3.1 Kolmogorov Axioms of Probability

3.1.1 Axiom 1

$(\mathbb{P}(A) \geq 0)$ for any event $(A \subset S)$.

3.1.2 Axiom 2

$(\mathbb{P}(S) = 1)$.

3.1.3 Axiom 3

If the events $(A_1), (A_2), (A_3) \ldots$ are disjoint then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) \text{ for every } n, \tag{3.1}$$

and furthermore,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{3.2}$$

3.1.4 Four important propositions

We'll be using these later on a lot.

Proposition 1. *Let $E \cup E^c = S$. Then,*

$$1 = P(S) = P(E \cup E^c) = P(E) + P(E^c) \tag{3.3}$$

or:

$$P(E^c) = 1 - P(E) \tag{3.4}$$

Proposition 2. *If $E \subset F$ then $P(E) \leq P(F)$.*

Proposition 3.

$$P(E \cup F) = P(E) + P(F) - P(EF) \tag{3.5}$$

This result will be needed for a (to me) totally non-obvious outcome in multivariate distributions.

Proposition 4. *This is ugly if written in a formula, but easy to follow intuitively: the inclusion-exclusion identity.*

3.2 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	$n!/(n-k)!$	$\binom{n}{k}$

3.3 Permutations

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

$$\frac{n!}{n_1!n_2!\ldots n_r!} \tag{3.6}$$

3.4 Combinations

Choosing k distinct objects from n, when order irrelevant:

$$\binom{n}{k} = \frac{n!}{(n-r)!r!} \tag{3.7}$$

3.5 Binomial theorem

$$(x+y)^n = \sum_{n=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{3.8}$$

3.6 Conditional probability

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

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad \text{if } \mathbb{P}(A) > 0. \tag{3.9}$$

Theorem: For any fixed event A with $\mathbb{P}(A) > 0$,

1. $\mathbb{P}(B|A) \geq 0$, for all events $B \subset S$,
2. $\mathbb{P}(S|A) = 1$, and
3. If B_1, B_2, B_3,... are disjoint events,

then:

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} B_k \,\middle|\, A\right) = \sum_{k=1}^{\infty} \mathbb{P}(B_k|A). \tag{3.10}$$

In other words, $\mathbb{P}(\cdot|A)$ is a legitimate probability function. With this fact in mind, the following properties are immediate:

For any events A, B, and C with $\mathbb{P}(A) > 0$,

1. $\mathbb{P}(B^c|A) = 1 - \mathbb{P}(B|A)$.
2. If $B \subset C$ then $\mathbb{P}(B|A) \leq \mathbb{P}(C|A)$.
3. $\mathbb{P}[(B \cup C)|A] = \mathbb{P}(B|A) + \mathbb{P}(C|A) - \mathbb{P}[(B \cap C|A)]$.
4. The Multiplication Rule. For any two events A and B,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A). \tag{3.11}$$

And more generally, for events A_1, A_2, A_3,..., A_n,

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\cdots\mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}). \tag{3.12}$$

3.6.1 Some important results that keep turning up

1. Conditional probability $A \mid B$ stated in terms of the complement $A' \mid B$:

$$\begin{aligned} P(B) =& P(AB) + P(A'B) \\ =& P(A \mid B)P(B) + P(A' \mid B)P(B) \\ \Leftrightarrow& P(B) = P(B)[P(A \mid B) + P(A' \mid B)] \\ \Leftrightarrow& 1 = P(A \mid B) + P(A' \mid B) \\ \Leftrightarrow& P(A \mid B) = 1 - P(A' \mid B) \end{aligned} \tag{3.13}$$

2. Eq 3.1 in Ross:

$$P(A) = P(A \mid C)P(C) + P(A \mid C')P(C') \tag{3.14}$$

Note also (the equation below is the same as the second line in the first item above):

$$P(C) = P(A \mid C)P(C) + P(A' \mid C)P(C) \tag{3.15}$$

3. A product P(AB) can be written:

$$P(AB) = P(A \mid B)P(B) = P(B \mid A)P(A) \tag{3.16}$$

3.6.2 Iterated conditional probability

to-do: see Cameron book (prob-1.pdf in ExtraReading)

3.6.3 Independence of events

[Taken nearly verbatim from [5].]

Definition 5. *Events A and B are said to be independent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{3.17}$$

Otherwise, the events are said to be dependent.

We know that when $\mathbb{P}(B) > 0$ we may write

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{3.18}$$

In the case that A and B are independent, the numerator of the fraction factors so that $\mathbb{P}(B)$ cancels, with the result:

$$\mathbb{P}(A|B) = \mathbb{P}(A) \text{ when } A, B \text{ are independent.} \tag{3.19}$$

Proposition:

If E and F are independent events, then so are E and F^c, E^c and F, and E^c and F^c.

Proof:

Assume E and F are independent. Since $E = EF \cup EF^c$ and EF and EF^c are mutually exclusive,

$$\begin{aligned} P(E) =& P(EF) + P(EF^c) \\ =& P(E)P(F) + P(EF^c) \end{aligned} \tag{3.20}$$

Equivalently:

$$\begin{aligned} P(EF^c) =& P(E)[1-P(F)] \\ =& P(E)P(F^c) \end{aligned} \tag{3.21}$$

3.6.4 Bayes' rule

[Cited verbatim from [5].]

Theorem 2. ***Bayes' Rule***. *Let B_1, B_2, ..., B_n be mutually exclusive and exhaustive and let A be an event with $\mathbb{P}(A) > 0$. Then*

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(B_k)\mathbb{P}(A|B_k)}{\sum_{i=1}^{n}\mathbb{P}(B_i)\mathbb{P}(A|B_i)}, \quad k = 1,2,\ldots,n. \tag{3.22}$$

The proof follows from looking at $\mathbb{P}(B_k \cap A)$ in two different ways. For simplicity, suppose that $P(B_k) > 0$ for all k. Then

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

Since $\mathbb{P}(A) > 0$ we may divide through to obtain

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(B_k)\mathbb{P}(A|B_k)}{\mathbb{P}(A)}.$$

Now remembering that $\{B_k\}$ is a partition, the Theorem of Total Probability gives the denominator of the last expression to be

$$\mathbb{P}(A) = \sum_{k=1}^{n}\mathbb{P}(B_k \cap A) = \sum_{k=1}^{n}\mathbb{P}(B_k)\mathbb{P}(A|B_k).$$

■

See great example in [5] on misfiling assistants.

3.7 Discrete random variables; Expectation

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., $x \in S_X$. It seems we can also sloppily write $X \in S_X$ (not sure about this).

Good example: number of coin tosses till H

- $X : \omega \to x$
- ω: H, TH, TTH,... (infinite)
- $x = 0, 1, 2, \ldots; x \in S_X$

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

$$p_X : S_X \to [0, 1] \tag{3.23}$$

defined by

$$p_X(x) = P(X(\omega) = x), x \in S_X \tag{3.24}$$

[**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

$$F(a) = \sum_{\text{all } x \leq a} p(x) \tag{3.25}$$

Basic results:

$$E[X] = \sum_{i=1}^{n} x_i p(x_i) \tag{3.26}$$

$$E[g(X)] = \sum_{i=1}^{n} g(x_i) p(x_i) \tag{3.27}$$

$$Var(X) = E[(X - \mu)^2] \tag{3.28}$$

$$Var(X) = E[X^2] - (E[X])^2 \tag{3.29}$$

$$Var(aX + b) = a^2 Var(X) \tag{3.30}$$

$$SD(X) = \sqrt{Var(X)} \tag{3.31}$$

For two independent random variables X and Y,

$$E[XY] = E[X]E[Y] \tag{3.32}$$

Covariance of two random variables:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] \tag{3.33}$$

Note that Cov(X,Y)=0 if X and Y are independent.
Corollary in 4.1 of Ross:

$$E[aX + b] = aE[X] + b \tag{3.34}$$

A related result is about **linear combinations of RVs**:
Theorem. Given two **not necessarily independent** random variables X and Y:

$$E[aX + bY] = aE[X] + bE[Y] \tag{3.35}$$

If X and Y are independent,

$$Var(X + Y) = Var[X] + Var[Y] \tag{3.36}$$

and

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) \tag{3.37}$$

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

$$Var(X - Y) = Var(X) + Var(Y) \tag{3.38}$$

If X and Y are not independent, then

$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X,Y) \tag{3.39}$$

3.7.1 Discrete probability distributions

to-do

3.7.1.1 Binomial

3.7.1.2 Poisson

3.7.1.3 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,

$$P(X = n) = (1-p)^{n-1}p \quad n = 1, 2, \ldots \tag{3.40}$$

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

$$\mu = \frac{1-p}{p} = \frac{q}{p} \text{ and } \sigma^2 = \frac{q}{p^2}. \tag{3.41}$$

to-do: example

3.7.1.4 Negative binomial

[Taken nearly verbatim from [5].]

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

$$f_X(x) = \binom{r+x-1}{r-1} p^r (1-p)^x, \quad x = 0, 1, 2, \ldots \tag{3.42}$$

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

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:

$$\frac{1}{1-t} = \sum_{k=0}^{\infty} t^k, \quad \text{for } -1 < t < 1, \text{ and} \tag{3.43}$$

$$\frac{1}{(1-t)^r} = \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} t^k, \quad \text{for } -1 < t < 1. \tag{3.44}$$

Therefore

$$\sum_{x=0}^{\infty} f_X(x) = p^r \sum_{x=0}^{\infty} \binom{r+x-1}{r-1} q^x = p^r (1-q)^{-r} = 1, \tag{3.45}$$

since $|q| = |1-p| < 1$.
to-do examples

3.7.1.5 Hypergeometric

3.8 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[1] 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,

$$P\{X \in B\} = \int_B f(x)\,dx \tag{3.46}$$

Kerns has the following to add about the above:

> Continuous random variables have supports that look like
>
> $$S_X = [a,b] \text{ or } (a,b), \tag{3.47}$$
>
> or unions of intervals of the above form. Examples of random variables that are often taken to be continuous are:

[1] Is this supposed to be an iff? Note that there are pathological continuous RVs that have no PDF [5, 138]. Kerns says that regardless of this fact, it can be proved that the CDF always exists. Kerns provides no references, so I have to look this up (to-do).

- 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

$$\mathbb{P}(X \in A) = \int_c^d f_X(x)\,\mathrm{d}x. \tag{3.48}$$

- 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:

$$\mathbb{P}(a \le X \le b) = \mathbb{P}(a < X \le b) = \mathbb{P}(a \le X < b) = \mathbb{P}(a < X < b). \tag{3.49}$$

- 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

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)\,dx \tag{3.50}$$

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

$$P\{a \leq X \leq b\} = \int_{a}^{b} f(x)\,dx \tag{3.51}$$

If $a = b$, we get

$$P\{X = a\} = \int_{a}^{a} f(x)\,dx = 0 \tag{3.52}$$

Hence, for any continuous random variable,

$$P\{X < a\} = P\{X \leq a\} = F(a) = \int_{-\infty}^{a} f(x)\,dx \tag{3.53}$$

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

$$\frac{dF(a)}{da} = f(a) \tag{3.54}$$

The density (PDF) is the derivative of the CDF. In the discrete case [5, 128]:

$$f_X(x) = F_X(x) - \lim_{t \to x^-} F_X(t) \tag{3.55}$$

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

$$P\{a - \frac{\varepsilon}{2} \leq X \leq a + \frac{\varepsilon}{2}\} = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)\,dx \approx \varepsilon f(a) \tag{3.56}$$

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.
$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx \tag{3.57}$$

2.
$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \tag{3.58}$$

3.
$$E[aX + b] = aE[X] + b \tag{3.59}$$

4.
$$Var[X] = E[(X-\mu)^2] = E[X^2] - (E[X])^2 \tag{3.60}$$

5.
$$Var(aX + b) = a^2 Var(X) \tag{3.61}$$

3.9 Important classes of continuous random variables

3.9.1 Uniform random variable

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

$$f_X(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha < x < \beta, \\ 0, & \text{otherwise} \end{cases} \tag{3.62}$$

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

$$F_X(a) = \begin{cases} 0, & a < 0, \\ \frac{a-\alpha}{\beta-\alpha}, & \alpha \le t < \beta, \\ 1, & a \ge \beta. \end{cases} \tag{3.63}$$

$$E[X] = \frac{\beta + \alpha}{2} \tag{3.64}$$

$$Var(X) = \frac{(\beta - \alpha)^2}{12} \tag{3.65}$$

3.9.2 Normal random variable

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty. \tag{3.66}$$

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

```
> plot(function(x) dnorm(x), -3, 3,
         main = "Normal density",ylim=c(0,.4),
                ylab="density",xlab="X")
```

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)
0.9545 with absolute error < 1.8e-11
> ## alternatively:
> pnorm(2)-pnorm(-2)
[1] 0.9545
> integrate(function(x) dnorm(x, mean = 0, sd = 1),
            lower=-1,upper=1)
0.68269 with absolute error < 7.6e-15
> ## alternatively:
> pnorm(1)-pnorm(-1)
[1] 0.68269
```

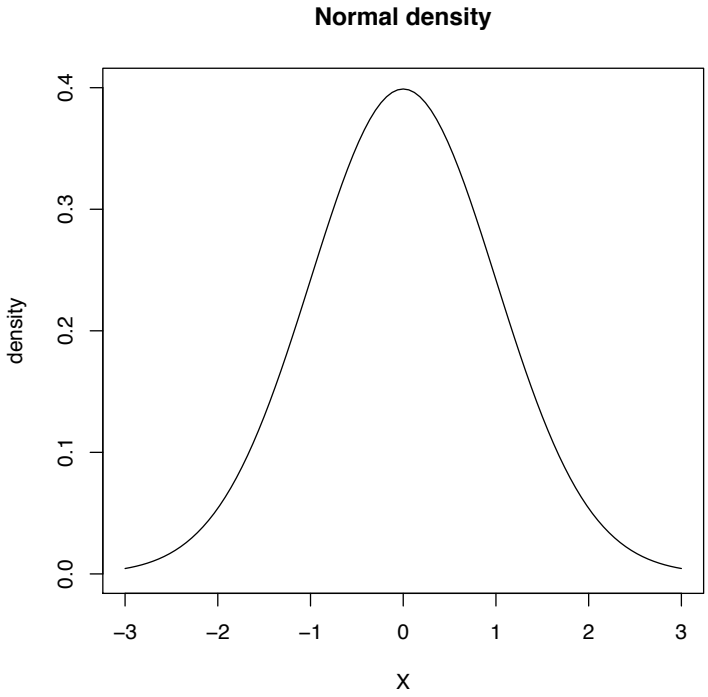

Figure 3.1: Normal distribution.

3.9.2.1 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:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}} dy \quad \text{where} y = (x-\mu)/\sigma \tag{3.67}$$

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

$$\Phi(-x) = 1 - \Phi(x), \quad -\infty < x < \infty \tag{3.68}$$

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

$$p\{Z \leq -x\} = P\{Z > x\}, \quad -\infty < x < \infty \tag{3.69}$$

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:

$$F_X(a) = P\{X \leq a\} = P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{3.70}$$

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:

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

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

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

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

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

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

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

[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 $N(10, 4/\sqrt{4})$.

to-do: print out exercises in chapters 5 and 6 of Grinstead and Snell and work through them.

3.9.3 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.]

3.9.3.1 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:

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

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,

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

3.9.4 Gamma distribution

[The text is an amalgam of [5] 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:

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

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 $X \sim \mathsf{exp}(\texttt{rate} = \lambda)$. 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 $X \sim \mathsf{exp}(\texttt{rate} = \lambda)$. If we let Y measure the length of time until the α^{th} event occurs then $Y \sim \mathsf{gamma}(\texttt{shape} = \alpha, \texttt{rate} = \lambda)$. When α is an integer this distribution is also known as the **Erlang** distribution.

```
> ## fn refers to the fact that it is a function in R, it does n
> gamma.fn<-function(x){
            lambda<-1
            alpha<-1
            (lambda * exp(1)^(-lambda*x) *
            (lambda*x)^(alpha-1))/gamma(alpha)
   }
> x<-seq(0,4,by=.01)
> plot(x,gamma.fn(x),type="l")
```

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

```
> gamma.fn<-function(x){
            lambda<-1/2
            alpha<-8/2 ## n=4
            (lambda * (exp(1)^(-lambda*x)) *
            (lambda*x)^(alpha-1))/gamma(alpha)
   }
> x<-seq(0,100,by=.01)
> plot(x,gamma.fn(x),type="l")
```

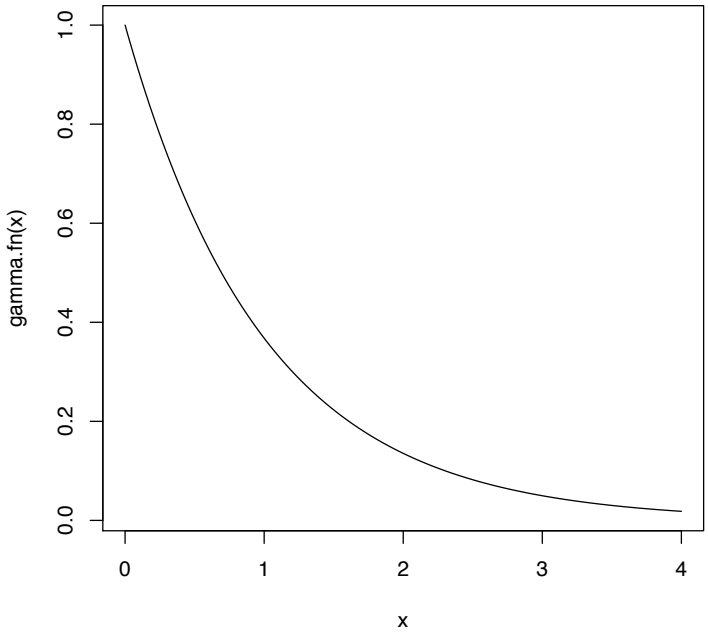

Figure 3.2: The gamma distribution.

.

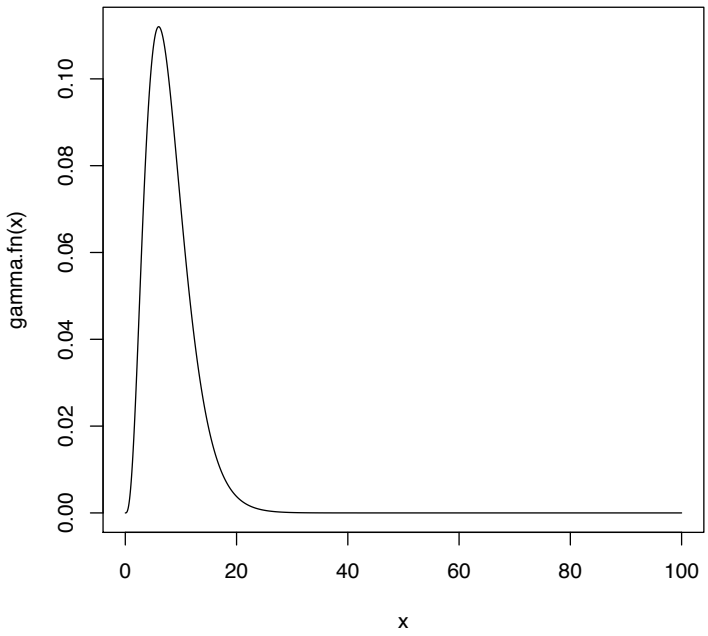

Figure 3.3: The chi-squared distribution.

3.9.4.1 Mean and variance of gamma distribution

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

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

It is easy to show (exercise) that

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

3.9.5 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:

```
> fn<-function(x,lambda){
            lambda*exp(1)^(-lambda*x)
   }
> x<-seq(0,1000,by=1)
> plot(x,fn(x,lambda=1/100),type="l")
> abline(v=200,col=3,lty=3)
> abline(v=300,col=1,lty=3)


> x1<-seq(300,1300,by=1)
> plot(x1,fn(x1,lambda=1/100),type="l")
```

3.9.5.1 Examples of memorylessness

[problem 2 in P-Ass3]

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

$$P(X > t+a \mid X > a) = \frac{P(X > t+a)}{P(X > a)} \tag{3.71}$$

hence establishing the 'absence of memory' property:

$$P(X > t+a \mid X > a) = P(X > t) \tag{3.72}$$

Proof:

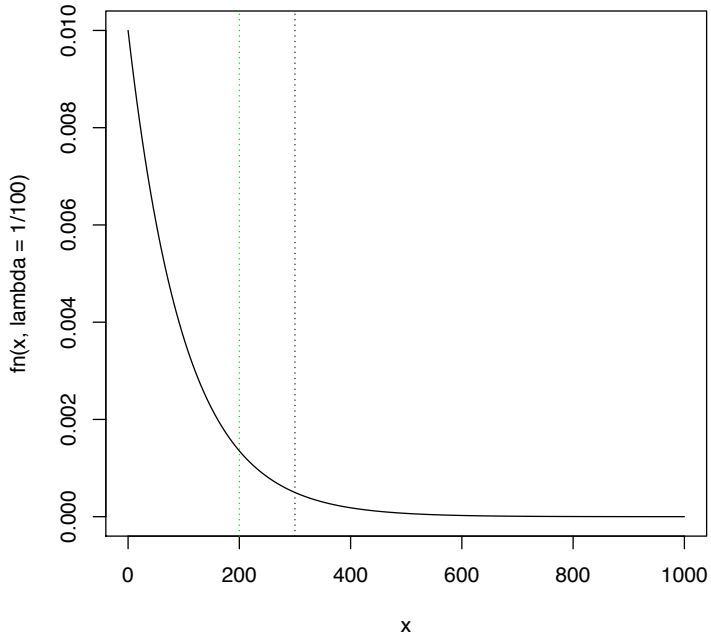

Figure 3.4: The memoryless property of the exponential distribution. The graph after point 300 is an exact copy of the original graph (this is not obvious from the graph, but redoing the graph starting from 300 makes this clear, see figure 3.5 below).

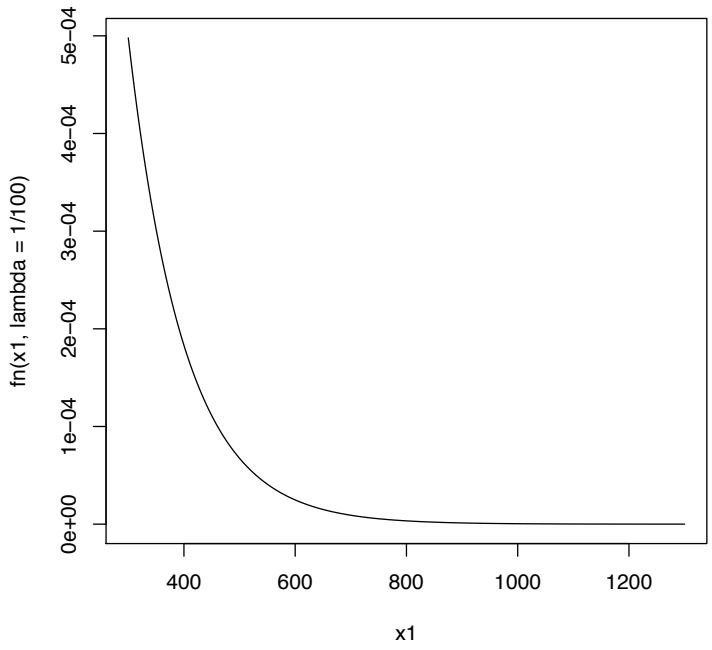

Figure 3.5: Replotting the distribution starting from 300 instead of 0, and extending the x-axis to 1300 instead of 1000 (the number in figure 3.4) gives us an exact copy of original. This is the meaning of the memoryless property of the distribution.

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

$$(1-p)^{x-1}p \tag{3.73}$$

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

$$P(X > a) = (1-p)^a \tag{3.74}$$

[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,

$$P(X > t) = (1-p)^t \tag{3.75}$$

and

$$P(X > t+a) = (1-p)^{t+a} \tag{3.76}$$

Now, we plug in the values for the right-hand side in equation 3.71, repeated below:

$$P(X > t+a \mid X > a) = \frac{P(X > t+a)}{P(X > a)} = \frac{(1-p)^{t+a}}{(1-p)^a} = (1-p)^t \tag{3.77}$$

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

$$P(X > t+a \mid X > a) = P(X > t) \tag{3.78}$$

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

■

3.9.5.2 Prove the memorylessness property for Gamma and Exponential distributions

Exponential:

The CDF is:

$$P(a) = 1 - e^{-\lambda a} \tag{3.79}$$

Therefore:

$$P(X > s+t) = 1 - P(s+t) = 1 - (1 - e^{-\lambda(s+t)}) = e^{-\lambda(s+t)} = e^{-\lambda s} e^{-\lambda t} = P(X > s)P(X > t) \tag{3.80}$$

The above is the definition of memorylessness.

■

Gamma distribution:
The CDF (not sure how this comes about, see Ross) is

$$F(x;\alpha,\beta) = 1 - \sum_{i=0}^{\alpha-1} \frac{1}{i!}(\beta x)^i e^{-\beta x} \tag{3.81}$$

Therefore,

$$P(X > s+t) = 1 - P(X < s+t) = 1 - (1 - \sum_{i=0}^{\alpha-1} \frac{1}{i!}(\beta(s+t))^i e^{-\beta(s+t)}) = \sum_{i=0}^{\alpha-1} \frac{1}{i!}(\beta(s+t))^i e^{-\beta(s+t)} \tag{3.82}$$

3.9.6 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

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1. \tag{3.83}$$

The mean and variance are

$$E[X] = \frac{a}{a+b} \text{ and } Var(X) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{3.84}$$

to-do: plot beta with different a,b.

3.9.7 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:

$$Y = g(X) \tag{3.85}$$

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:

$$\int_x^{x+\Delta x} f_X(s)\,ds = \int_y^{y+\Delta y} f_Y(t)\,dt \tag{3.86}$$

Therefore, for small Δx:

$$f_Y(y)\Delta y \approx f_X(x)\Delta x \tag{3.87}$$

Dividing by Δy we get:

$$f_Y(y) \approx f_X(x)\frac{\Delta x}{\Delta y} \tag{3.88}$$

Theorem 3 (Theorem 7.1 in Ross). *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 $g(y-y)$.

[to-do: Ross writes $\frac{d}{dx} g^{-1}(y)$ as an absolute $\mid \frac{d}{dx} g^{-1}(y) \mid$. Need to understand why.]

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

$$\begin{aligned} F_Y(y) =& P(g(X) \leq y) \\ =& P(X \leq g^{-1}(y)) \\ =& F_X(g^{-1}(y)) \end{aligned} \tag{3.89}$$

Differentiation gives

$$f_Y(y) = f_X(g^{-1}(y)) \frac{d(g^{-1}(y))}{dy} \tag{3.90}$$

Detailed explanation for the above equation: Since

$$\begin{aligned} F_Y(y) =& F_X(g^{-1}(y)) \\ =& \int f_X(g^{-1}(y))\, dy \end{aligned} \tag{3.91}$$

Differentiating:

$$\frac{d(F_Y(y))}{dy} = \frac{d}{dy}(F_X(g^{-1}(y))) \tag{3.92}$$

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. $Y = aX + b$ (see document gst2.pdf)

3.9.8 χ^2 distribution

to-do

3.9.9 t distribution

to-do

3.9.10 F distribution

to-do

3.9.11 The Poisson distribution

As Kerns [5] 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

$$f_X(x) = \mathbb{P}(X = x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0,1,2,\ldots \tag{3.93}$$

3.9.11.1 Poisson conditional probability and binomial

If $X_1 \sim Pois(\lambda_1)$ and $X_2 \sim Pois(\lambda_2)$ are independent and $Y = X_1 + X_2$, then the distribution of X_1 conditional on $Y = y$ is a binomial. Specifically, $X_1 \mid Y = y \sim Binom(y, \lambda_1)/(\lambda_1, \lambda_2)$. 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

$$X_i \mid \sum_{j=1}^{n} X_j \sim Binom(\sum_{j=1}^{n} X_j, \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}) \tag{3.94}$$

[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.

3.9.12 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:

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

3.9.12.1 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].

3.9.13 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 $P(X > n - 0.5)$
to-do: show graphically why.

3.10 Limit theorems

3.10.1 Chebyshev's inequality

Chebyshev's inequality states that if X is a random variable with finite mean μ and variance σ^2, then, for any value $k > 0$,

$$P(|\, X - \mu \,|) \geq k) \leq \frac{\sigma^2}{k^2} \tag{3.95}$$

3.10.2 Central Limit Theorem

The Central Limit Theorem is as follows:

Let $X_1, X_2, \ldots$ be a sequence of iid random variables, each having mean μ and variance σ^2. Then the distribution of

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \tag{3.96}$$

tends to the standard normal as $n \to \infty$. That is, $-\infty < a < \infty$.

$$P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} \quad \text{as } n \to \infty \tag{3.97}$$

3.11 Jointly distributed random variables

3.11.1 Joint distribution functions

3.11.1.1 Discrete case

[This section is an extract from [5].]

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

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y), \quad \text{for } (x,y) \in S_{X,Y}. \tag{3.98}$$

Every joint PMF satisfies

$$f_{X,Y}(x,y) > 0 \text{ for all } (x,y) \in S_{X,Y}, \tag{3.99}$$

and

$$\sum_{(x,y)\in S_{X,Y}} f_{X,Y}(x,y) = 1. \tag{3.100}$$

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

$$\begin{aligned} f_X(x) &= \mathbb{P}(X = x), & (3.101)\\ &= \sum_{y\in S_Y} \mathbb{P}(X = x, Y = y), & (3.102)\\ &= \sum_{y\in S_Y} f_{X,Y}(x,y). & (3.103) \end{aligned}$$

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

$$f_Y(y) = \sum_{x\in S_X} f_{X,Y}(x,y). \tag{3.104}$$

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.[2]

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

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

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 3.98. The joint CDF is typically not used in practice due to its inconvenient form; one can usually get by with the joint PMF alone.

[2] We are not at a total loss, however. There are Frechet bounds which pose limits on how large (and small) the joint distribution must be at each point.

Examples from [5]: 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 [5], worth study.

3.11.1.2 Continuous case

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

$$F(a,b) = P(X \le a, Y \le b) \quad -\infty < a,b < \infty \tag{3.105}$$

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

1. The CDF of X:

$$F_X(a) = P(X \le a) = F_X(a,\infty) \tag{3.106}$$

2. The CDF of Y:

$$F_Y(a) = P(Y \le b) = F_Y(\infty,b) \tag{3.107}$$

Definition 6. ***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:*

$$P((X,Y) \in C) = \iint\limits_{(x,y)\in C} f(x,y)\,dx\,dy \tag{3.108}$$

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

$$f(x,y) \geq 0 \text{ for all } (x,y) \in S_{X,Y}, \tag{3.109}$$

and

$$\iint\limits_{S_{X,Y}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1. \tag{3.110}$$

For any sets of real numbers A and B, and if $C = \{(x,y) : x \in A, y \in B\}$, it follows from equation 3.108 that

$$P((X \in A, Y \in B) \in C) = \int_B \int_A f(x,y)\,dx\,dy \tag{3.111}$$

Note that

$$F(a,b) = P(X \in (-\infty,a]), Y \in (-\infty,b])) = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)\,dx\,dy \tag{3.112}$$

Differentiating, we get the joint pdf:

$$f(a,b) = \frac{\partial^2}{\partial a \partial b} F(a,b) \tag{3.113}$$

One way to understand the joint PDF:

$$P(a < X < a+da, b < Y < b+db) = \int_b^{d+db}\int_a^{a+da} f(x,y)\,dx\,dy \approx f(a,b)dadb \tag{3.114}$$

[to-do: show this graphically]

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

3.11.1.3 Marginal probability distribution functions

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

$$\begin{aligned} P(X \in A) =& P(X \in A, Y \in (-\infty, \infty)) \\ =& \int_A \int_{-\infty}^{\infty} f(x,y)\, dy\, dx \\ =& \int_A f_X(x)\, dx \end{aligned} \tag{3.115}$$

where

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy \tag{3.116}$$

Similarly:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx \tag{3.117}$$

3.11.1.4 Independent random variables

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

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{3.118}$$

In the jointly continuous case:

$$f(x,y) = f_X(x) f_Y(y) \quad \text{for all } x, y \tag{3.119}$$

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 5.

Easy-to-understand example from [5]: 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

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

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

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

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 ($f_X(x)f_Y(y)$).

3.11.1.5 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:

$$\begin{aligned} F_{X+Y}(a) =& P(X+Y \leq a) \\ =& \iint_{x+y\leq a} f_{XY}(x,y)\,dx\,dy \\ =& \iint_{x+y\leq a} f_X(x) f_Y(y)\,dx\,dy \\ =& \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x) f_Y(y)\,dx\,dy \\ =& \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)\,dx f_Y(y)\,dy \\ =& \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\,dy \end{aligned} \tag{3.120}$$

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}:

$$\begin{aligned} f_{X+Y} =& \frac{d}{dx}\int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\,dy \\ =& \int_{-\infty}^{\infty} \frac{d}{dx} F_X(a-y) f_Y(y)\,dy \\ =& \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\,dy \end{aligned} \tag{3.121}$$

to-do: don't know how a differential can be moved inside an integral (never seen that before and I didn't know that was possible to do).

to-do: examples of diff. distributions

3.11.2 Conditional distributions

3.11.2.1 Discrete case

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

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad \text{if } \mathbb{P}(A) > 0. \tag{3.122}$$

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

$$\begin{aligned} p_{X|Y}(x \mid y) =& P(X = x \mid Y = y) \\ =& \frac{P(X = x, Y = y)}{P(Y = y)} \\ =& \frac{p(x,y)}{p_Y(y)} \end{aligned} \tag{3.123}$$

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:

$$\begin{aligned} F_{X|Y} =& P(X \leq x \mid Y = y) \\ =& \sum_{a \leq x} p_{X|Y}(a \mid y) \end{aligned} \tag{3.124}$$

If X and Y are independent then

$$p_{X|Y}(x \mid y) = P(X = x) = p_X(x) \tag{3.125}$$

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".

3.11.2.2 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

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)} \tag{3.126}$$

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

$$\begin{aligned} f_{X|Y}(x \mid y)\,dx &= \frac{f(x,y)}{f_Y(y)}\frac{dxdy}{dy} \\ &= \frac{f(x,y)dxdy}{f_Y(y)dy} \\ &= \frac{P(x < X < d+dx, y < Y < y+dy)}{y < P < y+dy} \end{aligned} \tag{3.127}$$

3.11.3 Joint and marginal expectation

[Taken nearly verbatim from [5].]

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:

$$\mathbb{E}\,g(X,Y) = \sum_{(x,y)\in S_{X,Y}}\sum g(x,y)\,f_{X,Y}(x,y). \tag{3.128}$$

Continuous case:

$$\mathbb{E}\,g(X,Y) = \iint_{S_{X,Y}} g(x,y)\,f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y, \tag{3.129}$$

3.11.3.1 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 7. *The **covariance** of X and Y is*

$$Cov(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y). \tag{3.130}$$

Shortcut formula for covariance:

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y). \tag{3.131}$$

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

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{3.132}$$

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

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

and the expected value of Y is

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

Finally, the expected value of XY is

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

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

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

3.11.4 Conditional expectation

Recall that

$$f_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \tag{3.133}$$

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

It follows that

$$\begin{aligned} E[X \mid Y=y] &= \sum_x x P(X=x \mid Y=y) \\ &= \sum_x x p_{X|Y}(x \mid y) \end{aligned} \tag{3.134}$$

$E[X \mid Y]$ 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.

3.11.4.1 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 (to-do: spell out the other equations):

Example 1: to-do: develop some specific examples.

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

Example 2:

$$E\left[\sum_{i=1}^n X_i \mid Y=y\right] = \sum_{i=1}^n E[X_i \mid Y=y] \tag{3.135}$$

Proposition 6. ***Expectation of the conditional expectation***

$$E[X] = E[E[X \mid Y]] \tag{3.136}$$

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

$$E[X] = \sum_y E[X \mid Y=y] P(Y=y) \tag{3.137}$$

3.11.5 Multinomial coefficients and multinomial distributions

[Taken almost verbatim from [5], 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

$$f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \binom{n}{x_1\, x_2 \cdots x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \tag{3.138}$$

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

$$(X_1, \ldots, X_k) \sim \mathsf{multinom}(\mathtt{size} = n, \mathtt{prob} = \mathbf{p}_{\mathrm{k}\times 1}). \tag{3.139}$$

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

$$\binom{n}{x_1\, x_2 \cdots x_k} = \frac{n!}{x_1!\, x_2! \cdots x_k!} \tag{3.140}$$

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 $i = 1,\ldots,6$, 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:

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \binom{n}{x_1 x_2 \cdots x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \quad (3.141)$$

Plugging in the values:

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \binom{9}{3\,2\,2\,1\,1\,0} \frac{1}{6}^3 \frac{1}{6}^2 \frac{1}{6}^2 \frac{1}{6}^1 \frac{1}{6}^1 \frac{1}{6}^0 \quad (3.142)$$

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

3.11.6 Multivariate normal distributions

Recall that in the univariate case:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}e\{-\frac{(\frac{(x-\mu)}{\sigma})^2}{2}\}} \quad -\infty < x < \infty \quad (3.143)$$

We can write the power of the exponential as:

$$(\frac{(x-\mu)}{\sigma})^2 = (x-\mu)(x-\mu)(\sigma^2)^{-1} = (x-\mu)(\sigma^2)^{-1}(x-\mu) = Q \quad (3.144)$$

Generalizing this to the multivariate case:

$$Q = (x-\mu)'\Sigma^{-1}(x-\mu) \quad (3.145)$$

So, for multivariate case:

$$f(x) = \frac{1}{\sqrt{2\pi det\Sigma}e\{-Q/2\}} \quad -\infty < x_i < \infty, i = 1,\ldots,n \quad (3.146)$$

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.

Chapter 4

Statistics

4.1 Histograms by hand

$$\text{density} = \frac{frequency}{classwidth} \tag{4.1}$$

4.2 Means for grouped data

$$\bar{x} = \frac{\sum_j f_i y_j}{\sum_j f_j} = \sum (relfreq)_j y_j \tag{4.2}$$

f_j=freq of class j; y_j midpoint of class j; relfreq$_j$ relative freq. of class j.

4.3 Standard deviation shortcut for ungrouped data

$$s^2 = \frac{1}{n-1}\left\{ \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \right\} = \frac{s_{xx}}{n-1} \tag{4.3}$$

$$s_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = \sum (x-\bar{x})^2 \tag{4.4}$$

4.4 Variance approximation for grouped data

The variance for grouped data can be computed/approximated by the formula:

$$s^2 = \frac{1}{n-1}(\sum_j f_j y_j^2 - \frac{(\sum f_j y_j)^2}{n}) \tag{4.5}$$

4.5 Pearson correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.6}$$

The left-hand side in the denominator is

$$s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4.7}$$

The right-hand side in the denominator is

$$s_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{4.8}$$

The term in the numerator is hybrid of the other two, and is called s_{xy}:

$$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \tag{4.9}$$

Note that $\frac{\sum x_i \sum y_i}{n} = n\bar{x}\bar{y}$.
It follows that the formula for r can be written as:

$$r = \frac{s_{xy}}{\sqrt{s_{xx} \times s_{yy}}} \tag{4.10}$$

Linear regression parameter estimates by hand:
Given a regression line

$$y = a + b.x \tag{4.11}$$

Moore and McCabe have the slope of a regression line b as:

$$b = r \times \frac{s_y}{s_x} \tag{4.12}$$

We can rewrite this as

$$b = \frac{s_{xy}}{s_{xx}} \tag{4.13}$$

The intercept a is

$$a = \bar{y} - b.\bar{x} \tag{4.14}$$

4.6 Contingency tables

marginal distribution, joint distribution, and conditional distribution.

4.7 The distribution of the mean

Definition 8. *If $X_1,\ldots,X_n$ are independent and identically distributed random variables, we say that they constitute a **random sample** from the infinite population given by their common distribution.*

Definition 9. *If $X_1,\ldots,X_n$ constitute a random sample, then*

$$\bar{X} = \sum X_i/n \tag{4.15}$$

*is the **sample mean** and*

$$S^2 = \sum (X_i - \bar{X})^2/n - 1 \tag{4.16}$$

*is the **sample variance**.*
*The sample mean and sample variance are called **statistics**.*

Theorem 4. *If $X_1,\ldots,X_n$ constitute a random sample from an infinite population with mean μ and variance σ^2, then*

$$E[\bar{X}] = \mu \tag{4.17}$$

and

$$Var(\bar{X}) = \frac{\sigma^2}{n} \tag{4.18}$$

[See p. 266 of [6] for proof.]

It is customary to write $E[\bar{X}]$ as $\mu_{\bar{X}}$ and $Var(\bar{X})$ as $\sigma^2_{\bar{X}}$ and to call $\sigma_{\bar{X}}$ the **standard error of the mean**.

4.8 Point estimation

$\bar{X}$ is the point estimator (a random variable; a statistic), and μ is the point (not an RC) we are trying to estimate, and $\bar{x}$ is a point estimate of μ (in one single sample, this is a single value).

S^2 is the point estimator (a random variable; a statistic), and σ^2 is the point (not an RV) we are trying to estimate, and s^2 is a point estimate of σ^2 (in one single sample, this is a single value).

Properties of estimators:

1. Unbiased
2. Efficient (Minimum variance)
3. Consistent
4. Sufficient
5. Robust

4.8.1 Unbiased estimators

4.8.1.1 S^2

Theorem 5. *If S^2 is the variance of a random sample from an infinite population with the finite variance σ^2, then $E(S^2) = \sigma^2$*

[Proof on p. 321 of Freund.]

Note that S^2 is not an unbiased estimator of σ^2 if the population is finite, and in both infinite and finite population cases, S is not an unbiased estimator of σ.

4.8.1.2 Efficiency, minimum variance

If we have to choose between several unbiased estimators of a given parameter, we usually take the one with the minimum variance; i.e., we check whether the unbiased estimator is the minimum variance unbiased estimator (best unbiased estimator).

Fact 1. ***Cramér-Rao inequality****:*
If $\hat{\theta}$ is an unbiased estimator of θ, the variance of $\hat{\theta}$ must satisfy the inequality

$$var(\hat{\theta}) \geq \frac{1}{n\dot{E}\left[\left(\frac{\partial \ln f(X)}{\partial \theta}\right)^2\right]} \tag{4.19}$$

where $f(x)$ is the value of the population density at X and n is the size of the random sample.

Theorem 6. *$\hat{\theta}$ is a minimum variance unbiased estimator of θ if $\hat{\theta}$ is an unbiased estimator of θ and*

$$var(\hat{\theta}) = \frac{1}{n\dot{E}\left[\left(\frac{\partial \ln f(X)}{\partial \theta}\right)^2\right]} \tag{4.20}$$

The denominator $\frac{1}{n\dot{E}\left[\left(\frac{\partial \ln f(X)}{\partial \theta}\right)^2\right]}$ is the information about θ that is supplied by the sample. The smaller the variance, the greater the information.

Given two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of the parameter θ of a given population, and the variance of the first is less than the variance of the second, we say that the first one is relatively more efficient. We use the ratio of the variances as a measure of relative efficiency.

$$\frac{var(\hat{\theta}_1)}{var(\hat{\theta}_2)} \tag{4.21}$$

4.9 Type I, II, power

Definitions:

Type I error probability:

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true}) = P(R \mid H_0) \tag{4.22}$$

Type II error probability:

$$\beta = P(\text{accept } H_0 \mid H_0 \text{ false}) = P(A \mid not H_0) \tag{4.23}$$

Power:

$$1-\beta = P(\text{reject } H_0 \mid H_0 \text{ false}) = P(R \mid not H_0) \tag{4.24}$$

4.9.1 Computing the power function

As an example, let $H_0 : \mu = 93$, and let $H_1 : \mu \neq 93$. Assume that population sd σ and sample size n are given. Note that in realistic situations we don't know σ but we can estimate it using s.

We can get a sample mean that is greater than μ or one that is smaller than μ. Call these $\bar{x}_g$ and $\bar{x}_s$ respectively.

In the case where we know σ, the test **under the null hypothesis** is:

$$\frac{\bar{x}_g - 93}{\sigma/\sqrt{n}} > 1.96 \quad \text{or} \quad \frac{\bar{x}_s - 93}{\sigma/\sqrt{n}} > -1.96 \tag{4.25}$$

Solving for the two $\bar{x}$'s, we get:

$$\bar{x}_g > 1.96\frac{\sigma}{\sqrt{n}} + 93 \quad \text{or} \quad \bar{x}_s > -1.96\frac{\sigma}{\sqrt{n}} + 93 \tag{4.26}$$

Now, power is the probability of rejecting the null hypothesis when the mean is whatever the alternative hypothesis mean is (say some specific value μ).

That, the test **under the alternative hypothesis** is:

$$\frac{\bar{x}_g - \mu}{\sigma/\sqrt{n}} > 1.96 \quad or \quad \frac{\bar{x}_s - \mu}{\sigma/\sqrt{n}} < -1.96 \tag{4.27}$$

We can replace the $\bar{x}_g$ with its full form, and do the same with $\bar{x}_s$.

$$\frac{1.96\frac{\sigma}{\sqrt{n}} + 93 - \mu}{\sigma/\sqrt{n}} > 1.96 \quad or \quad \frac{-1.96\frac{\sigma}{\sqrt{n}} + 93 - \mu}{\sigma/\sqrt{n}} < -1.96 \tag{4.28}$$

I can rewrite the above as:

$$\frac{1.96\frac{\sigma}{\sqrt{n}} - (\mu - 93)}{\sigma/\sqrt{n}} > 1.96 \quad or \quad \frac{-1.96\frac{\sigma}{\sqrt{n}} - (\mu - 93}{\sigma/\sqrt{n}} < -1.96 \tag{4.29}$$

Simplifying:

$$1.96 - \frac{(\mu - 93)}{\sigma/\sqrt{n}} > 1.96 \quad or \quad -1.96 - \frac{(\mu - 93)}{\sigma/\sqrt{n}} < -1.96 \tag{4.30}$$

This is now easy to solve! I will use R's pnorm function in the equation below, simply because we haven't introduced a symbol for pnorm in this course. We can rewrite the above expression as:

$$[1 - pnorm(1.96 - \frac{(\mu - 93)}{\sigma/\sqrt{n}})] + pnorm(-1.96 - \frac{(\mu - 93)}{\sigma/\sqrt{n}}) \tag{4.31}$$

The above equation allows us to

- compute sample size for any given null and alternative hypotheses, provided I have the population standard deviation.
- compute power given a null and alternative hypothesis, population standard deviation, and sample size.

Example: suppose I need power of 0.99 for $H_0 : \mu = 93$ and $H_1 : \mu = 98$, $\sigma = 5$.

For this example, what sample size do I need? I take the above equation and fill in the values:

$$[1 - pnorm(1.96 - \frac{(98 - 93)}{5/\sqrt{n}})] + pnorm(-1.96 - \frac{(98 - 93)}{5/\sqrt{n}}) \tag{4.32}$$

Simplifying, this gives us:

$$[1 - pnorm(1.96 - \sqrt{n})] + pnorm(-1.96 - \sqrt{n}) \tag{4.33}$$

Note that the second term will be effectively zero for some reasonable n like 10:

```
> pnorm(-1.96-sqrt(10))
[1] 1.5093e-07
```

So we can concentrate on the first term:

$$[1 - pnorm(1.96 - \sqrt{n})] \tag{4.34}$$

If the above has to be equal to 0.99, then

$$pnorm(1.96 - \sqrt{n}) = 0.01 \tag{4.35}$$

So, we just need to find the value of the z-score that will give us a probability of approximately 0.01. You can do this analytically (exercise), but you could also play with some values of n to see what you get. The answer is $n = 18$.

```
> pnorm(1.96-sqrt(18))
[1] 0.011226
```

4.10 Methods of inference

4.10.1 Methods of Moments

The methods of moments consists of equating the first few moments of a population to the corresponding moments of a sample, thus getting as many equations as are needed to solve for the unknown parameters of the population.

Definition 10. *The kth **sample moment** of a set of observations $x_1, x_2, \ldots, x_n$ is the mean of their kth powers and it is denoted by m'_k.*

$$m'_k = \frac{\sum_{i=1}^{n} x_i^k}{n} \tag{4.36}$$

If a population has r parameters, the method of moments consists of solving the following system of equations for the r parameters:

$$m'_k = \mu'_k \quad k = 1, 2, \ldots, r \tag{4.37}$$

4.10.2 Method of maximum likelihood

Here, 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 $x_1, x_2, \ldots, x_n$. The probability of getting them is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta) \quad (4.38)$$

i.e., the function f is the value of the joint probability **distribution** of the random variables $X_1, \ldots, X_n$ at $X_1 = x_1, \ldots, X_n = x_n$. Since the sample values have been observed and are fixed, $f(x_1, \ldots, x_n; \theta)$ 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 11. *If $x_1, x_2, \ldots, x_n$ are the values of a random sample from a population with parameter θ, the **likelihood function** of the sample is given by*

$$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta) \quad (4.39)$$

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.

4.10.2.1 Finding maximum likelihood estimates for different distributions

4.10.2.1.1 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.

$$L(\sigma) = \prod f(x_i; \sigma) = \frac{1}{(2\sigma)^n} exp(-\sum \frac{|\, x_i \,|}{\sigma}) \quad (4.40)$$

Let ℓ be log likelihood. Then:

$$\ell(x; \sigma) = \sum \left[-\log 2 - \log \sigma - \frac{|\, x_i \,|}{\sigma} \right] \quad (4.41)$$

Differentiating and equating to zero to find maximum:

$$\ell'(\sigma) = \sum \left[-\frac{1}{\sigma} + \frac{|\, x_i \,|}{\sigma^2} \right] = -\frac{n}{\sigma} + \frac{|\, x_i \,|}{\sigma^2} = 0 \tag{4.42}$$

Rearranging the above, the MLE for σ is:

$$\hat{\sigma} = \frac{\sum |\, x_i \,|}{n} \tag{4.43}$$

4.10.2.1.2 Exponential

$$f(x;\lambda) = \lambda exp(-\lambda x) \tag{4.44}$$

Log likelihood:

$$\ell = n \log \lambda - \sum \lambda x_i \tag{4.45}$$

Differentiating:

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum x_i = 0 \tag{4.46}$$

$$\frac{n}{\lambda} = \sum x_i \tag{4.47}$$

I.e.,

$$\frac{1}{\hat{\lambda}} = \frac{\sum x_i}{n} \tag{4.48}$$

4.10.2.1.3 Cauchy

Given:

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x-a)^2} \tag{4.49}$$

$$\ell(a) = n \log \pi - \sum [1 + (x_i - a)^2] \tag{4.50}$$

Differentiating:

$$\ell'(a) = \sum \frac{2(x_i - a)^2}{1 + (x_i - a)^2} = 0 \tag{4.51}$$

"This is an equation of degree $2N-1$ in a, and up to $2N-1$ different solutions may exist, N of which will correspond to maxima of the Likelihood Function. Usually the best value corresponding to the highest maximum of L is near to the sample median. This median may therefore be taken as the starting value in a iterative search for the maximum of L." (source: notes found on internet, course5.pdf)

4.10.2.1.4 Poisson

$$L(\mu;x) = \prod \frac{\exp^{-\mu}\mu^{x_i}}{x_i!} \tag{4.52}$$

$$= \exp^{-\mu}\mu^{\sum x_i}\frac{1}{\prod x_i!} \tag{4.53}$$

Log lik:

$$\ell(\mu;x) = -n\mu + \sum x_i \log\mu - \sum \log y! \tag{4.54}$$

Differentiating:

$$\ell'(\mu) = -n + \frac{\sum x_i}{\mu} = 0 \tag{4.55}$$

Therefore:

$$\hat{\lambda} = \frac{\sum x_i}{n} \tag{4.56}$$

4.10.2.1.5 Uniform

4.10.2.1.6 Binomial

$$L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{4.57}$$

Log lik:

$$\ell(\theta) = \log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta) \tag{4.58}$$

Differentiating:

$$\ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \tag{4.59}$$

Thus:

$$\hat{\theta} = \frac{x}{n} \tag{4.60}$$

4.10.2.1.7 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.

$$L(\mu; \sigma^2) = \prod N(x_i; \mu, \sigma) \tag{4.61}$$

$$= (\tfrac{1}{\sigma\sqrt{2\pi}})^n \exp(-\tfrac{1}{2\sigma^2}\sigma(x_i\mu)^2) \tag{4.62}$$

$$\tag{4.63}$$

Taking logs and differentiating with respect to μ and σ, we get:

$$\hat{\mu} = \frac{1}{n}\sum x_i = \bar{x} \tag{4.64}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 \tag{4.65}$$

Note that we did not show that $\hat{\sigma}$ is an MLE of σ. But MLEs have the invariance property: if $\hat{\Theta}$ is a maximum likelihood estimator of θ, and the function $g(\hat{\Theta})$ is continuous, then $g(\hat{\Theta})$ is also an ML estimator of $g(\theta)$.

4.10.2.1.8 Geometric

$$f(x; p) = (1-p)^{x-1} p \tag{4.66}$$

$$L(p) = p^n (1-p)^{\sum x - n} \tag{4.67}$$

Log lik:

$$\ell(p) = n \log p + (\sum x - n) \log(1-p) \tag{4.68}$$

Differentiating:

$$\ell'(p) \frac{n}{p} - \frac{\sum x - n}{1-p} = 0 \tag{4.69}$$

$$\hat{p} = \frac{1}{\bar{x}} \tag{4.70}$$

4.10.2.1.9 Rayleigh

4.10.2.1.10 Pareto

4.11 Hypothesis testing

4.11.1 Neyman-Pearson lemma

[Taken almost verbatim from this (best) presentation I could find: https://onlinecourses.science.psu.edu/stat414

The Neyman Pearson Lemma guarantees to us that each of the tests we use is the most powerful test for testing statistical hypotheses about the parameter under the assumed probability distribution.

Simple vs composite hypotheses: $H_0 : \mu = \mu_0$ is simple, $\mu > \mu_0$ composite.

Best critical region:

Consider the test of the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_A : \theta = \theta_A$. Let C and D be critical regions of size α, that is, let:

$$\alpha = P(C; \theta_0) \text{ and } \alpha = P(D; \theta_0) \tag{4.71}$$

Then, C is a best critical region of size α if the power of the test at $\theta = \theta_0$ is the largest among all possible hypothesis tests. More formally, C is the best critical region of size α if, for every other critical region D of size α, we have:

$$P(C; \theta_a) \geq P(D; \theta_a) \tag{4.72}$$

That is, C is the best critical region of size α if the power of C is at least as great as the power of every other critical region D of size α. We say that C is the most powerful size α test.

Neyman-Pearson Lemma

The Neyman Pearson Lemma. Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a probability distribution with parameter θ. Then, if C is a critical region of size α and k is a constant such that:

$$L(\theta_0)/L(\theta_a) \leq k \text{ inside the critical region C} \tag{4.73}$$

and:

$$L(\theta_0)/L(\theta_a) \geq k \text{ outside the critical region C} \tag{4.74}$$

then C is the best, that is, most powerful, critical region for testing the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_A : \theta = \theta_a$.

to-do: proof, also of the discrete case

Example to clarify what the Lemma means:

Suppose X is a single observation (that's one data point!) from a normal population with unknown mean μ and known standard deviation $\sigma = 1/3$. Then, we can apply the Neyman Pearson Lemma when testing the simple null hypothesis $H_0 : \mu = 3$ against the simple alternative hypothesis $H_A : \mu = 4$. The lemma tells us that, in order to be the most powerful test, the ratio of the likelihoods:

$$L(\theta_0)/L(\theta_a) = L(3)/L(4) \tag{4.75}$$

should be small for sample points X inside the critical region C ("less than or equal to some constant k") and large for sample points X outside of the critical region ("greater than or equal to some constant k"). In this case, because we are dealing with just one observation X, the ratio of the likelihoods equals the ratio of the normal probability curves:

$$L(3)/L(4) = f(x;3,1/9)/f(x;4,1/9) \tag{4.76}$$

Then, the following drawing summarizes the situation:

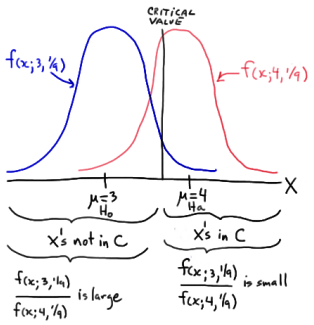

Figure 4.1: Illustration of Neyman-Pearson lemma (figure from the PSU website mentioned above).

In short, it makes intuitive sense that we would want to reject $H_0 : \mu = 3$ in favor of $H_A : \mu = 4$ if our observed x is large, that is, if our observed x falls in the critical region C. Well, as the drawing illustrates, it is those large X values in C

for which the ratio of the likelihoods is small; and, it is for the small X values not in C for which the ratio of the likelihoods is large. Just as the Neyman Pearson Lemma suggests!

Two examples of how the Lemma is used:
to-do

4.11.1.1 Uniformly most powerful tests

to-do

4.11.2 Likelihood ratio tests

Chapter 5

Notes from Statistical Inference by Juarez

p. 4 notes on notation for $P(x \mid \theta)$.

5.1 Likelihood vs probability

In a pure likelihood framework, 'likelihood (L)' is not the same as 'probability P', and 'based on' (;) is not the same as 'given' (|), although there are similarities. $L(\theta;x)$ is a function of θ for given x. It is not a probability distribution, and can sum or integrate to any positive value at all. By contrast, $f(x \mid \theta)$ is a function of x, it is the probability distribution of X for given θ. It sums or integrates to one:

$$\int_X f(x \mid \theta)\, dx = 1 \quad \text{or} \sum_X p(x \mid \theta) = 1 \tag{5.1}$$

Chapter 6

Linear modelling notes (6003)

$$\begin{aligned} L(\beta,\sigma^2;y) =& f(y;\beta,\sigma^2) \\ =& \frac{1}{2\pi\sigma^2}^{n/2} exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)\right) \end{aligned} \tag{6.1}$$

Taking logs and dropping constant:

$$\ell \propto -n\log\sigma - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) \tag{6.2}$$

Differentiating with respect to σ and equating to zero:

$$\frac{d\ell}{d\sigma} \propto -\frac{n}{\sigma} + \frac{\cancel{2}}{\cancel{2}\sigma^3}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) = 0 \tag{6.3}$$

Rearranging terms:

$$\frac{1}{\sigma^3}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) = \frac{n}{\sigma} \tag{6.4}$$

This gives us:

$$\sigma^2 = (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)n^{-1} \tag{6.5}$$

6.1 Chapter 2 notes

Given the tractor data, let's say we have the model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \tag{6.6}$$

```
> data<-read.table("data/tractordata.txt",header=T)
> colnames(data)<-c("Age","Maint")
> fm0<-lm(Maint~Age+I(Age^2)+I(Age^3)+I(Age^4),data)
> #summary(fm0)
```

Three ways to test the hypothesis $H_0 : \beta_3 = \beta_4 = 0$.

6.1.1 Method 1: Using contrast matrix

```
> ## Method 1: Using C matrix specification:
> y<-data$Maint
> X<-model.matrix(fm0)
> (XT.X <- t (X)%*%X)
            (Intercept)      Age  I(Age^2)
(Intercept)       16.00    57.50    269.25
Age               57.50   269.25   1350.88
I(Age^2)         269.25  1350.88   6973.31
I(Age^3)        1350.88  6973.31  36725.47
I(Age^4)        6973.31 36725.47 196766.20
             I(Age^3)  I(Age^4)
(Intercept)    1350.9    6973.3
Age            6973.3   36725.5
I(Age^2)      36725.5  196766.2
I(Age^3)     196766.2 1070379.4
I(Age^4)    1070379.4 5901359.9
> (XT.y <- t(X)%*%y)
               [,1]
(Intercept)   13049
Age           55411
I(Age^2)     271562
I(Age^3)    1389848
I(Age^4)    7268377
```

```
> (G<-solve(XT.X))
             (Intercept)       Age I(Age^2)
(Intercept)    5.409674 -10.54936   5.5024
Age          -10.549364  22.19135 -12.3049
I(Age^2)       5.502425 -12.30492   7.4502
I(Age^3)      -1.096444   2.56774  -1.6603
I(Age^4)       0.074665  -0.18109   0.1228
             I(Age^3)   I(Age^4)
(Intercept) -1.096444  0.0746653
Age          2.567744 -0.1810924
I(Age^2)    -1.660282  0.1228036
I(Age^3)     0.386870 -0.0294960
I(Age^4)    -0.029496  0.0022943
> (beta.hat <- G%*%XT.y)
                 [,1]
(Intercept)  -903.125
Age          2958.284
I(Age^2)    -1775.951
I(Age^3)      406.647
I(Age^4)      -30.654
> C<-matrix(c(0,0,0,1,0,0,0,0,0,1),byrow=T,ncol=5)
> c<-matrix(c(0,0),byrow=T,ncol=1)
> (yT.y <- t(y)%*%y)
         [,1]
[1,] 12910953
> (bT.XT.y<-t(beta.hat)%*%t(X)%*%y)
         [,1]
[1,] 12231043
> ## by definition of S_r:
> (Sr <- yT.y - bT.XT.y)
       [,1]
[1,] 679910
> n<-17
> p<-5  ## num parameters
> (sigma.hat.2 <- (1/(n-p)) * Sr)
```

```
      [,1]
[1,] 56659
> sqrt(sigma.hat.2)
       [,1]
[1,] 238.03
> (num<-t(C %*% beta.hat - c) %*%
             solve(C %*% G %*% t(C) ) %*%
             (C %*% beta.hat -c ))
       [,1]
[1,] 430135
> q<-2
> (F<-num/(sigma.hat.2*q))
       [,1]
[1,] 3.7958
```

The above answer matches the lecture notes.

6.1.2 Method 2: Using model comparison

```
> ## Method 2: using model comparison:
> fm0a<-lm(Maint~Age+I(Age^2),data)
> fm0b<-fm0
> (aov.output<-anova(fm0a,fm0b))
Analysis of Variance Table

Model 1: Maint ~ Age + I(Age^2)
Model 2: Maint ~ Age + I(Age^2) + I(Age^3) + I(Age^4)
  Res.Df     RSS Df Sum of Sq    F Pr(>F)
1     13 1110045
2     11  679910  2    430135 3.48  0.067
```

6.1.3 Method 3: Using ANOVA

Here, I get an answer that is slightly different from the lecture notes, and it has to do with the fact that I use exact values from the anova output.

```
> ## H0: b3=b4=0
> ## Method 2: using anova:
> ss<-anova(fm0)$"Sum Sq"
> df<-anova(fm0)$Df
> ## getting exact nos. gives a slightly different result:
> (f<-(ss[3]+ss[4]/2)/(ss[5]/12))
[1] 3.9774
> 1-pf(f,2,12)
[1] 0.047294
> ## cf. taking ceiling of each number as in lecture notes:
> f.2<-((31195+405902)/2)/(766079/12)
> 1-pf(f.2,2,12)
[1] 0.066629
```

Does it matter if I have this much deviation from the lecture notes results?

6.1.4 Residuals, leverage, outliers

- If sample size is small, then use scaled or standardized residuals.
- MSc lecture notes: "If the [QQ] plot is clearly bowed, it suggests a skew distribution (whereas the normal distribution is symmetric). If the plotted points curve down at the left and up at the right, it suggests a distribution with heavier tails than the normal (and therefore one that is prone to produce outliers)."
- Histogram plot is useful only for larger samples.
- Index plot or I-chart. MSc lecture notes: "plots residual against observation number, and indicates whether the observations might be correlated. Specifically, it helps to assess whether adjacent observations are correlated. Such a correlation would show up by observations that are adjacent in the sequence having very similar residuals. The plot would then tend to move slowly up and down. Like most diagnostics, it is not very sensitive for a sample as small as the tractor data, and in this case there is no reason to suspect this kind of correlation. However, sometimes we might suspect that observations made adjacently in time or space would show correlation, and the index plot will show this if we order the observations appropriately."

  An alternative is to compute correlation of adjacent residuals.

- Residuals against fitted values: Checks homoscedasticity. MSc lecture notes: "The most common form of heteroscedasticity arises when larger observations are also more variable. This often happens when the response variable is necessarily positive. In this case, if the response is just above 0 its variance is likely to be less than if the response was much larger since the response is bounded below by zero. We would observe this kind of heteroscedasticity in the plot by seeing the residuals appearing to fan out as we move from left to right. Other patterns might indicate other ways in which the response variance its related to its mean."

  Bartlett's test. (MSc lecture notes: "This statistic is officially for testing the equality of variances of a number of normal samples, and so the standard case is the one factor model. However, it can be applied as an approximation to variances derived from groups of residuals in a more general linear model. An example might be dividing the data into two groups according to the size of the fitted values. Note, however, that you should not use the data to identify groupings to be tested.")

- Plot residuals against explanatory variables. Here we look for extra regression structures. From lecture notes:

  "Note, however, than we would never expect to see a straight line relationship when plotting residuals against any explanatory variable used in the fitted model. This is because of the fundamental property that $X^T e = 0$. This means that if we multiply residuals by any column in the X matrix and sum, the result must always be zero. So the residuals and the explanatory variables are always uncorrelated."

- Leverage:

  "leverage measures the *potential* influence of an observation to affect the parameter estimates but we need to take into account whether the observation is outlying to assess its *actual* influence."

  "An observation will generally have high influence if it is both outlying and of high leverage. It is possible for an observation to have high influence if it is extremely outlying but not of high leverage or of extremely high leverage but not very outlying; though this is somewhat less likely."

Bibliography

[1] Leonard Evens. *A brief course in linear algebra (version 0.1)*. 2002.

[2] J. Fox. *A mathematical primer for social statistics*. Number 159. Sage Publications, Inc, 2009.

[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] G. Jay Kerns. *Introduction to Probability and Statistics Using R*. 2010.

[6] I. Miller and M. Miller. *John E. Freund's Mathematical Statistics with Applications*. Prentice Hall, 2004.

[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] Michael Spivak. *Calculus*. Cambridge, 2010.