A Primer of Probability & Statistics

Contents

0. Preface

This document provides an overview of mathematical and statistical results useful in molecular evolution, phylogenetics, and population genetics. It is prepared for two purposes. First it is to be used as the pre-course reading material for the Computational Molecular Evolution (CoME) Workshop organized by Nick Goldman, Aidan Budd, Alexandros Stamatakis, and me that runs annually in Hinxton, Cambridge or Heraklion, Crete (http://abacus.gene.ucl.ac.uk/CoME/). Second it is used as reading material for biologist readers of my books *Computational Molecular Evolution* (Yang 2006, OUP) and *Molecular Evolution: A Statistical Approach* (Yang 2014, OUP).

The notes are not comprehensive and are mostly for review. I have added R code for the graphs. You may install R, and copy the R code to reproduce the graphs. Comments and suggestions are welcome. In particular, please let me know which parts are too hard, and I will try to add more examples.

I would like to thank the following colleagues for providing assistance, making comments, correcting mistakes or helping with the translation of the glossary: Aidan Budd, Hiro Kishino, Fengrong Ren, Alexandros Stamatakis, Veronika Boskova and Tianqi Zhu.

Ziheng Yang
September 2011, updated March 2013
Last updated, May 2014

1. Mathematical preliminaries

1.1. A few words about notation

Journal publishers are increasingly leaving it to authors to do the typesetting, so it is useful to know how to write professional-looking mathematical equations using Microsoft Word.

Mathematical variables are by convention written using italic English letters, such as a, b, x, or y. It is also common to use x, y, z, w for unknowns, a, b, c for constants, i, j, k for integers or indexes, and m and n for counts.

An vector may be written in bold letters, such as $\mathbf{x} = \{x_i\}$, and a matrix in bold capital letters, such as $\mathbf{P}(t) = \{p_{ij}(t)\}$. Alternatively one may write a vector using an italic English letter and a matrix using an italic capital English letter, such as $x = \{x_i\}$, and $A = \{a_{ij}\}$, and $Q = \{q_{ij}\}$. Either is fine but you should be consistent.

There should be a small space on both sides of operators +, –, ×, and =. Arabic numerals should be written in Roman font and not italic. For example write $y = a_0 - x^2$ instead of y = a_0 - x^2 or $y=a_0-x^2$. Note the difference between '–' (minus sign) and a dash. (If you type "1 - 2" with spaces around the dash, MS Word will "auto-correct" the dash into minus.)

In statistics, small Greek letters are often used to represent parameters while English letters are for random variables. Both should ideally be written in italic. For example, the mean and variance of the normal distribution are μ and σ^2, while the shape and scale parameters of the gamma distribution are α and β. The collection (vector) of parameters in the model is often written as θ or Θ. For example, if the data are a sample from the normal distribution with mean μ and variance σ^2, we may write $\theta = \{\mu, \sigma^2\}$ as the parameters in the model. Whether θ is a scalar or a vector is typically clear from the context.

We may write $x \sim \mathrm{N}(\mu, \sigma^2)$ to mean that the random variable x follows the normal distribution with mean μ and variance σ^2. Statistical books use capital English letters (e.g., X and Y) to represent random variables, and the corresponding small letters (e.g., x and y) for their realized values. Thus $\Pr\{X < 3\}$ is the probability that random variable X is less than 3 and $\Pr\{X < x\}$ is the probability that random variable X is less than x. In non-statistical publications, we rarely bother with this distinction. Thus we write $x \sim \mathrm{N}(\mu, \sigma^2)$ and $\Pr\{x < 3\}$, and try to avoid using $\Pr\{X < x\}$.

Hotkeys for MS Word. Press Ctrl-i for *italic*, Ctrl-b for **bold**, Ctrl-+ (Ctrl-Shift-=) for superscript, and Ctrl-= for subscript. All these hotkeys are switches, so you press the same key again to remove the formatting. Try Ctrl-i (a) when some text is highlighted, (b) when the cursor is inside a word but nothing is highlighted, and (c) when the cursor is at a space and nothing is highlighted (and start typing). Use Insert-Symbol to insert Greek and mathematical symbols such as α, ×, and ∞. The symbol ℓ (for log likelihood) is in the font set MT Extra. Do not write a and then change its font to Symbol to get α. You will see the difference when you change the font for the whole paragraph from Times New Roman to Arial, say. Use Insert-Object-Microsoft Equation Editor to write equations. Do not write a symbol as an equation.

1.2. Summation and product symbols

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \,. \tag{1.1}$$

$$\sum_{i=1}^{n} cx_i = cx_1 + cx_2 + \cdots + cx_n = c\sum_{i=1}^{n} x_i \ .$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_i x_j = \left(\sum_{i=1}^{m} x_i\right)\left(\sum_{j=1}^{n} x_j\right).$$

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \ldots \cdot x_n \,. \qquad (1.2)$$

$$\prod_{i=1}^{n} i = 1 \cdot 2 \cdot \cdots \cdot n = n!.$$

$$\prod_{i=1}^{n} ax_i = (ax_1)\cdot(ax_2)\cdot\ldots\cdot(ax_n) = a^n \prod_{i=1}^{n} x_i \ .$$

$$\prod_{i=1}^{n} \mathrm{e}^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} = \exp\left\{-\tfrac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right\}.$$

$$\log\left\{\prod_{i=1}^{N} p_i^{x_i}\right\} = \log(p_1^{x_1} p_2^{x_2} \cdots p_N^{x_N}) = x_1\log(p_1) + x_2\log(p_2) + \cdots + x_N\log(p_N) = \sum_{i=1}^{N} x_i \log(p_i)\,.$$

Note that the logarithm in this note is the natural logarithm, with base e.

Summing up n positive numbers x_i as in equation (1.1) when they are too small or too large may run into numerical problems called overflows or underflows. On most computer systems we use today, the smallest nonzero double-precision floating number is 2.25×10^{-308}, so any number smaller than this will become 0 on the computer, causing an underflow. Similarly the largest double number is 1.79×10^{307}, so any number greater than this will not be possible to represent on the computer (Inf or Nan), causing an overflow. One solution is to store the logarithms of the numbers. The question is then how to calculate the logarithm of the sum given the logarithms of the numbers? In other words, given $y_i = \log\{x_i\}$, $i = 1, 2, \ldots, n$, how do we calculate $\log\left\{\sum_{i=1}^{n} x_i\right\}$?

The following table illustrates the procedure, with the x_i to be 10^{-10}, 10^{-15}, 10^{-20}, 10^{-25} (these numbers are not that small and are used for illustration here). In the computer we store the logarithms (y_i) (row 2). We find the largest of the y_i to be $y^* = -23.02585$. This is used for scaling: it does not have to be equal to but should be similar to the largest among the y_i. Subtract y^* from each y_i (row 3). Take the exponential and sum up (row 4). The final result is $\log\left\{\sum_{i=1}^{n} x_i\right\} = y^* + \log(s) = -23.02584$.

(1) x_i	10^{-10}	10^{-15}	10^{-20}	10^{-25}	
(2) $y_i = \log\{x_i\}$	–23.02585	–34.53878	–46.05170	–57.56463	Largest $y^* =$ –23.02585
(3) $z_i = y_i - y^*$	0	–11.51283	–23.02575	–34.53868	
(4) $\exp\{z_i\}$	1	10^{-5}	10^{-10}	10^{-15}	Sum s = 1.000010

Exercise 1 (Summing up large numbers). Use the above procedure to calculate the logarithm of the sum $\mathrm{e}^{1000} + \mathrm{e}^{1200} + \mathrm{e}^{1215} + \mathrm{e}^{1216}$. [The answer is 1216.31326.]

1.3. Linear algebra

An $m \times n$ matrix is written as $A = \{a_{ij}\}$. A column vector is an $n \times 1$ matrix while a row vector is a $1 \times n$ matrix. For example,

$$x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

where the superscript T stands for transpose, is a column vector, and

$$y = (y_1, y_2, \ldots, y_n)$$

is a row vector. If vectors x and y are of the same size, their *inner-product* is defined as

$$x \cdot y = \sum_{i=1}^{n} x_i y_i \,. \tag{1.3}$$

If both matrices A and B are of the same size (that is, if they have the same number of rows and the same number of columns), one can define their sum $C = A + B$, where $C = \{c_{ij}\}$ with $c_{ij} = a_{ij} + b_{ij}$.

$$I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad D = \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \end{bmatrix} \quad Q = \begin{bmatrix} q_{11} & a & b & c \\ a & q_{22} & d & e \\ b & d & q_{33} & f \\ c & e & f & q_{44} \end{bmatrix} \quad 0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Identity matrix I_4 — Diagonal matrix $D = \mathrm{diag}\{a, b, c, d\}$ — Symmetrical matrix Q, in which $q_{ij} = q_{ji}$. — A null vector

If A is $m \times n$ and B is $n \times k$, then their product exists, $C = AB$, which is of size $m \times k$. The ijth element of C is the inner product of the ith row in A and the jth column in B:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \,. \tag{1.4}$$

Exercise 2 (matrix addition and multiplication). Suppose

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \quad I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad D = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix}$$

Confirm that $IA = AI = A$. Calculate DA and AD. What pattern did you see? Calculate D^2, and D^n, for any natural number n.

Let A be a square matrix of size $n \times n$. If there exists an $n \times n$ matrix B such that

$$AB = BA = I_n, \tag{1.5}$$

where I_n is the identity matrix of size $n \times n$, then A is said to be *non-singular*, and B is the inverse of A, also written as $B = A^{-1}$. The inverse of a matrix, if it exists, is unique. A square matrix that does not have an inverse is said to be *singular* or *degenerate*. A square matrix is singular if and only if its determinant $|A| = 0$.

The system of linear equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned} \tag{1.6}$$

is written in matrix form as

$$Ax = b, \tag{1.7}$$

where x and b are column vectors of size n. The solution, if it exists, is given as

$$x = A^{-1}b. \tag{1.8}$$

Matrix diagonalization or spectral decomposition. Let A be an $n \times n$ matrix. If there exists a non-singular $n \times n$ matrix U and a diagonal matrix $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ such that

$$A = U\Lambda U^{-1}, \tag{1.9}$$

then A is said to be *diagonalizable*, $\lambda_1, \lambda_2, \ldots, \lambda_n$ are called the *eigenvalues*, *eigenroots*, *latent roots* or *characteristic roots* of A, and column vectors of U are called the right eigenvectors of A, and rows of U^{-1} are the left eigenvectors. Equation (1.9) is called the *spectral decomposition* of matrix A. Note that not all square matrices can be diagonalized.

To find the eigenvalues and eigenvectors, we solve the system of linear equations

$$Ax = \lambda x, \tag{1.10}$$

or

$$(A - \lambda I)x = 0. \tag{1.11}$$

Note here λ is a scalar, A and I are $n \times n$ matrices, and x is a non-null column vector. First we find the eigenvalues by solving the nth-order polynomial equation

$$|A - \lambda I| = 0. \tag{1.12}$$

There should be n roots, but some of them may be identical. Then use each of the n roots to find the corresponding right eigenvector x from equation (1.11), which will be a column in U. Note that if λ and x satisfy equation (1.10), so will λ and cx for any $c \neq 0$, so the eigenvectors are not unique.

Example (eigensolution for Kimura's 2-parameter model of nucleotide substitution). Find the eigensolution (spectral decomposition) of

$$Q = \begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix}.$$

One can use equation (1.12) to find the four eigenvalues to be $\lambda_1 = 0$ (the first one is always 0), $\lambda_2 = -4\beta$, and $\lambda_3 = \lambda_4 = -2(\alpha + \beta)$. Then equation (1.10) can be used with each of the eigenvalues to find the corresponding right eigenvector x, which should be the corresponding column in U. The solution is

$$Q = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -4\beta & 0 & 0 \\ 0 & 0 & -2(\alpha+\beta) & 0 \\ 0 & 0 & 0 & -2(\alpha+\beta) \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & 0 & 0 \end{bmatrix}. \tag{1.13}$$

Exercise 3 (matrix inversion). Find A^{-1} where

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

[Hint. Let $A^{-1} = \begin{bmatrix} r & s \\ t & u \end{bmatrix}$. Find r, s, t, w by solving the equations $AA^{-1} = \mathrm{I}$.]

Exercise 4 (eigensolution for the Jukes & Cantor model). Find the eigensolution of

$$Q = \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix}.$$

Hint: Use the result from the example for the K80 model.

Algebraic functions of a matrix. It is easy to see that the *n*th power of a diagonal matrix $\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is also a diagonal matrix: $\Lambda^n = \mathrm{diag}\{\lambda_1^n, \lambda_2^n, \cdots, \lambda_n^n\}$. With the spectral decomposition of A (equation 1.9), it is easy to calculate the *n*th power of A:

$$A^n = (U\Lambda U^{-1})^n = (U\Lambda U^{-1})(U\Lambda U^{-1})\cdots(U\Lambda U^{-1}) = U\Lambda^n U^{-1}, \qquad (1.14)$$

An algebraic function of a matrix is typically defined as a limit. Here we are interested in the exponential of a square matrix, which is useful in Markov chain models of nucleotide or amino acid substitution. Following the Taylor expansion of the exponential function of a scalar x

$$\mathrm{e}^x = 1 + x + \tfrac{1}{2!}x^2 + \tfrac{1}{3!}x^3 + \ldots. \qquad (1.15)$$

the exponential of an $n \times n$ matrix A is defined as

$$\mathrm{e}^A = I + A + \tfrac{1}{2!}A^2 + \tfrac{1}{3!}A^3 + \ldots. \qquad (1.16)$$

Now if the spectral decomposition of A is available (equation 1.9), we have

$$\mathrm{e}^A = U\,\mathrm{e}^\Lambda\,U^{-1}, \qquad (1.17)$$

where $\mathrm{e}^\Lambda = \mathrm{diag}\{\mathrm{e}^{\lambda_1}, \mathrm{e}^{\lambda_2}, \cdots, \mathrm{e}^{\lambda_n}\}$.

Exercise 5 (Transition probability matrix under Kimura's 2-parameter model).* Use the result of the Example (equation 1.13) to calculate e^{Qt}, where t is a scalar.

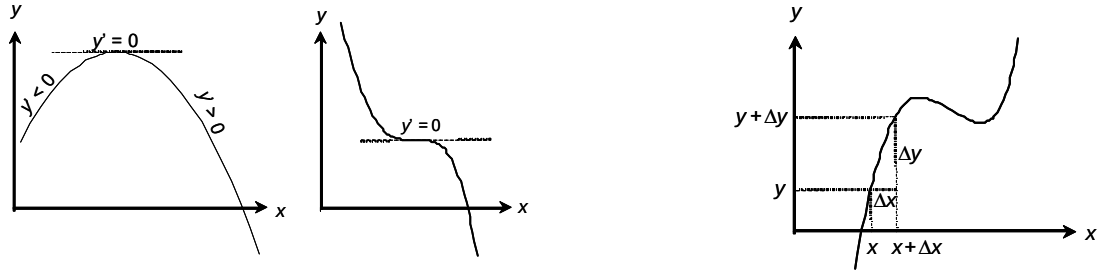
1.4. Differentiation

Unless stated otherwise, we assume that all functions discussed in this note are continuous and smooth. When we calculate the derivative of a function, the derivative is assumed to exist (or the function is differentiable). Our emphasis is on intuitive understanding rather than mathematical rigor.

Suppose $y = f(x)$ is a function of x. When x changes by Δx, y changes by Δy. Then the ratio $\Delta y/\Delta x$ measures the slope. For a straight line $y = a + bx$, the slope is b. For a curve, the slope may differ at different parts of the curve (that is, it may depend on x), and is defined as

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}. \qquad (1.18)$$

dy/dx is also written as y' or $f'(x)$. The function $y = f(x)$ is increasing if $y' > 0$ and decreasing if $y' < 0$. When a smooth function $f(x)$ reaches a minimum or maximum inside its domain, $y' = 0$. Nevertheless, the inverse is not true: $y' = 0$ does not necessarily mean a minimum or maximum (see figure). Also if the function is defined over the interval $a \le x \le b$ and if the minimum or maximum occurs at the boundary a or b, then y' may not be 0.



Suppose $y = x^2$. When x increases from x to $x + \Delta x$, y increases from x^2 to $(x + \Delta x)^2$, with $\Delta y = (x + \Delta x)^2 - x^2 = 2x\Delta x + (\Delta x)^2$, so that $\Delta y/\Delta x = 2x + \Delta x \to 2x$ when $\Delta x \to 0$. Thus $\dfrac{dx^2}{dx} = 2x$. Here are a few more examples.

$\dfrac{d}{dx}c = 0$ where c is a constant.

$\dfrac{d}{dx}x^{-1} = -x^{-2}$.

$\dfrac{d}{dx}e^x = e^x$.

$\dfrac{d}{dx}\sqrt{x} = \dfrac{d}{dx}x^{\frac{1}{2}} = \frac{1}{2} \times x^{-\frac{1}{2}} = \dfrac{1}{2\sqrt{x}}$.

$\dfrac{d}{dx}\log\{x\} = \dfrac{1}{x}$.

$\dfrac{d}{dx}a^x = a^x \log\{a\}$.

$\dfrac{d}{dx}x^a = ax^{a-1}$.

$\dfrac{d}{dx}\sin x = \cos x$.

Note that log means the natural logarithm, with base e.

The product rule. If both u and v are functions of x, then

$$(uv)' = uv' + u'v, \quad \text{or} \quad \frac{d(uv)}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}. \tag{1.19}$$

The chain rule. If y is a function of u, and u a function of x, then

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}. \tag{1.20}$$

Here are a few examples.

$$\frac{d}{dx}e^{-\frac{1}{2}x^2} = e^{-\frac{1}{2}x^2} \cdot \frac{d(-\frac{1}{2}x^2)}{dx} = e^{-\frac{1}{2}x^2}(-x).$$

If $y = x^{\alpha-1}e^{-\beta x}$, then

$$\frac{dy}{dx} = x^{\alpha-1}\frac{de^{-\beta x}}{dx} + e^{-\beta x}\frac{dx^{\alpha-1}}{dx} = x^{\alpha-1}e^{-\beta x}(-\beta) + e^{-\beta x}(\alpha-1)x^{\alpha-2} = x^{\alpha-2}e^{-\beta x}(-x\beta + \alpha - 1). \tag{1.21}$$

)

Suppose $y = x^{-\alpha-1}e^{-\beta/x}$. Calculate

$$\frac{dy}{dx} =$$

Since $f'(x)$ is a function of x, we can take its derivative, to get the second derivative $y'' = f''(x) = \frac{df'(x)}{dx}$. The nth derivative is written as $y^{(n)}$ or $f^{(n)}(x)$.

Taylor expansion. One can use a polynomial to approximate an arbitrary function $f(x)$. The Taylor expansion of the function $f(x)$ around $x = a$ is

$$f(x) = f(a) + f'(a)(x-a) + \tfrac{1}{2} f''(a)(x-a)^2 + \tfrac{1}{3!} f^{(3)}(a)(x-a)^2 + \cdots . \qquad (1.22)$$

Here $f(a)$ is the function $f(x)$ evaluated at $x = a$, $f'(a)$ is the first derivative evaluated at $x = a$: that is, $f'(a) = f'(x)\big|_{x=a}$, and so on. Two examples are

$$e^x = 1 + x + \tfrac{1}{2}x^2 + \tfrac{1}{3!}x^3 + \tfrac{1}{4!}x^4 + \ldots . \qquad (1.23)$$

$$\log(1+x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \tfrac{1}{4}x^4 + \ldots . \qquad (1.24)$$

One can use the first two or three terms of the Taylor expansion to approximate the function $f(x)$ in the neighbourhood of $x = a$, giving

$$f(x) \approx f(a) + f'(a)(x-a) . \qquad (1.25)$$

$$f(x) \approx f(a) + f'(a)(x-a) + \tfrac{1}{2} f''(a)(x-a)^2 . \qquad (1.26)$$

For example, when x is close to 0 (if $|x| \ll 1$)

$$e^x \approx 1 + x. \qquad (1.27)$$

Similarly if x is close to 0

$$\log(1 + x) \approx x. \qquad (1.28)$$

In statistics, the Taylor expansion is sometimes used to approximate the log likelihood function $\ell(\theta) = \log\{L(\theta)\}$. Suppose $\ell(\theta)$ has a mode (maximum) at $\hat{\theta}$, at which the gradient $\ell'(\hat{\theta}) = 0$. Then equation (1.26) gives the approximate value of $\ell(\theta)$ in the neighbourhood of $\hat{\theta}$ as

$$\ell(\theta) \approx \ell(\hat{\theta}) + \tfrac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2 . \qquad (1.29)$$

This means that the likelihood $L(\theta)$ is approximated by the density of a normal distribution

$$L(\theta) = e^{\ell(\theta)} \approx e^{\ell(\hat{\theta})} e^{\frac{1}{2}\ell''(\hat{\theta})(\theta-\hat{\theta})^2}, \qquad (1.30)$$

where the variance of the normal distribution is $-1/\ell''(\hat{\theta})$. The density function of the normal distribution is given later in eq. (2.21).

1.5. Integration

Integration is the inverse of differentiation.

The indefinite integral. If $F'(x) = \frac{dF(x)}{dx} = f(x)$, then

$$\int f(x)\,\mathrm{d}x = F(x) + C, \tag{1.31}$$

where C is a constant. The function $f(x)$ is called the integrand. Thus

$$\int \mathrm{e}^x \mathrm{d}x = \mathrm{e}^x + C.$$

$$\int \frac{1}{x}\mathrm{d}x = \log(x) + C.$$

$$\int x^n \mathrm{d}x = \tfrac{1}{n+1} x^{n+1} + C\ ,\ n \neq -1.$$

Integration by parts $$\int u\,\mathrm{d}v = uv - \int v\,\mathrm{d}u\,. \tag{1.32}$$

For example,

$$\begin{aligned}\int x^2 \log(x)\,\mathrm{d}x &= \int \log(x)\,\mathrm{d}(\tfrac{1}{3}x^3)\\ &= \tfrac{1}{3}x^3\log(x) - \int \tfrac{1}{3}x^3\,\mathrm{d}\log(x)\\ &= \tfrac{1}{3}x^3\log(x) - \tfrac{1}{3}\int x^3\,\tfrac{1}{x}\mathrm{d}x = \tfrac{1}{3}x^3\log(x) - \tfrac{1}{3}\int x^2\,\mathrm{d}x\\ &= \tfrac{1}{3}x^3\log(x) - \tfrac{1}{9}x^3 + C.\end{aligned}$$

Change of variables

Calculate $\int x\mathrm{e}^{-x^2/2}\mathrm{d}x$.

Let $u = -x^2/2$. Then $\mathrm{d}u = -x\,\mathrm{d}x$. Thus

$$\int x\mathrm{e}^{-x^2/2}\mathrm{d}x = -\int \mathrm{e}^u \mathrm{d}u = -\mathrm{e}^u + C = -\mathrm{e}^{-x^2/2} + C\,.$$

or one can write

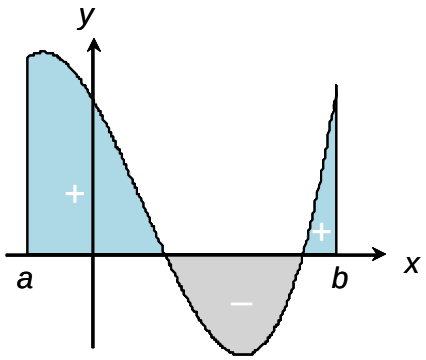
$$\int x\mathrm{e}^{-x^2/2}\mathrm{d}x = -\int \mathrm{e}^{-x^2/2}\mathrm{d}(\tfrac{-x^2}{2}) = -\int \mathrm{d}\mathrm{e}^{-x^2/2} = -\mathrm{e}^{-x^2/2} + C\,.$$

The definite integral. $I = \int_a^b f(x)\,\mathrm{d}x$ is the area under the curve $f(x)$ between $x = a$ and $x = b$, but the area below the x axis is negative. Here $f(x)$ is known as the integrand, and a and b are the integration limits. Some integrals can be calculated analytically. If $F'(x) = \dfrac{\mathrm{d}F(x)}{\mathrm{d}x} = f(x)$, then

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a). \tag{1.33}$$

For example,

$$\int_0^\infty \tfrac{1}{\mu}\mathrm{e}^{-x/\mu}\mathrm{d}x = -\mathrm{e}^{-x/\mu}\Big|_0^\infty = (-\mathrm{e}^{-\infty}) - (-\mathrm{e}^0) = 1.$$



A definite integral of a function $f(x)$ over (a, b) represents the signed area of the region bounded by the curve of the function and the x axis. In most of our applications, $f(x) > 0$, so the integral represents the area above the x-axis under the curve between a and b.

```
fx <- function(x) (x+1)*(x-2)*(x-4)
xstart=-1;  xend=4.5;  by=0.01;
x = seq(xstart, xend, by)
y = fx(x);
plot(x, y, type="n", xlab="", ylab="", frame.plot=0,
   tck=-.02, xaxp=c(0,xend,1));

x = seq(xstart, 2, by)
y = fx(x);
polygon(c(xstart,x,2),c(0,y,0), col='lightblue');
x = seq(4, xend, by)
y = fx(x);
polygon(c(4,x,xend),c(0,y,0), col='lightblue');
x = seq(2, 4, by)
y = fx(x);
polygon(c(2,x,4),c(0,y,0), col='lightgray');
```

Integrals involving the probability density function of a continuous distribution. We will describe the probability density functions for continuous distributions in section 2.3 but one thing to note here about them is that they integrate to 1, or the area under the probability density curve is 1. This result is very useful in calculating integrals involving continuous random variables. Here we give two examples.

The first involves the gamma distribution, which has density

$$g(x;\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}, \quad x > 0, \tag{1.34}$$

with $\int_{-\infty}^{\infty} g(x;\alpha,\beta)\mathrm{d}x = 1$. Suppose now we want to calculate the following integral

$$I = \int_0^\infty e^{-cx} g(x;\alpha,\beta)\,\mathrm{d}x = \int_0^\infty e^{-cx} \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}\,\mathrm{d}x\,. \tag{1.35}$$

This integral arrives in the so-called gamma distances. We rearrange the integrand so that it becomes a gamma density

$$I = \left(\frac{\beta+c}{\beta}\right)^{-\alpha} \int_0^\infty \left[\frac{(\beta+c)^\alpha}{\Gamma(\alpha)} e^{-(\beta+c)x} x^{\alpha-1} \right] \mathrm{d}x = \left(\frac{\beta+c}{\beta}\right)^{-\alpha}. \tag{1.36}$$

Note that now the integrand in brackets is the gamma density $g(x;\, \alpha, \beta + c)$ so the integral is 1.

A second example involves the normal distribution, which has density

$$\phi(x;\, \mu,\, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty, \tag{1.37}$$

with $\int_{-\infty}^{\infty} \phi(x;\mu,\sigma^2)\mathrm{d}x = 1$. Suppose we want to calculate the integral

$$M = \int_{-\infty}^{\infty} \phi(\bar{x};\mu,\tfrac{1}{n})\phi(\mu;\mu_0,\sigma^2)\mathrm{d}\mu = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi/n}} e^{-\frac{n}{2}(\bar{x}-\mu)^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mu-\mu_0)^2}\,\mathrm{d}\mu\,. \tag{1.38}$$

Here M is the marginal likelihood in Bayesian analysis of the normal model, in which the parameter is the population mean μ, and we assign a prior $\mu \sim \mathrm{N}(0, 1)$, while the data are a sample of size n from the population. The variance of the population is given as 1. In attacking this integral, treat n and $\bar{x}$ as given constants and μ as the only unknown variable or parameter. The exponent in the integrand of equation (1.38) is a quadratic form of μ, so

we "complete the square" and have

$$n(\bar{x}-\mu)^2+\mu^2=(n+1)\left(\mu-\tfrac{n\bar{x}}{n+1}\right)^2+\tfrac{n}{n+1}\bar{x}^2. \tag{1.39}$$

Thus

$$\begin{aligned}
M &= \tfrac{1}{\sqrt{2\pi/n}}\times\tfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[n(\bar{x}-\mu)^2+\mu^2\right]}\,d\mu \\
&= \tfrac{1}{\sqrt{2\pi/n}}\times\tfrac{1}{\sqrt{2\pi}}\times\left[\int_{-\infty}^{\infty} e^{-\frac{1}{2}(n+1)\left(\mu-\frac{n\bar{x}}{n+1}\right)^2}\times e^{-\frac{n}{2(n+1)}\bar{x}^2}\,d\mu\right] \quad \leftarrow \text{The red part is independent of } \mu \\
&= \tfrac{1}{\sqrt{2\pi/n}}\times\tfrac{1}{\sqrt{2\pi}}\times\sqrt{2\pi/(n+1)}\times e^{-\frac{n}{2(n+1)}\bar{x}^2}\times\left[\int_{-\infty}^{\infty}\tfrac{1}{\sqrt{2\pi/(n+1)}} e^{-\frac{1}{2}(n+1)\left(\mu-\frac{n\bar{x}}{n+1}\right)^2}\,d\mu\right] \quad \leftarrow \text{Integrand is normal density.} \\
&= \frac{1}{\sqrt{2\pi(\frac{1}{n}+1)}} e^{-\frac{1}{2(\frac{1}{n}+1)}\bar{x}^2} = \phi(\bar{x};0,\tfrac{1}{n}+1).
\end{aligned} \tag{1.40}$$

Many integrals in statistical applications cannot be calculated analytically. We mention two important ones here.

The first is the cumulative distribution function (CDF) of the standard normal distribution (see below for the definition of CDF)

$$\Phi(z)=\int_{-\infty}^{z}\tfrac{1}{\sqrt{2\pi}}e^{-t^2/2}dt. \tag{1.41}$$

The integrand $\phi(x)=\tfrac{1}{\sqrt{2\pi}}e^{-x^2/2}$ is the probability density function (PDF) of the standard normal distribution. The CDF is $\Phi(z) = \Pr\{Z < z\}$ where $Z \sim N(0, 1)$. Here are a few familiar values: $\Phi(1.96) = 0.975$, and $\Phi(2.58) = 0.995$.

The second integral is called the gamma function

$$\Gamma(\alpha)=\int_0^{\infty} x^{\alpha-1}e^{-x}dx. \tag{1.42}$$

Only the case of $\alpha > 0$ concerns us here. When α is a whole number,

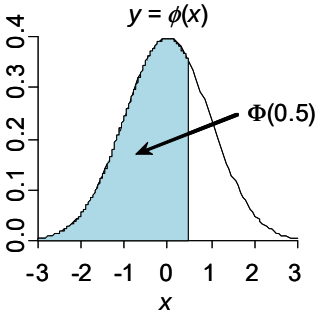
$$\Gamma(n)=(n-1)!, \tag{1.43}$$

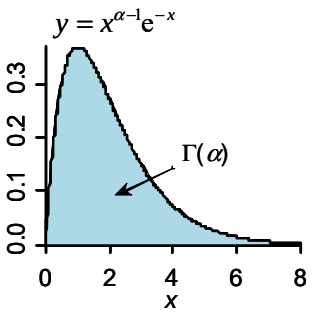
so the gamma function is a generalization of the factorial to non-integer numbers. When α is not an integer, numerical methods are used to calculate $\Gamma(\alpha)$.

$$\Gamma(\alpha+1)=\alpha\Gamma(\alpha),$$

$$\Gamma(\tfrac{1}{2})=\sqrt{\pi}.$$

Numerical algorithms exist for calculating $\Phi(z)$ and $\Gamma(\alpha)$.






```
#### Normal CDF function
a=-3; b=3; threshold=0.5;
curve(dnorm(x, 0,1), xlim=c(a,b),
  frame.plot=0);
x=seq(a,threshold,0.01);
y=dnorm(x,0,1);
polygon(c(a,x,threshold), c(0,y,0),
  col='lightblue');

#### Gamma function
gammaf <- function(x, a, b)  x^(a-1)*exp(-b*x)
a=2; b=1; range=8;
curve(gammaf(x, a, b), from=0, to=range,
  frame.plot=0);
x=seq(0,range,0.01);
y=gammaf(x,a,b);
polygon(c(0,x,range), c(0,y,0),
```

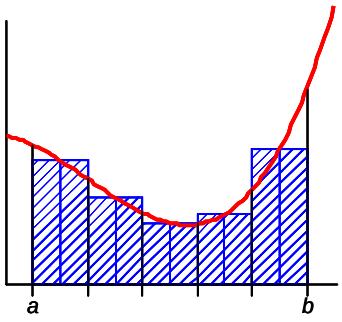
The CDF $\Phi(x)$ of the standard normal distribution and the $\Gamma(x)$ function are the shaded areas in the plots.

```
col='lightblue');

plot(x,y,  type='h', frame.plot=0,
  col='lightblue');
```

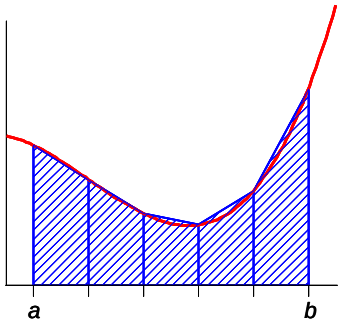
Numerical integration. We break the interval (a, b) into n pieces, each of width $h = (b - a)/n$. We approximate the area of each piece by the area of the rectangle, with the height given by the function value at the mid-point of the interval. Label the n mid-points as $x_1 = a + h/2$, $x_2 = a + 3h/2$, …, $x_n = a + (2n - 1)h/2$, and let $y_i = f(x_i)$, $i = 1, 2, \ldots, n$. Then

$$\int_a^b f(x)\,dx \approx \frac{(b-a)}{n}(y_1 + y_2 + \cdots + y_n) = \frac{b-a}{n}\sum_{i=1}^{n} f\left(a + (i - \tfrac{1}{2})\frac{(b-a)}{n}\right). \qquad (1.44)$$



Midpoint method

```
#### Numerical integration, midpoint method
fx <- function(x) 0.5*x^3  - 3*x^2 + x + 20
a0=0.5; b0=6.5; ymax=35;
a=1; b=6;
xmid = seq(a+0.5, b-0.5, 1);
ymid = fx(xmid);
plot(xmid, ymid, type='h', lwd=2, col="blue", xlim=c(a0,b0),
   ylim=c(0,ymax), xaxs='i', yaxs='i')
for(i in 1:5) {
   xb = c(a+i-1, a+i-1, a+i, a+i);
   mid = fx(a+i-0.5);
   yb = c(0, mid, mid, 0);
   polygon(xb, yb, density=15, col="blue");
}
curve(fx, from=a0, to=b0, lwd=3, col="red", add=T);
x = seq(a, b, 1);
y = fx(x);
lines(x,y, type='h');
```



Trapezoid method

```
#### Numerical integration, trapezoid method
fx <- function(x) 0.5*x^3  - 3*x^2 + x + 20
a0=0.5; b0=6.5; ymax=35;
a=1; b=6;
x = seq(a, b, 1);
y = fx(x);
plot(x, y, type="h", lwd=2, col="blue", xlim=c(a0,b0),
   ylim=c(0,ymax), xaxs='i', yaxs='i');
lines(x, y, lwd=2, col="blue")
curve(fx, from=a0, to=b0, lwd=3, col="red", add=T);
polygon(c(a,x,b),c(0,y,0), density=15, col="blue");
```

The trapezoid method. We break the interval (a, b) into n pieces, each of width $h = (b - a)/n$. We label the $(n + 1)$ points as $x_0 = a$, $x_1 = a + h$, $x_2 = a + 2h$, …, $x_n = b$, and let $y_i = f(x_i)$, $i = 0, 1, 2, \ldots, n$. Then we approximate the area of each piece by the area of the trapezoid, to get

$$\begin{aligned}\int_a^b f(x)\,dx &\approx \tfrac{h}{2}(y_0 + y_1) + \tfrac{h}{2}(y_1 + y_2) + \cdots \tfrac{h}{2}(y_{n-1} + y_n) \\ &= \tfrac{h}{2}(y_0 + 2y_1 + 2y_2 + 2y_{n-1} + y_n).\end{aligned} \qquad (1.45)$$

More sophisticated methods may use a curve (polynomial) to approximate the function $f(x)$. Note that a polynomial $p(x) = a_0 + a_1x + a_2x^2 + \ldots + a_mx^m = \sum_{i=0}^{m} a_i x^m$ is integrable, with $\int p(x)\,dx = \sum_{i=0}^{m} \frac{1}{m+1} a_i x^{m+1}$. If $f(x)$ is a straight line, $f(x) = a_0 + a_1x$, two points will be enough to achieve a perfect fit. In general, one can fit a polynomial of order n with $n + 1$ points. A class of numerical integration methods known as Gaussian quadrature try to approximate the integrand using those forms. Similarly if the integrand can be written in the forms $f(x) = p(x)e^{ax}$, $f(x) = p(x)\sin x$, etc., with $p(x)$ to be a polynomial, one can calculate the

integral analytically.

The integral $\int_a^b \int_c^d f(x,y)\,dy\,dx$ is the volume between the x-y plane and the surface $f(x, y)$ over the rectangle defined by $a < x < b$ and $c < y < d$. [[Add a graph to show volume.]]

Higher-dimensional integrals can be calculated numerically as well, just like the 1-D integrals. While the amount of computation for 1-D integrals is proportional to the number of points n, the amount of computation for integrals of k dimensions is proportional to n^k, which becomes quickly unmanageable with the increase of the dimension k. Calculation of high-dimensional integrals is a major problem in many applications in physics and statistics. This difficulty has motivated the development of modern computational algorithms such as Markov chain Monte Carlo.

1.6. Analytical and numerical methods

We use the example of finding the maximum of the following function to illustrate the difference between an analytical method and a numerical method (iterative algorithm).

$$L(\theta) = \theta^k (1-\theta)^{n-k} = \theta^{10}(1-\theta)^{90}, \quad 0 \le \theta \le 1,$$

with $n = 100$, $k = 10$. When θ changes, L will change as well. We want to find the value of θ that makes L achieve its maximum.

As will be explained later around equation 3.2, L is the likelihood function for binomial data while θ is the probability parameter, and the value of θ that maximizes L is called the maximum likelihood estimate (MLE). You can use the following R code to plot the curve.

```
L=function(x) x^10*(1-x)^90
curve(L(x), xlim=c(0,1))
```

Analytical solution. When L achieves its maximum, the slope of the curve is 0, so we set the first derivative to 0: $L'(\theta) = 0$ and solve the resulting equation. In our case this means

$$L' = k\theta^{k-1}(1-\theta)^{n-k} - \theta^k (n-k)(1-\theta)^{n-k-1} = \theta^{k-1}(1-\theta)^{n-k-1}[k(1-\theta) - (n-k)\theta] = 0. \quad (1.46)$$

We get three roots: $\theta = 0$, 1, and $k/n = 0.1$. The root $\theta = 0.1$ maximizes L and is the solution we seek. (In theory we should also check that the second derivative $L'' < 0$.)

Numerical solution. If we cannot calculate the derivatives or solve the equation, we may use a numerical algorithm to find the θ that maximizes L iteratively. One such algorithm is Newton's method, which uses a second-order polynomial (a parabola) to approximate the curve

$$\tilde{L} = a\theta^2 + b\theta + c. \quad (1.47)$$

If $a < 0$, $\tilde{L}$ has a maximum at $\theta^* = -b/(2a)$. We use the first three terms of the Taylor expansion to approximate $L(\theta)$ around the current value θ_k

$$\tilde{L}(\theta) = L(\theta_k) + L'(\theta_k)(\theta - \theta_k) + \tfrac{1}{2}L''(\theta_k)(\theta - \theta_k)^2. \quad (1.48)$$

This is a quadratic function in θ, in the form of equation (1.47), with $a = L''(\theta_k)/2$ and $b = L'(\theta_k) - L''(\theta_k)\theta_k$. If $L''(\theta_k) < 0$, the quadratic (1.48) achieves its maximum at

$$\theta_{k+1} = -\frac{b}{2a} = \theta_k - \frac{L'(\theta_k)}{L''(\theta_k)}. \quad (1.49)$$

In our problem, the first and second derivatives are given as

$$L' = k\theta^{k-1} \times (1-\theta)^{n-k} - \theta^k \times (n-k)(1-\theta)^{n-k-1},$$

$$L'' = k(k-1)\theta^{k-2} \times (1-\theta)^{n-k} - 2k(n-k)\theta^{k-1}(1-\theta)^{n-k-1} + (n-k))n-k-1)x^k(1-\theta)^{n-k-2}$$

We apply this algorithm to our problem, with the initial value $\theta_0 = 0.08$. The iteration is shown in the table below. The algorithm converges to the correct value very quickly, with L' approaching 0 very fast.

Two drawbacks of Newton's method are (i) that it requires the calculation of the first and second derivatives which may be expensive or impossible to calculate; and (ii) it may diverge so that good initial values are very important: in our example, the algorithm converges only if the initial value is in the neighbourhood of the MLE, in the narrow range (0.08, 0.11).

Iteration	θ_k	$L'(\theta_k)$	$L''(\theta_k)$	θ_{k+1}	$L(\theta_k)$
0	0.08	1.607×10^{-13}	-5.501×10^{-12}	0.1092	5.913×10^{-15}
1	0.1092	-6.894×10^{-14}	-6.282×10^{-12}	0.0982	7.285×10^{-15}
2	0.0982	1.518×10^{-14}	-8.692×10^{-12}	0.1000	7.604×10^{-15}
3	0.1000	1.849×10^{-16}	-8.467×10^{-12}	0.1000	7.618×10^{-15}
4	0.1000	3.579×10^{-20}	-8.464×10^{-12}	0.1000	7.618×10^{-15}
5	0.1000	1.414×10^{-27}	-8.464×10^{-12}	0.1000	7.618×10^{-15}

```
L <- function(x, n, k) (x^k)*((1-x)^(n-k))
dL <- function(x, n, k)  k*((1-x)^(n-k))*(x^(k-1)) - (n-k)*((1-x)^(n-k-1))*(x^k)
ddL <- function(x, n, k)  k*(k-1)*((1-x)^(n-k))*(x^(k-2)) - 2*k*(n-k)*((1-x)^(n-k-1))*(x^(k-
1)) + (n-k)*(n-k-1)*((1-x)^(n-k-2))*(x^k)

n = 100
k = 10
x = 0.08

for(i in 1:9){
  d = dL(x, n, k)
  dd = ddL(x, n, k)
  xnew = x - d/dd
  cat(i-1, sprintf("%.4f",x), sprintf("%.3e", d), sprintf("%.3e", dd),
      sprintf("%.4f",xnew), sprintf("%.3e", L(x, n, k)), sep = "   ", "\n")
  x = xnew
}
```

2. Probability

2.1. Two probability concepts

In *classical (also called Frequentist) statistics*, probability is defined as the limit of a frequency in a long-running experiment. For example, the frequency of heads in many tosses of a "fair" coin is believed to approach ½ when the number of coin tosses approaches infinity. The value ½ is thus the probability of heads for the coin.

In *Bayesian statistics*, probability is an expression of one's degree of belief. According to the Bayesian view (from the time of Laplace), the physical world is fully deterministic and the only uncertainty is our knowledge of it. Bayesian statistics attempts to use probability distributions to represent our uncertain knowledge of the world.

2.2. Discrete random variables

A random variable is a variable whose value is a measurement or observation of a random process. It may represent the possible outcome of an experiment to be performed, or the potential value of a quantity whose value is fixed but uncertain, due to incomplete information or imprecise measurement. For example, the random variable X may represent the result of a coin toss, with 0 for heads and 1 for tails, or X may represent the outcome of

the throw of a die: 1, 2, …, 6.

A random variable can be discrete or continuous. A discrete random variable may take a finite number of values or a countably infinite many values (such as all the natural numbers). A continuous random variable may assume any numerical value in an interval or a collection of intervals.

A probability distribution describes the probabilities of different values that a random variable may take. For a discrete random variable X, we list the probabilities for the individual values that X may take. Suppose X can take k possible values: x_i for $i = 1, 2, \ldots, k$, where k can be ∞. The distribution is then specified by the probabilities p_i for those values, with $\sum_i p_i = 1$.

x_i	x_1	x_2	…	x_k
p_i	p_1	p_2	…	p_k

For a continuous random variable, its distribution is characterized using the probability density function, to be explained later.

The average value of the random variable is known as the (mathematical) *expectation*. For a discrete variable X, it is defined as

$$\mu = E(X) = \sum_{i=1}^{k} p_i x_i \,. \tag{2.1}$$

The variance is defined as

$$V = E(x-\mu)^2 = \sum_{i=1}^{k} p_i (x_i - \mu)^2 = E(x^2) - \big(E(x)\big)^2 \,. \tag{2.2}$$

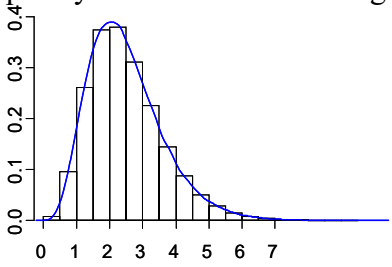
For example, the number of points in a throw of a fair dice has the following distribution

x_i	1	2	3	4	5	6
p_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The expectation is thus $E(x) = 1\times\frac{1}{6} + 2\times\frac{1}{6} + 3\times\frac{1}{6} + 4\times\frac{1}{6} + 5\times\frac{1}{6} + 6\times\frac{1}{6} = 3.5$, and the variance is $E(x) = \sum_{i=1}^{6} \frac{1}{6}(i-3.5)^2 = ?$.

2.3. Continuous random variables

Histogram. A histogram is an estimate of the probability distribution of a continuous random variable. We break the range of the variable into many equal-sized bins and plot the frequency of observations falling in each bin against the mid-value of the bin.



A histogram

```
#### histogram
x=rgamma(100000, 5, 2)
hist(x, xlim=c(0,6.9), ylim=c(0,0.45), freq=F, xaxs="i",
   yaxs="i", bty='l')
lines(density(x), col='blue', lwd=2)
```

Probability density function (PDF). Imagine now that we take a very large sample from a

continuous distribution, with billions of points, and construct a histogram with many small bins. A typical bin may be $(x, x + \Delta x)$. As Δx is very small, the mid value is nearly x. When the sample size increases and the bin size Δx decreases, the histogram will look more and more smooth. Furthermore we require that the total area under the curve is 1. This resulting curve is known as the probability density function or PDF and is written $f(x)$. The PDF has the following interpretation: $f(x)\Delta x$, for small Δx, is the proportion of the sample that lies in the bin $(x, x + \Delta x)$.

The (cumulative) distribution function (CDF), written as $F(x)$, is the proportion of points in all bins left of x. More formally, it is defined as

$$F(x) = \Pr(X < x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t. \tag{2.3}$$

Note that $F(-\infty) = 0$ and $F(\infty) = 1$.

For a continuous random variable X, its mean and variance are defined as

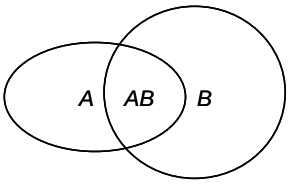
$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x\,. \tag{2.4}$$

$$V = E(x-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x\,. \tag{2.5}$$

2.4. Conditional, joint, and marginal probabilities

We consider the discrete case first, and then the continuous case.

Discrete variables



Set operations. $a \in A$ means an item a belongs to the set A. Ω is the whole set, $\varnothing$ is the empty set. $\cap$ means the intersection, $\cup$ means the union, and $\bar{}$ means negation.

Suppose A is the event that it will rain tomorrow, and B the event that it will not rain tomorrow. Then $A \cap B = \varnothing$ since both A and B cannot occur, while $A \cup B = \Omega$ since one of A and B must occur. Here $B = \bar{A}$.

Let A and B be two events. Then $P(A \cup B)$ is the probability that at least one of A and B occurs, and $P(A \cap B)$ is the probability that both A and B occur.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\,. \tag{2.6}$$

This should be obvious from the Venn diagram.

The *conditional probability* of A given B, written as $P(A|B)$ and read "probability of A given B", is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}\,, \tag{2.7}$$

under the assumption that $P(B) > 0$. Again for a proof, look at the Venn diagram.

For two discrete random variables X and Y, $p_{ij} = P(X = i, Y = j)$ is called the *joint probability*, while

$$\begin{aligned} P(X = x_i) &= \sum_{j=1}^{\infty} P(X = x_i, Y = y_j), \\ P(Y = y_i) &= \sum_{i=1}^{\infty} P(X = x_i, Y = y_j) \end{aligned} \tag{2.8}$$

are the *marginal probabilities* of X and Y, respectively.

Suppose $\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$ and $A_i \cap A_j = \varnothing$ for every i and j. In other words, one and only one of $A_1, A_2, \ldots, A_n$ must occur. Then $A_1, A_2, \ldots, A_n$ are said to form a mutually exclusive partition of the sample space. Then $\sum_{i=1}^{n} P(A_i) = 1$. The probability of any event B is then given as

$$P(B) = \sum_{i=1}^{n} P(A_i, B) = \sum_{i=1}^{n} P(A_i) P(B \mid A_i) . \qquad (2.9)$$

This is called *the law of total probability*. The conditional probability of A_i given that B has occurred is given by the *Bayes theorem* or the *inverse-probability theorem*:

$$P(A_i \mid B) = \frac{P(A_i, B)}{P(B)} = \frac{P(A_i) P(B \mid A_i)}{\sum_{j=1}^{n} P(A_j) P(B \mid A_j)}, \quad i = 1, 2, \ldots, n. \qquad (2.10)$$

Example (False positives of a test) (from Yang 2006: page 147). Suppose a new clinical test has been developed to screen for an infection in the population. If a person has the infection, the test accurately reports a positive 99% of the time, and if a person does not have the infection, the test falsely reports a positive only 2% of the time. Suppose that 0.1% of the population have the infection. What is the probability that a person who has tested positive actually has the infection?

Let A be the event that a person has the infection, and $\bar{A}$ no infection. Let B stand for test-positive. Then P(A) = 0.001, $P(\bar{A}) = 0.999$, $P(B|A) = 0.99$, $P(B \mid \bar{A}) = 0.02$. The probability that a random person from the population tests positive is, according to equation (2.10),

$$P(B) = P(A)P(B \mid A) + P(\bar{A})P(B \mid \bar{A}) = 0.001 \times 0.99 + 0.999 \times 0.02 = 0.02097.$$

This is close to the proportion among the noninfected individuals of the population. Equation (2.10) then gives the probability that a person who has tested positive has the infection as

$$P(A|B) = \frac{P(A)P(B \mid A)}{P(B)} = \frac{0.001 \times 0.99}{0.02097} = 0.0420.$$

This indicates a perhaps surprisingly poor test. Because the infection is rare and most individuals are healthy, most of the people testing positive are actually healthy.

Exercise 6 The Monty Hall problem is a probability puzzle based on the US television game show *Let's Make a Deal*, originally hosted by Monty Hall. It is also called the Monty Hall paradox. Suppose you are given the choice of three doors: Behind one door is a car; behind the other two, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Calculate the probability of winning if you do not switch and if you switch.

Hint: Define A_1: behind door 1 is a car; A_2: behind door 1 is a goat. Define B: winning. Then apply the law of total probability (eq. 2.9) for each of the two options (switching and no switching).

Continuous variables

For continuous random variables x and y, $f(x, y)$ is called the *joint probability density function* of (X, Y), while

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,\mathrm{d}u\,\mathrm{d}v \tag{2.11}$$

is the cumulative distribution function. The marginal density functions of X and Y are

$$\begin{aligned} f(x) &= \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y, \\ f(y) &= \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x. \end{aligned} \tag{2.12}$$

The conditional probability density function

$$f(x\,|\,y) = \frac{f(x,y)}{f(y)} = \frac{f(x)\,f(y\,|\,x)}{f(y)}. \tag{2.13}$$

The joint probability density is thus written as

$$f(x,y) = f(y)\,f(x\,|\,y) = f(x)\,f(y\,|\,x). \tag{2.14}$$

Distributions of more than two variables are defined similarly.

Similar to the discrete case, we also have the law of total probability and the Bayes theorem, except that the results are stated using the probability density functions (PDFs) instead of probabilities, and the summations are replaced by integrals.

Law of total probability: $$f(y) = \int_{-\infty}^{\infty} f(y\,|\,x)\,f(x)\,\mathrm{d}x. \tag{2.15}$$

Bayes theorem: $$f(x\,|\,y) = \frac{f(y\,|\,x)\,f(x)}{f(y)} = \frac{f(y\,|\,x)\,f(x)}{\int_{-\infty}^{\infty} f(y\,|\,x)\,f(x)\,\mathrm{d}x}. \tag{2.16}$$

[[Add examples to illustrate the use of conditioning to derive probabilities.]]

2.5. Common discrete distributions

Below we illustrate a few commonly used discrete distributions. For each we give the probability distribution as well as the mean (expectation) and variance.

Binomial distribution. Suppose a coin is biased with the probability of heads to be p. The number of heads x in n tosses of the coin has the binomial distribution

$$p_x = \binom{n}{x} p^x (1-p)^{n-x},\ \ x = 0,1,\cdots n, \tag{2.17}$$

where $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$. We write $x \sim \text{bino}(n, p)$. The expectation and variance are

$$E(x) = np,\ \ \text{Var}(x) = np(1-p). \tag{2.18}$$

Note that the proportion x/n has mean p and variance $p(1-p)/n$.

Poisson distribution. The Poisson distribution has a parameter $\lambda > 0$, which is the expected number of a particular event. The number of such events has the probability

$$p_x = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda},\ x = 0,1,\cdots. \tag{2.19}$$

Both the expectation and variance are $E(x) = \mathrm{Var}(x) = \lambda$.

Another common formulation is based on the *Poisson process*. Suppose a particular event (earthquake, lightening, arrival of a customer at McDonald's, a mutation in a gene etc.) occurs at the rate λ, so that the expected number of events over time t is λt. Then the number of events over time t is a random variable from the Poisson distribution with parameter λt:

$$p_x = \frac{(\lambda t)^x}{x!}\mathrm{e}^{-\lambda t},\ x = 0,1,\cdots. \tag{2.20}$$

Note that the number of events (x) is a random variable, and its mean is λt.

2.6. Common continuous distributions

We give the PDF of a few commonly-used continuous distributions, together with their expectation (mean) and variance.

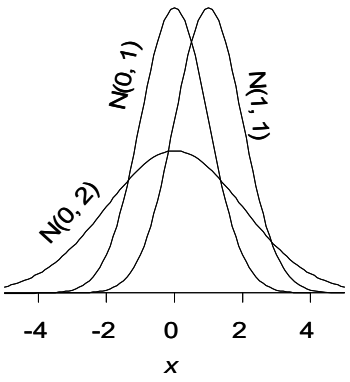
Normal distribution: $X \sim N(\mu, \sigma^2)$.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},\qquad -\infty < x < +\infty. \tag{2.21}$$

The expectation and variance are $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

If $\mu = 0$ and $\sigma^2 = 1$, $X \sim N(0, 1)$ is said to follow the standard normal distribution. The PDF and CDF for the standard normal distribution are often written as ϕ and Φ.

$$\begin{aligned}\phi(x) &= \tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2},\\ \Phi(x) &= \int_{-\infty}^{x}\tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-t^2/2}\mathrm{d}t.\end{aligned} \tag{2.22}$$



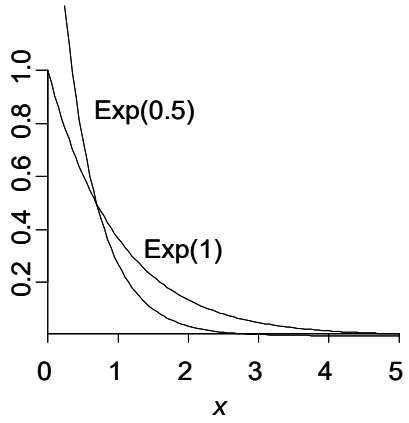
The probability density functions of normal distribution with different parameters.

```
#### Normal distributions
a=-5; b=5;
curve(dnorm(x, 0, 1), xlim=c(a,b), xaxs="i",
   yaxs="i", bty="l");
curve(dnorm(x, 0, 2), xlim=c(a,b), add=T);
curve(dnorm(x, 1, 1), xlim=c(a,b), add=T);
```

Exponential distribution. The density function is

$$f(x) = \lambda\mathrm{e}^{-\lambda x},\quad x \geq 0. \tag{2.23}$$

This is sometimes written as $X \sim \mathrm{Exp}(\lambda)$. The mean and variance are $E(X) = 1/\lambda$ and $\mathrm{Var}(X) = 1/\lambda^2$.




```
#### exponential distributions
a=0; b=5;
curve(dexp(x, 1), xlim=c(a,b), xaxs="i", yaxs="i",
   bty="l");
curve(dexp(x, 2), xlim=c(a,b), add=T);
```

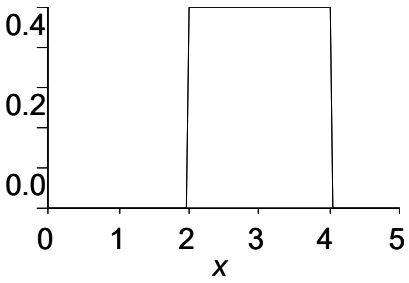
The probability density function of the exponential distribution

One reason that the exponential distribution is important is that the waiting time of a *Poisson process* is an exponential random variable. Suppose bus N arrives according to a Poisson process, with rate $\lambda = 0.1$ per minute. The probability that the bus will arrive in the next half minute (we think a half minute is a small time interval) is $\lambda\Delta t = 0.1 \times 0.5 = 0.05$. The number of bus arrivals over the time interval $(t_0, t_0 + t)$ has a Poisson distribution with mean λt: if $t = 60$ minutes, the average number of bus arrivals in one hour is 6. The waiting time or inter-arrival time has the exponential distribution with mean $1/\lambda = 10$ minutes. The Poisson process has no memory. The probability that the bus will arrive in the next 2 minutes is independent of the amount of time that we have waited for the bus; the waiting time until the bus arrives has the same exponential distribution with mean 10 minutes, whether we have waited for 1 minute or 30 minutes.

Uniform distribution: X ~ U(a, b). The density function is

$$f(x) = \frac{1}{b-a}, \quad a < x < b. \tag{2.24}$$

The mean and variance are $E(x) = (a + b)/2$ and $\text{Var}(x) = (b - a)^2/2$.




```
#### uniform distribution
a=0; b=5;
curve(dunif(x,2,4), xlim=c(a,b), type='l', xaxs="i",
   yaxs="i", bty="l");
```

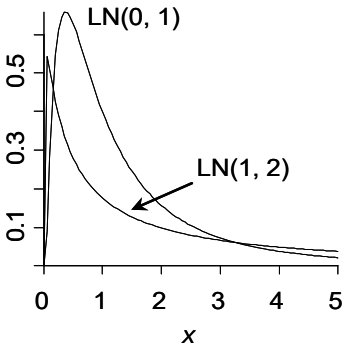
The probability density function for $U(a, b)$

The $U(0, 1)$ random variable is fundamental to computer simulation, and is known as a *random number*. A *random number generator* is a mathematical algorithm that generates a sequence of numbers that look like random variables from the $U(0, 1)$ distribution.

Log-normal distribution. If y is a random variable with a normal distribution, $Y \sim N(\mu, \sigma^2)$, then $X = e^Y$ has a log-normal distribution. The density function is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(\log\{x\} - \mu)^2}{2\sigma^2} \right\}, \quad x > 0. \tag{2.25}$$

The mean and variance are $E(X) = e^{\mu + \frac{1}{2}\sigma^2}$ and $V(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$. Note that while the mean of Y is μ, the mean of $X = e^Y$ is not e^{μ}.



The probability density function of the log-normal distribution

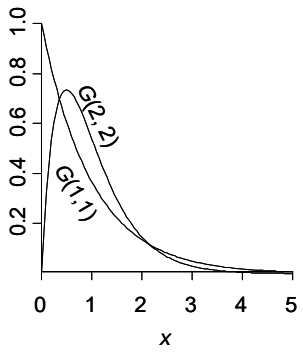
```
#### Log-normal distribution
a=0; b=5;
curve(dlnorm(x,0,1), xlim=c(a,b), type='l',
   xaxs="i", yaxs="i", bty="l");
curve(dlnorm(x,1,2), xlim=c(a,b), add=T);
```

Gamma distribution. $X \sim G(\alpha, \beta)$. The density function is

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0, \tag{2.26}$$

where $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$ is the gamma function (equation 1.42). The mean and variance are $E(X) = \alpha/\beta$ and $\mathrm{Var}(X) = \alpha/\beta^2$. Here α is the shape parameter and β is the rate parameter. When $\alpha \le 1$, the density has a L shape. If $\alpha > 1$, it has a mode in the middle. When α is large, the gamma distribution is approximated by the normal distribution.

Some authors use an alternative notation in which the scale parameter (also written as β) instead of the rate parameter is used: note that the scale parameter is the reciprocal of the rate parameter. Check the mean to make sure which notation is used.



The probability density function for the gamma distribution $X \sim G(\alpha, \beta)$

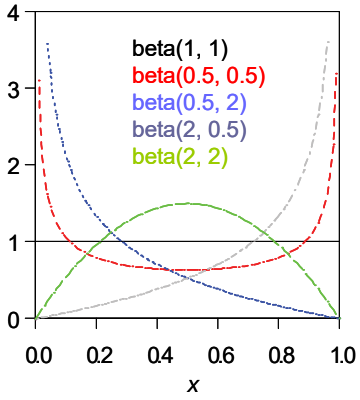
```
#### Gamma distribution
a=0; b=5;
curve(dgamma(x,1,1), xlim=c(a,b), type='l',
   xaxs="i", yaxs="i", bty="l");
curve(dgamma(x,2,2), xlim=c(a,b), add=T);
```

When $\alpha = 1$, the gamma distribution reduces to the exponential distribution: $G(1, \beta) = \mathrm{Exp}(\beta)$. Also when $\alpha = n/2$ and $\beta = ½$, the gamma distribution is known as the χ^2 distribution with n degrees of freedom, written as χ^2_n: that is, $G\left(\frac{n}{2}, \frac{1}{2}\right) = \chi^2_n$.

Beta distribution. $X \sim \mathrm{beta}(a, b)$. The density function is

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1 \tag{2.27}$$

In particular, beta(1, 1) is the $U(0, 1)$ distribution.
The mean and variance are $E(X) = a/(a + b)$ and $V(X) = ab/[(a + b)^2 (a + b + 1)]$.



The probability density function for the beta distribution $x \sim$ beta(a, b).

```
#### Beta distributions
a=0; b=1;
curve(dbeta(x,1,1), xlim=c(a,b), ylim=c(0,4),
   type='l', xaxs="i", yaxs="i");
curve(dbeta(x,0.5,0.5), xlim=c(a,b), col='red',
   lty=2, lwd=2, add=T);
curve(dbeta(x,0.5,2), xlim=c(a+0.04,b), col='blue',
   lty=3, lwd=2, add=T);
curve(dbeta(x,2,0.5), xlim=c(a,b-0.04), col='gray',
   lty=4, lwd=2, add=T);
curve(dbeta(x,2,2), xlim=c(a,b), col='green', lty=5,
   lwd=2, add=T);
```

2.7. Functions of random variables

Suppose the random variable X has the probability density function $f_X(x)$ and $y = y(x)$ is a one-to-one transform of x, with the inverse transform $x = x(y)$. Then Y is a random variable with density

$$f_Y(y) = f_X(x(y)) \times \left| \frac{dx}{dy} \right|. \tag{2.28}$$

Here $f_Y(y)$ is the PDF of Y evaluated at $Y = y$ and $f_X(x)$ is the PDF of X evaluated at $X = x$.

To appreciate how this formula works, imagine we have a huge number of x values and a histogram for x. We then apply the transform $y = y(x)$ to calculate the corresponding y values and construct a histogram for y. In the histogram for x, the proportion of points in the bin $(x, x + \Delta x)$ is $f_X(x)\Delta x$. The transform converts x into $y = y(x)$ and $x + \Delta x$ into $y + \Delta y = y(x + \Delta x)$, with $\Delta y = \Delta x \cdot \frac{dy}{dx}$. In other words, after the transform all points in the $(x, x + \Delta x)$ bin, which has width Δx, will be in the $(y, y + \Delta y)$ bin, which has width Δy. Since those points constitute the same proportion of the whole sample, we have

$$f_X(x)\Delta x = f_Y(y)\Delta y, \tag{2.29}$$

which means $f_Y(y) = f_X(x)\Delta x/\Delta y = f_X(x)dx/dy$. If the transform $y = y(x)$ is monotonically decreasing, we have $f_Y(y) = f_X(x)|dx/dy|$ since only the width of the bin matters in the argument.

Example (Normal distribution). Suppose z has a standard normal distribution with density

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}. \tag{2.30}$$

Let $x = \mu + \sigma z$, so that $z = (x - \mu)/\sigma$ and $dz/dx = 1/\sigma$. Then x has the density

$$f(x) = \phi\left(\frac{x-\mu}{\sigma}\right) \times \frac{1}{\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}. \tag{2.31}$$

Thus x has a normal distribution with mean μ and variance σ^2.

Example (Log-normal distribution). Suppose $y \sim N(\mu, \sigma^2)$, and $x = e^y$.

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < +\infty \tag{2.32}$$

We have $y = \log(x)$ and $dy/dx = 1/x$. Thus the density function of x is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\log\{x\} - \mu)^2}{2\sigma^2}\right\} \times \frac{1}{x} = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\log\{x\} - \mu)^2}{2\sigma^2}\right\}, \tag{2.33}$$

which is equation (2.25).

Example (Exponential distribution). Suppose $u \sim U(0, 1)$, and $x = -\log(u)$. Then x has an exponential distribution with mean 1. To see this, note that $f(u) = 1$ for $0 < u < 1$, and $du/dx = -e^{-x}$, so that

$$f(x) = 1 \times e^{-x}, \quad 0 < x < \infty. \tag{2.34}$$

Thus if u is a random number, which follows the $U(0, 1)$ distribution, then $x = -\log(u)$ will have an exponential distribution with mean 1. Thus to simulate an exponential random variable with parameter λ (and mean $1/\lambda$), we first generate a random number u, and then apply the transform $x = -\log(u)/\lambda$.

Exercise 7 (inverse gamma distribution) Suppose

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \; 0 < x < \infty.$$

Let $y = 1/x$. Derive the density of y.

Exercise 8 Suppose

$$f(x) = \tfrac{1}{\mu} e^{-x/\mu}, \; 0 < x < \infty.$$

Let $y = 1 - e^{-x/\mu}$. Show that y has the uniform distribution by deriving the density of y. [Hint. $F(x) = 1 - e^{-x/\mu}$ is the C.D.F. of x. First determine the range of y.]

3. Statistics

Statistics is a science that aims to make inference about the population based on a sample. The sample is assumed to involve random errors, generated under a probabilistic model. For example, we may assume that our data are a sample taken from a normal distribution and we are interested in estimating the mean μ and variance σ^2 of the normal distribution or in testing the hypothesis that $\mu = 0$.

3.1. Estimation

A *parameter* is a constant that describes the population we are interested in, for example, the mean μ and variance σ^2 of the normal distribution. A *statistic* is a quantity we can calculate using the observed data. For example the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a statistic. We use a statistic to estimate the parameter, and it is called an estimator. For example, the sample mean $\bar{x}$ is an estimator of parameter μ, sometimes written as $\hat{\mu} = \bar{x}$.

The statistic or estimator is a random variable. It fluctuates among datasets if we take many different samples (datasets) from the population. If the average of the estimator over datasets equals the parameter value,

$$E(\hat{\mu}) = \mu, \tag{3.1}$$

the estimator is said to be *unbiased*. Otherwise it is said to be biased and the difference $E(\hat{\mu}) - \mu$ is called the *bias*.

Note that unbiased estimators are in general not invariant to different parametrizations, except that the transform $h(\theta)$ is linear. In other words, if $E(\hat{\theta}) = \theta$, then $E(a + b\hat{\theta}) = a + b\theta$. However, if $E(\hat{\theta}) = \theta$, then $E(\hat{\theta}^2) \neq \theta^2$.

Likelihood function and MLE

The probability of observing the data given the parameters, viewed as a function of the parameters, is called the likelihood function. Values of parameters that maximize the likelihood function are called maximum likelihood estimates (MLEs)

Maximum likelihood (ML) is a methodology for estimating parameters and testing statistical hypotheses. Suppose we have an unknown parameter θ in the model. The probability of observing the data D is considered a function of θ, and is called the *likelihood function*. According to the *likelihood principle*, the likelihood function contains all information from the data about θ. We use two simple examples to introduce the methodology.

Example. Binomial model (red fish and blue fish). There are a lot of red and blue fish in a pond. Suppose we take a sample of $n = 100$ fish and found $k = 10$ red and $n - k = 90$ blue. What is our best estimate of the proportion of red fish (p)?

With this simple case, we know the answer: the estimate is $k/n = 0.1$. To use ML, note

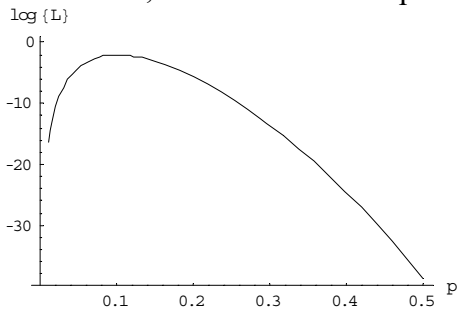
$$\Pr(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{100}{10} p^{10} (1-p)^{90}. \tag{3.2}$$

Since the data are observed, we view this probability as a function of p; let it be $L(p)$. L is usually very small, so it is more convenient to work with the log likelihood $\ell = \log\{L\}$.

$$\ell(p) = \log\{L(p)\} = \log\binom{100}{10} + 10\log(p) + 90\log(1-p).$$

This is plotted against p below. $\ell(p)$ and $L(p)$ reach their maxima when $p = k/n$, so $\hat{p} = k/n = 0.1$ is the MLE.

The constant $\binom{100}{10}$ is often ignored. Note that dropping it does not affect our evaluation of different values of p. The likelihood function is then $p^{10}(1-p)^{90}$, given by multiplying 10 ps (corresponding to the 10 red fish, each of which has probability p of occurrence) and 90 $(1-p)$'s (corresponding to the 90 blue fish, each of which has probability $1-p$ of occurrence).



Example. Estimation of sequence distance under JC69. Suppose we want to estimate the sequence distance d under the JC69 model using a pair of sequences with $n = 100$ sites and $k = 10$ differences. The unknown parameter is d. Consider any site in the sequence. The probability that this site is different between the two sequences separated by distance d is

$$p = \frac{3}{4}\left(1 - e^{-\frac{4}{3}d}\right). \tag{3.3}$$

The likelihood function $L(d)$ or the probability of observing k differences out of n sites is given by the binomial probability

$$L(d) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{100}{10}\left(\frac{3}{4} - \frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{10}\left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right)^{90}. \tag{3.4}$$

The log likelihood function is

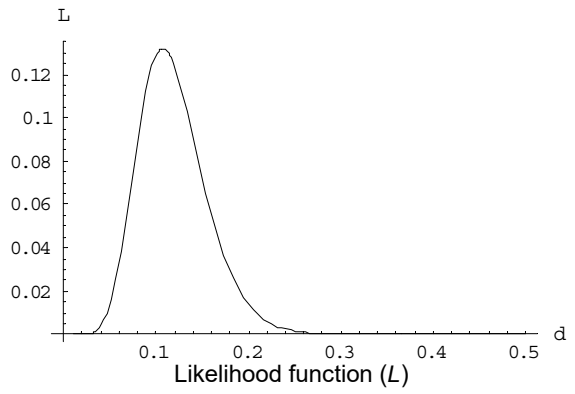
$$\ell(d) = \log\{L(d)\} = \log\binom{100}{10} + 10\log\left(\frac{3}{4} - \frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right) + 90\log\left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-\frac{4}{3}d}\right). \tag{3.5}$$

These are plotted below. They achieve their highest values at the same parameter value $\hat{d} = 0.1073$, corresponding to $\hat{p} = k/n = 10/100$. Thus the MLE of d is given by the Jukes-Cantor formula.

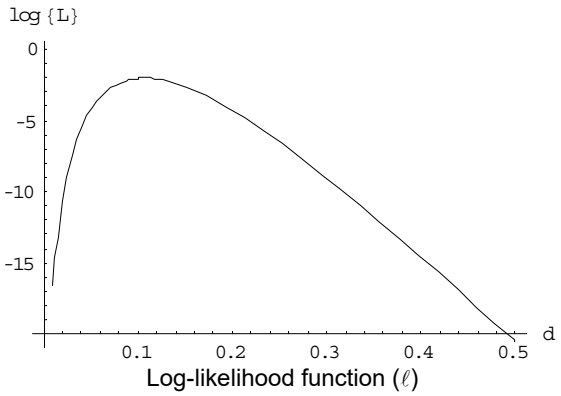
$$\hat{d} = -\tfrac{3}{4}\log(1 - \tfrac{4}{3}\hat{p}), \tag{3.6}$$

where the base of the logarithm is the constant e. You can also obtain this analytically by solving the *likelihood equation*

$$\frac{\mathrm{d}\ell}{\mathrm{d}d} = 0. \tag{3.7}$$



Likelihood function (L)



Log-likelihood function (ℓ)

Again note that after the constant is dropped, the likelihood is given by multiplying the probabilities across all sites: 10 differences and 90 identities.

The invariance of the MLEs. Note that if the MLE of parameter θ is $\hat{\theta}$, and $h(\theta)$ is a monotonic function, then the MLE of $h(\theta)$ is $h(\hat{\theta})$. For example, if θ is the side of a square and its MLE is 10 m. Then the MLE of the area of the square $h(\theta) = \theta^2$ is simply 100 m^2. Suppose we are building a model, in which we can measure the size of the square by either its side or its area. These two options will be different *parametrizations*. ML is *invariant* to different parametrizations: no matter what parameters you choose to use, you will obtain the same conclusions. In this regard, one may view the proportion of different sites p as a different parametrization in the Jukes-Cantor model, and its MLE is $\hat{p} = k/n$. Then $d = h(p)$ is considered a one-to-one mapping from p to d through equation (3.3), which can be used to derive the MLE of d as

$$\hat{d} = h(\hat{p}) = -\tfrac{3}{4}\log(1 - \tfrac{4}{3}\hat{p}). \qquad (3.8)$$

Thus in general MLEs may have a bias in small datasets, which is tolerated. When the dataset is large, MLEs are asymptotically unbiased. They are also asymptotically normally distributed and asymptotically most efficient (they have the smallest variance).

Under these simple models, analytical solutions are available, and the MLEs agree with our intuition. In more complicated models, our intuition may fail. Then ML provides a general methodology to allow us to proceed. When analytical solutions are not possible, numerical optimisation algorithms are used to obtain the MLEs.

Example. Estimation of allele frequencies for the ABO blood groups. The A, B, and O blood groups were discovered in 1900 and 1901 at the University of Vienna by Karl Landsteiner in the process of trying to learn why blood transfusions sometimes caused death and at other times saved a patient. In 1930, he received the Nobel Prize for this discovery. Two alleles A and B code for two antigens that bind with two antibodies. We will use the following table to estimate the frequencies of the two alleles p and q, with $r = 1 - p - q$.

Phenotypes	Genotypes	Probability	Sample	Frequency
A	AA + AO	$p^2 + 2pr$	$n_A = 44$	0.269939
B	BB + BO	$q^2 + 2qr$	$n_B = 27$	0.165644
AB	AB	$2pq$	$n_{AB} = 4$	0.024540
O	OO	r^2	$n_O = 88$	0.539877
Sum			$n = 163$	1

Here we have two parameters p and q in the model. Write down the likelihood, that is, the probability of observing the counts of people with different blood groups as a function of p and q. In this case it is not possible to obtain the MLEs analytically, and numerical methods

have to be used. The solution is $\hat{p} = 0.1605$, $\hat{q} = 0.1004$, $\hat{r} = 1 - \hat{p} - \hat{q} = 0.7392$, with $\ell = -175.448$. [[Insert some R code here for the optimization.]]

3.2. Hypothesis testing

Consider a sample from the normal distribution. Suppose we know the variance from past data and are interested in whether the population mean μ deviates from a specific value $\mu_0 = 0$. Here $\mu_0 = 0$ may represent a lack of treatment effect. Thus H_0: $\mu_0 = 0$ is the null hypothesis. If the observed data are seriously at odds with the predictions of H_0, we will be forced to reject H_0 and accept the more general hypothesis: H_1: $\mu > \mu_0$. Here H_1 is the alternative hypothesis. Note that the two hypotheses H_0 and H_1 do not have the same role. If they both fit the data nearly equally well, we will prefer H_0. Only when H_0 does not fit the data and H_1 fits the data much better are we prepared to reject H_0. The error of rejecting H_0 when it is true is called the *type-I error* or *false positive error*. The error of accepting H_0 when H_0 is false (and H_1 is true) is the *type-II error* or *false negative error*. Type-I errors are considered to be more serious than type-II errors.

We use $\bar{x}$ as the test statistic. Suppose the data indicates a positive effect, with $\bar{x} > 0$. To determine whether this result could have arisen by chance even if H_0 is true, we derive the distribution of the test statistic under H_0 and calculate the probability that the test statistic is even greater than the observed statistic (the sample mean). This will be the p value. A small p value (say, <5% or 1%) means rejection of H_0. The p value may be the most confusing concept in statistics. All the following statements are incorrect:

(a) The p value is the probability that the null hypothesis is correct.
(b) The p value is the probability of the data given the null hypothesis.
(c) …

Likelihood ratio test. We are interested in comparing two parametric models H_0 and H_1, with H_0 to be a special case of H_1. The two models are said to be nested. Suppose the null model H_0 has q parameters and its optimized log likelihood is ℓ_0, and the alternative model H_1 has p parameters and its optimized log likelihood is ℓ_1. Then twice the log likelihood difference $2\Delta\ell = 2(\ell_1 - \ell_0)$ can be compared with the χ^2 distribution with $p - q$ degrees of freedom to decide whether H_1 fits the data significantly better than H_0. The null distribution here is reliable when the dataset is large. This test is known as the *likelihood ratio test*, as $2\Delta\ell = 2(\ell_1 - \ell_0) = 2\log\frac{L_1}{L_0}$.

4. Markov chains

4.1. Discrete-time Markov chains

Suppose there are three kinds of weather: sunny (S ☼), cloudy (C) and raining (R). We write the probability for tomorrow's weather given today's weather in the form of a matrix, as follows

$$P = \{p_{ij}\} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}, \tag{4.1}$$

where p_{ij} is the probability that tomorrow's weather will be j given that today's is i. For example, $p_{SS} = 0.7$ means that given it is sunny today, the probability that it will be sunny tomorrow is 0.7. Note that each row sums to 1. There is a tendency for tomorrow's weather to stay the same as today's.

Let X_n be the weather on the nth day, which takes values of S, C or R. We assume that given today's weather, the probabilities for tomorrow's weather do not depend on the weather yesterday or earlier. In words, given the present, the future does not depend on the past. This memory-less property is known as *the Markovian property*. More formally

$$P(X_{n+1} \mid X_0, X_1, \ldots, X_n) = P(X_{n+1} \mid X_n). \tag{4.2}$$

Then $X_0, X_1, X_2, \ldots$ form a Markov chain.

Back to the weather, we assume no seasons, so that the transition probability matrix P is independent of time; that is

$$P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i) = p_{ij}. \tag{4.3}$$

Markov chains satisfying this condition are said to be *time-homogeneous*.

The dynamics of a time-homogeneous Markov chain is characterized by the (one-step) transition matrix P. We now try to calculate the two-step transition probability, for example, the probability that it will be sunny the day after tomorrow given that it is sunny today. We have to average (sum) over all possible states for tomorrow's weather

$$p_{SS}^{(2)} = p_{SS}\,p_{SS} + p_{SC}\,p_{CS} + p_{SR}\,p_{RS} = 0.7 \times 0.7 + 0.2 \times 0.1 + 0.1 \times 0.3 = 0.54. \tag{4.4}$$

Note that this is a straightforward application of the law of total probabilities (2.9), conditioning on tomorrow's weather.

Other two-step transition probabilities can be calculated similarly. If we let $P^{(2)} = \{p_{ij}^{(2)}\}$ be the two-step transition matrix. Equation (4.4) implies

$$P^{(2)} = P \cdot P = P^2. \tag{4.5}$$

By induction, the n-step transition probability matrix $P^{(n)} = \{p_{ij}^{(n)}\}$ is

$$P^{(n)} = P^{(n-1)} \cdot P = P^n. \tag{4.6}$$

A generalization of equation (4.4) is the *Chapman-Kolmogorov equation*:

$$p_{ij}^{(n+m)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}^{(m)} \quad \text{for any } i, j, \text{ and for any } m, n \geq 0. \tag{4.7}$$

This can also be written in matrix form as

$$P^{(m+n)} = P^{(m)} \cdot P^{(n)} \tag{4.8}$$

for any $m, n \geq 0$.

Exercise 9. Calculate P^2 using equation (4.1).

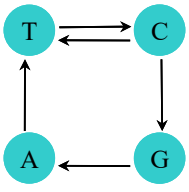
The n-step transition matrix P^n can be calculated through the diagonalization of P (equation 1.14).

Exercise 10 (Jukes & Cantor model of DNA sequence evolution).* The evolution of a nucleotide site in a DNA sequence is described by a Markov chain. The four states are the nucleotides T, C, A, G. In every generation the nucleotide changes to one of the three other nucleotides with probability λ. The transition matrix is thus

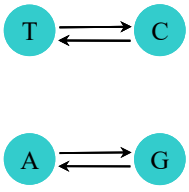
$$P = \begin{bmatrix} 1-3\lambda & \lambda & \lambda & \lambda \\ \lambda & 1-3\lambda & \lambda & \lambda \\ \lambda & \lambda & 1-3\lambda & \lambda \\ \lambda & \lambda & \lambda & 1-3\lambda \end{bmatrix}. \tag{4.9}$$

The nucleotides are ordered T, C, A, and G. Calculate P^n. To be specific, consider the evolution of a site in a DNA sequence in the human-chimpanzee ancestor down to the modern human. Let $\lambda = \frac{1}{3}\times 10^{-8}$ per generation, with 500,000 generations from the common ancestor to the present (assuming 10 years in one generation).

Certain Markov chains may have states with $P_{ii} = 1$, so that the chain will remain in those states as soon as it enters them. Such states are called *absorbing* states. Some population genetics models involve absorbing states, but they are not often used in models in molecular evolution. Some Markov chains are *periodic*. For example, a Markov chain with three states 1, 2, 3, and with transitions $1 \to 2 \to 3 \to 1$ has the period 3. If we start in state 1, it is possible to get back to state 1 in 3 steps, 6 steps etc., but not in 5 steps. Please convince yourself that the Markov chain illustrated below has a period of 2. In some Markov chains, it is impossible (with probability 0) to go from some states to some other states. Those states are said to be *non-commutative*. For example, in the Markov chain represented by the graph below, the states T and C are *commutative*, as are A and G, but T and A are not commutative. We do not consider those types of chains. Instead we try to avoid them when we build models in molecular evolution.



A Markov chain with period 2



A Markov chain with non-commutative states

Markov chains that we will use have a finite number of states. All states are "connected" or *commutative*, and every state is visited an infinite number of times and is said to be *recurrent*. The chain is said to be *irreducible*. Furthermore, the chain does not have a period and is *aperiodic*. In Markov chain Monte Carlo (MCMC) algorithms in Bayesian computation, we should make sure that the Markov chain we generate is aperiodic and

irreducible.

Limiting and stationary distributions. When $n \to \infty$, all elements in the same column of the matrix P^n will be identical: that is, $p_{ij}^{(n)} \to \pi_j$ when $n \to \infty$, irrespective of the starting state i. The vector $\{\pi_1, \pi_2, \ldots, \pi_k\}$, where k is the number of states, is called the *limiting distribution*. When the number of transitions n is large, the chain will have lost the memory of the initial state so that the probability that the chain is in state j after n transitions is close to π_j, independent of the initial state i.

The row vector $\pi = \{\pi_1, \pi_2, \ldots, \pi_k\}$ is called the *stationary distribution* if the π_j's are ≥ 0 and sum to 1, and if they satisfy

$$\pi_j = \sum_i \pi_i p_{ij} \tag{4.10}$$

or in matrix notation

$$\pi = \pi P. \tag{4.11}$$

Exercise 11: Calculate the stationary distribution of the Markov chain specified by equation (4.1). Use equation (4.11) to form two linear equations. Use them together with $\pi_1 + \pi_2 + \pi_3 = 1$ to solve the three unknowns π_1, π_2, π_3. The stationary distribution gives us the proportions of sunny, cloudy and raining days.

4.2. Continuous-time Markov chains

There are several ways of characterizing the continuous-time Markov chain. First we can view it as the limit or approximation of a discrete-time Markov chain. We consider a small time interval Δt as one step in the discrete-time chain, and let $\Delta t \to 0$. The state at any time t is then represented by $X(t)$.

With the Jukes-Cantor mutation model, we can write the one-step transition matrix as

$$P(\Delta t) = \begin{bmatrix} 1-3\lambda\Delta t & \lambda\Delta t & \lambda\Delta t & \lambda\Delta t \\ \lambda\Delta t & 1-3\lambda\Delta t & \lambda\Delta t & \lambda\Delta t \\ \lambda\Delta t & \lambda\Delta t & 1-3\lambda\Delta t & \lambda\Delta t \\ \lambda\Delta t & \lambda\Delta t & \lambda\Delta t & 1-3\lambda\Delta t \end{bmatrix} = I + Q\Delta t\,, \tag{4.12}$$

where I is the identity matrix and

$$Q = \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix}. \tag{4.13}$$

We are interested in the transition probability matrix over time t, $P(t) = \{p_{ij}(t)\}$, where $p_{ij}(t)$ is the probability that given the chain is in state i now, it will be in state j time t later. We calculate $P(t)$ as the transition probability matrix over $n = t/\Delta t$ steps.

$$P(t) = [P(\Delta t)]^n = (I + Qt/n)^n \approx e^{Qt}. \tag{4.14}$$

The last approximation should look familiar if you remember $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e$ and

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x \text{ for a scalar } x. \tag{4.15}$$

For most models of nucleotide substitution, the matrix exponential e^{Qt} can be calculated by diagonalizing the matrix Q.

$Q = \{q_{ij}\}$ in equation (4.13) is known as the generator for the continuous-time Markov chain. In the molecular evolution literature, it is commonly known as the rate matrix, since q_{ij}, $i \neq j$, is the rate at which nucleotide i mutates (changes) into nucleotide j. More precisely, the probability that given the current state i, the chain will be in state j a small time interval Δt later is $q_{ij}\Delta t$: $\Pr(X(t + \Delta t) = j | X(t) = i) = q_{ij}\Delta t$. The Q matrix has the following properties: (1) the off-diagonal elements are nonnegative, and (2) each row sums to 0, so that $-q_{ii}$ is the total rate of change for state i. Sometimes we write $q_i = -q_{ii}$.

Here is the Q matrix for the so-called HKY or HKY85 model

$$Q = \begin{bmatrix} -(\alpha\pi_C + \beta\pi_R) & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & -(\alpha\pi_T + \beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha\pi_G + \beta\pi_Y) & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & -(\alpha\pi_A + \beta\pi_Y) \end{bmatrix}, \qquad (4.16)$$

where $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. Again the nucleotides are ordered T, C, A, and G. And here is the most general Q matrix for 4 states.

$$Q = \begin{bmatrix} -q_T & q_{TC} & q_{TA} & q_{TG} \\ q_{CT} & -q_C & q_{CA} & q_{CG} \\ q_{AT} & q_{AC} & -q_A & q_{AG} \\ q_{GT} & q_{GC} & q_{GA} & -q_G \end{bmatrix}, \qquad (4.17)$$

where the diagonals are given by the requirement that each row sums to 0.

A second characterization of the Markov chain views the continuous-time Markov chain as a "waiting game". Given the current state i, the waiting time until the next event (transition) is an exponential variable with rate parameter $q_i = -q_{ii}$ or with mean $1/q_i$. When a transition occurs, the chain moves to the alternative states with probabilities proportional to their rates. In other words, given that a transition occurs, the moves are described by a discrete-time Markov chain with transition matrix (still using the transitions between the nucleotides as an example)

$$M = \begin{bmatrix} 0 & \frac{q_{TC}}{q_T} & \frac{q_{TA}}{q_T} & \frac{q_{TG}}{q_T} \\ \frac{q_{CT}}{q_C} & 0 & \frac{q_{CA}}{q_C} & \frac{q_{CG}}{q_C} \\ \frac{q_{AT}}{q_A} & \frac{q_{AC}}{q_A} & 0 & \frac{q_{AG}}{q_A} \\ \frac{q_{GT}}{q_G} & \frac{q_{GC}}{q_G} & \frac{q_{GA}}{q_G} & 0 \end{bmatrix}. \qquad (4.18)$$

For the Jukes-Cantor model, all the off-diagonal elements of M are 1/3.

In other words, if we ignore the waiting times between transitions, the sequence of states visited by the process constitutes a discrete-time Markov chain. This is called the *jump chain* or the *embedded Markov chain*.

The limiting and stationary distributions of the continuous-time Markov chain are defined similarly to the discrete case. To get the stationary distribution, we can solve the equation

$$\pi Q = 0, \qquad (4.19)$$

together with the constraint that the sum of the πs should be 1.

Glossary

English	Deutsch			日本語	中文
alternative hypothesis	Alternative Hypothese / Alternativhypothese	Hipótesis alternativa	Εναλλακτική υπόθεση	対立仮説	备择假设
aperiodic chain	Aperiodische/unperiodische Kette	Cadena aperiódica	Μη περιοδική αλυσίδα	非周期的連鎖	非周期链
Bayes's theorem	Satz von Bayes/Bayes-Regel	Teorema de Bayes	Μπεϋσιανό θεώρημα	ベイズの定理	贝叶斯(贝斯)定理
chain rule	Kettenregel	Regla de la cadena	Κανόνας αλυσίδας	連鎖則	链式法则
commutative states	Kommutative Zustände	Estados conmutativos		可換な状態	互通
conditional distribution	Bedingte Verteilung	Distribución condicional	Υπο όρους (δεσμευμένη κατανομή)	条件つき分布	条件分布
cumulative distribution function (CDF)	Kumulative Verteilungsfunktion	Función de distribución acumulada	αθροιστική συνάρτηση κατανομής	累積分布関数	(累计)分布函数
definite integral	Bestimmtes Integral	Integral definida	Ορισμένο ολοκλήρωμα	定積分	定积分
derivative	Ableitung	Derivada	Παράγωγος	導関数	导数
determinant	Determinante	Determinante	Ορίζουσα	行列式	行列式
diagonal matrix	Diagonalmatrix	Matriz diagonal	Διαγώνιος πίνακας	対角行列	对角矩阵
diagonalization	Diagonalisierung	Diagonalización	Διαγωνιοποίηση	対角化	(矩阵)对角化
embedded chain	Eingebettete Kette	Cadena embebida	Ενσωματωμένη αλυσίδα	埋め込まれた連鎖	嵌入链 (跳跃链)
ergodic chain	ergodische (Markov)-Kette	Cadena ergódica (regular)	Εργοδική αλυσίδα	エルゴード的マルコフ連鎖	遍历链
generator	Generator	Generador	γεννήτρια	生成作用素	生成元
gradient (slope)	Gradient/Steigung	Gradiente	κλίση	勾配	梯度(斜率)
Hessian matrix	Die Hesse-Matrix	Matriz Hessiana	Εσσιανό Μητρώο ή Εσσιανός Πίνακας	ヘシアン行列	Hessian 矩阵
identity matrix	Identitätsmatrix/Einheitsmatrix	Matriz identidad	Πίνακας ταυτότητας (Μοναδιαίος Πίνακας)	単位行列	单位矩阵

initial distribution	Anfangsverteilung	Distribución inicial	Αρχική κατανομή	初期分布	初始分布
integral	Integral	Integral	ολοκλήρωμα	積分	积分
irreducible chain	Nicht-reduzierbare Kette	Cadena irreducible	Αμείωτη αλυσίδα (αδιαχώριστη αλυσίδα ή μη αναγώγιμη αλυσίδα)	既約マルコフ連鎖	不可约链
joint distribution	Gemeinsame Verteilung	Distribución conjunta	Απο κοινού κατανομή	同時分布(結合分布)	联合分布
jump chain	Springende Markov-Kette/ Sprungkette		αλυσίδα μεταπτώσεων	ジャンプ連鎖	跳跃链 (嵌入链)
likelihood	Likelihood	Verosilimilitud	Πιθανοφάνεια	尤度	似然(值)
likelihood function	Likelihood-Funktion	Función de verosimilitud	Συναρτηση πιθανοφάνειας	尤度関数	似然函数
likelihood ratio test	Likelihood-Quotienten Test	Test de cociente de verosimilitud	τεστ λόγου πιθανοφάνειας	尤度比検定	似然比检验
marginal distribution	Die Randverteilung	Distribución marginal	περιθώρια κατανομής	周辺分布	边缘分布
Markov chain	Markov-Kette	Cadena Markov	Αλυσίδα Μαρκόφ (Μαρκοβιανή Αλυσίδα)	マルコフ連鎖	马尔可夫链
Markov chain Monte Carlo	Markov-Chain Monte Carlo Verfahren	MCMC	Μοντε Κάρλο μέθοδοι με χρήση Μαρκοβιανών Αλυσίδων	マルコフ連鎖モンテカルロ	马氏链蒙托卡罗
matrix	Matrix	Matriz	πίνακας	行列	矩阵
null hypothesis	Nullhypothese	Hipótesis nula	Μηδενική υπόθεση	帰無仮説	零假设
off-diagonal elements	Die nicht auf der Diagonale einer Matrix liegenden Elemente	Elementos no-diagonales	Στοιχεία εκτος διαγωνίου (μη διαγώνια στοιχεία)	非対角要素	非对角元素
period	Periode	Periodo	Περίοδος	周期	周期
polynomial	Das Polynom/ polynomial (adj.)	polinómico	Πολυώνημο	多項式	多项式
posterior distribution	A posteriori Verteilung	Distribución a-posteriori	posterior κατανομή ή εκ των υστέρων κατανομή	事後分布	后验分布
prior distribution	A priori Verteilung	Distribución a-priori		事前分布	先验分布
probability density function (PDF)	Wahrscheinlichkeitsdichtefunktion/Dichtefunktion	Función de densidad de probabilidad	Συνάρτηση πυκνότητας πιθανότητας	確率密度関数	概率密度函数

recurrent states	Wiederkehrende Zustände	Estado recurrentes		再帰状態	常返状态
Simpson's method	Simpsonregel (auch Keplersche Fassregel) zur numerischen Berechnung von Integralen	Método de Simpson	Μέθοδος Σίμσον	シンプソン法	Simpson方法
scalar	Skalar	escalar	μονόμετρος ή βαθμωτός	スカラー	标量
slope	Steigung	Gradiente / Pendiente	κλίση	勾配(傾斜)	斜率(梯度)
state space	Zustandsraum	Espacio de estados	χώρος κατάστασης	状態空間	状态空间
stationary distribution	Stationäre/gleichbleibende Verteilung	Distribución estacionaria	Στατική κατανομή (στάσιμη κατανομή)	定常分布	平稳分布
Taylor expansion	Taylorreihenentwicklung	Expansión de Taylor	επέκταση Taylor	テイラー展開	泰勒展开
time-homogeneous	Zeitlich homogen/ zeitlich homogener Prozess	Homogéneo en el tiempo	Χρονικά ομογενής (Ομογενής)	斉時的(マルコフ連鎖)	时齐(马尔可夫链)
transition matrix	Übergangsmatrix	Matriz de transición	Πίνακας μετάβασης	遷移行列	转移矩阵
transition probability	Übergangswahrscheinlichkeit	Probabilidad de transición	Πιθανότητα μετάβασης	遷移確率	转移概率
trapezoid method	Trapezregel-Methode zur Berechnung von Integralen	Regla del trapecio	Τραπεζοειδές μέθοδος (μέθοδος του τραπεζίου)	台形法	梯形法
type-I error	Fehler erster Art	Error tipo I	Σφάλμα τύπου 1	第一種の過誤	I-型错误
type-II error	Fehler zweiter Art	Error tipo II	Σφάλμα τύπου 2	第二種の過誤	II-型错误
vector	Vektor	Vector	διάνυσμα	ベクトル	向量(矢量)
Venn diagram	Venn-Diagramm/Mengendiagramm	Diagrama de Venn	Διάγραμμα Βενν	ベン図	封氏图