# 12 Appendix. Mathematical, computational and statistical background

Below you can find condensed descriptions of concepts and methods used in these notes. If you have a basic knowledge of some area these descriptions can serve as a repetition, but if some concepts are new to you, you presumably need to go to textbooks for more complete information. Nowadays quite useful information can also be obtained from the internet, for example from the Wikipedia pages.

## 12.1 Some matrix algebra

A matrix with $m$ rows and $n$ columns, or briefly a matrix of type $m \times n$, is a rectangular array

$$\begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix} \tag{94}$$

of numbers $a_{i,j}$, sometimes written $a_{ij}$, called matrix elements. If the type is understood we can write $A = [a_{i,j}]$. Row and column vectors are thin matrices with $m = 1$ and $n = 1$, respectively. If $m = n = 1$ the matrix is just a number. A square matrix has $m = n$.

Let $A$ be an $m \times n$ matrix. The transpose $A^T$ of $A$ is an $n \times m$ matrix obtained by making rows in $A$ into columns, that is the $(i, j)$ element in $A^T$ is the $(j, i)$ element in $A$. A matrix is symmetric if it equal to its transpose.

Matrices of the same type can be added by element-wise addition. If $A$ and $B$ are matrices of types $m \times n$ and $n \times k$, respectively, the product $C = AB$ is a matrix type $m \times k$ with elements $c_{i,j} = \sum_r a_{i,r} b_{r,j}$. A square $n \times n$ matrix A is called invertible (or non-singular) if there exists an inverse denoted $A^{-1}$ such that

$$AA^{-1} = A^{-1}A = I \tag{95}$$

where $I$ is the unit $n \times n$ matrix with diagonal elements $i_{j,j} = 1$ and off-diagonal elements $i_{j,k} = 0, j \neq k$.

Let us now define recursively the determinant $\det A$ of a square $n \times n$ matrix $A = [a_{i,j}]$. For $n = 1$ we define $\det A = a$ for the matrix $A = [a]$. Suppose that we have defined determinants for matrices of type $(n-1) \times (n-1)$ and let $A$ be a matrix of type $n \times n$. Let the minor $A_{i,j}$ be the determinant of the matrix obtained from $A$ by deleting row number $i$ and column number $j$. Then we put

$$\det A = \sum_{j=1}^{n} (-1)^{1+j} a_{1,j} A_{1,j}. \tag{96}$$

One can show that a square matrix $A$ is non-singular if and only if $\det A \neq 0$.

Let $A$ be a square matrix. We say that a real number $\lambda$ is an eigenvalue of $A$ and that a column vector $x$ is an eigenvector of $a$ if

$$Ax = \lambda x. \tag{97}$$

A symmetric real $n \times n$ matrix $A$ is said to be positive-definite or positive-semidefinite if $x^T A x > 0$ or $x^T A x \geq 0$, respectively, for each non-zero $n$-dimensional column vector $x$. One can show that a symmetric matrix is positive-definite or positive-semidefinite if all its eigenvalues are positive or nonnegative, respectively. Further, a positive definite matrix is invertible.

**Exercises**

*Exercise 11.1.* Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Determine $\det A$ by use of (96).

*Exercise 11.2.* Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with $ad - bc \neq 0$. Determine the inverse of $A$ by solving a linear equation system with four unknowns.

## 12.2 Optimization of a real funtion

Let us first consider Newton's method for optimization of a twice continuously differentiable real-valued function $f(x)$ of a real variable $x$. Suppose that $f$ has a maximum or minumum at $x^\star$. Then $f'(x^\star) = 0$. Newton's iterative method for locating $x^\star$ is to put

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}. \tag{98}$$

Assuming that $f''(x^\star) \neq 0$ and that we start close enough to $x^\star$ one can show that $x^k \to x^\star$ as $k \to \infty$.

Let us now consider Newton's method for optimization of a twice continuously differentiable real-valued function $f(x)$ of an $n$-dimensional column vector $x$. As above we suppose that $f$ has a maximum or minumum at $x^\star$. Let $\nabla f(x)$ denote the (column) gradient vector

$$\nabla f(x) = [\frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_n}]^T \tag{99}$$

and let $Hf(x)$ denote the Hessian matrix

$$Hf(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} \tag{100}$$

Newton's iterative method for locating $x^\star$ is to put

$$x^{k+1} = x^k - (Hf(x^k))^{-1} \nabla f(x^k) \tag{101}$$

Assuming that $Hf(x^\star)$ is positive-definite and thus invertible, and that we start close enough to $x^\star$ one can show that $x^k \to x^\star$ as $k \to \infty$.

Newton's method is quite efficient but has drawbacks. Computation of derivatives can require a lot of programming. One may use finite differences to compute approximate derivatives but that then it requires extra programming to find suitable step lengths. Often it is more efficient to use so called quasi-Newton methods where the Hessian is automatically estimated from successively computed gradient vectors, see for instance Press et al. (2007). In MATLAB the FMINUNC function uses a quasi-Newton metod for minimization.

The Newton and quasi-Newton methods typically work quite well if you start close to the optimum. A much slower but quite robust optimizer, which does not require computation of any derivates, is the simplex method of (Nelder & Mead, 1965) which is available in MATLAB as the function NELDER_MEAD. A good strategy in applications can often be to begin with the simplex metod to get an overview and suitable starting values and then to use a quasi-Newton method.

## 12.3 Discrete probability distributions

Discrete distributions for a random variable $X$ are characterized by the probability function $\Pr(X = x)$, $x \in V$, where $V$ is the finite or countable set of values that $X$ can take. For a real-valued discrete random variable the expectation $\mu$, standard deviation $\sigma$ and variance $\sigma^2$ are defined by $\mu = \mathbf{E}(X) = \sum_x x \Pr(X = x)$ and $\sigma^2 = \mathrm{var}(X) = \sum_x (x-\mu)^2 \Pr(X = x)$.

A random variable $X$ is said to be Poisson distributed with parameter $\lambda$ if

$$\Pr(X = n) = \frac{\lambda^n}{n!}\exp(-\lambda), \quad n = 0, 1, \ldots, \tag{102}$$

and for such a variable both the expectation and the variance are equal to $\lambda$.

A random variable $X$ is said to be binomial $(n,p)$ if

$$\Pr(X = k) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k (1-p)^{n-k}, \quad k = 0, \ldots, n, \tag{103}$$

and for such a variable the expectation is $np$ and the variance is $np(1-p)$.

## 12.4 Continuous probability distributions

Continuous distributions for a real-valued random variable $X$ are characterized by the probability density

$$f(x) = \frac{d}{dx}\Pr(X \le x), \quad x \in \mathbb{R}, \tag{104}$$

where $\mathbb{R} = (-\infty, \infty)$ is the set of real numbers. For a continuous random variable the expextation $\mu$, standard deviation $\sigma$ and variance $\sigma^2$ are defined by $\mu = \mathbf{E}(X) = \int_{\mathbb{R}} x f(x) dx$ and $\sigma^2 = \mathrm{var}(X) = \int_{\mathbb{R}} (x-\mu)^2 f(x) dx$.

A random variable $X$ is said to have a uniform distribution on the interval $(a, b)$ if the probability density is

$$f(x) = 1/(b-a), \quad a < x < b, \tag{105}$$

and $f(x) = 0$ for $x < a$ and $x > b$, and for such a variable the expectation is $(a+b)/2$ and the variance is $(b-a)^2/12$.

A random variable $X$ is said to have an exponential distribution with parameter $\beta$ if the probability density is

$$f(x) = \beta \exp(-\beta x), \;\; x > 0, \tag{106}$$

and $f(x) = 0$ for $x < 0$, and for such a variable the expectation is $1/\beta$ and the variance is $1/\beta^2$.

A random variable $X$ is said to be normal($\mu$,$\sigma^2$), or briefly $X \sim N(\mu,\sigma^2)$ if the probability density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x-\mu)^2/\sigma^2), \;\; x \in \mathbb{R}, \tag{107}$$

and for such a variable the expectation is $\mu$ and the variance is $\sigma^2$.

## 12.5 Multivariate probability distributions

Let $X_1, \ldots, X_d$ be real-valued random variables. Then $X = [X_1 \ldots X_d]^T$ is a $d$-dimensional random (column) vector. The expectation of a random vector (or a random matrix) is defined componentwise. Thus the expectation vector$\mu = \mu_X = \mathbf{E}(X)$ of a random column vector $X$ is the column vector with components $\mu_i = \mathbf{E}(X_i), i = 1, \ldots, d$. The covariance matrix $C = C_X = C(X)$ of $X$ is the symmetric $d \times d$ matrix

$$C = \mathbf{E}(X-\mu)(X-\mu)^T = \begin{bmatrix} E(X_1-\mu_1)(X_1-\mu_1) & \cdots & E(X_1-\mu_1)(X_d-\mu_d) \\ \vdots & & \vdots \\ E(X_d-\mu_d)(X_1-\mu_1) & \cdots & E(X_d-\mu_d)(X_d-\mu_d) \end{bmatrix}. \tag{108}$$

The $(i,j)$-element of the covariance matrix of $X$ is the covariance $\mathrm{cov}(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j)$ of the $i$th and $j$th components of $X$, which for $i = j$ is the variance of $X_i$.

The $d$-dimensional vector $X$ has a $d$-dimensional probability density $f = f_X$ if

$$\Pr(X \in A) = \int_A f(x)dx \tag{109}$$

for subsets $A$ of $d$-dimensional space $\mathbb{R}^d$ for which the integral in (109) is well-defined.

Let $\mu$ be a $d$-dimensional column vector and let $C$ be a positive-definite $d \times d$ matrix. The $d$-dimensional random vector $X$ is said to be normal($\mu$,$C$) or briefly $X \sim N(\mu,C)$ if $X$ has the $d$-dimensional density function

$$f_X(x) = \frac{1}{(2\pi)^{d/2}(\det C)^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)), \tag{110}$$

where $\det C$ denotes the determinant of the matrix $C$. One can show that then $X$ has expectation vector $\mu$ and covariance matrix $C$.

An important special case is the two-dimensional normal distribution. Regard $X = [X_1 \;\; X_2]^T$. Let $\mu_i$ and $\sigma_i^2$ denote the expectation and variance of $X_i$, $i = 1, 2$, and let

$\rho = \operatorname{cov}(X_1, X_2)/(\sigma_1\sigma_2)$ denote the correlation between the two components of $X$. Thus the covariance matrix of $X$ is

$$C = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{111}$$

One can then show that the two-dimensional density funcion of $X$ is

$$f(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}Q(x_1, x_2)\} \tag{112}$$

where

$$Q(x_1, x_2) = \frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho(\frac{x_1-\mu_1}{\sigma_1})(\frac{x_2-\mu_2}{\sigma_2}) + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \tag{113}$$

## 12.6 Random, Gaussian and Markov processes on the real line

A random process or stochastic process $X$ on the real line consists a set of random variables $X = (X_t)$ indexed by time $t \in T$, where $T$ is a subset of the real line $\mathbb{R}$. We suppose here that $T$ is either a set of consecutive integers or an interval and then we talk about a discrete time or continuous time random process, respectively. The set $V$ of values that $X_t$ can take we call the state space. A real-valued process has the real line or a subset of it as state space. A real-valued random process may be characterized by its mean value function,

$$m_t = \mathbf{E}X_t \tag{114}$$

and its covariance function

$$C(s,t) = \mathbf{E}(X_s - m_s)(X_t - m_t). \tag{115}$$

A random process is said to be normal or Gaussian if $(X_{t_1}, \ldots, X_{t_n})$ has an $n$-dimensional normal distribution for any choice of time points $t_1, \ldots, t_n$. One can show that a Gaussian process is fully specified by its mean value and covariance functions.

A random process $(X_t)$ is said to be stationary if its distribution is invariant under a translation $\tau$, more precisely if for each choice of $n \geq 1$ and $(t_1, \ldots, t_n)$ the distribution of the $n$-dimensional random vector $(X_{t_1+\tau}, \ldots, X_{t_n+\tau})$ does not depend on $\tau$. Consider the mean value and covariance functions of a stationary process. The mean value is a constant $m = \mathbf{E}X_t$ and the covariance function can be written as $C(s,t) = \sigma^2\rho(t-s)$ where the variance $\sigma^2 = C(t,t)$ and $\rho(t)$ is the correlation function.

We say that $(X_t, t \in T)$ is a Markov process if the conditional distribution of $X$ at a future time given the history up to time $t$ only depends on the value of $X$ at the current time $t$, more precisely if

$$\Pr(X_\tau \in A | X_s, s \leq t) = \Pr(X_\tau \in A | X_t), \ \ t < \tau. \tag{116}$$

A discrete time Markov process with finite state space $V$, for notational simplicity here denoted $V = \{1, \ldots v\}$, is determined by its transition probability matrix $P$ which is the $(v \times v)$ matrix with elements

$$p_{ij} = \Pr(X_{t+1} = j | X_t = i), \ \ 1 \leq i, j \leq v. \tag{117}$$

A zero-mean autoregressive process $(X_t)$ of order $p$ is recursively generated from

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \epsilon_t \tag{118}$$

where $\epsilon_t$ are independent and identically distributed random variables with zero mean and finite variance $\sigma^2$. Often $\epsilon_t$ is assumed to be normally distributed. Then $X_t$ is also normally distributed. An autoregressive process of order $p = 1$ is a Markov process. An autogressive process of order one is stationary if $|a_1| < 1$ and the starting value in (118) is suitably chosen.

An example of a continuous time Markov process is the Poisson process with intensity $\lambda$ which is characterized by the fact that the increment $X_t - X_s$ is Poisson distributed with expecation

$$\mathbf{E}(X_t - X_s) = \lambda(t - s), \;\; s < t, \tag{119}$$

and the increments over disjoint time intervals are independent.

Suppose that points are randomly placed on the real line such that

(i) the number of points in disjoint intervals are independent,

(ii) the probability that two points are placed in an interval of length $h$ tends to zero faster than the probability that one point is placed in the same interval when $h \to 0$ ,

(iii) the distribution of the number of points in an interval depends only on the length of the interval and not on where it is placed.

One can then show that if $X_t$ denotes the number of points in the interval $(0, t)$, then $(X_t, t > 0)$ is Poisson process with intensity $\lambda$ equal to the expected number of points in an interval of unit length. For an arbitrary time $t$ let further $W$ denote the waiting time for the first point after $t$. One can then show that $W$ has an exponential distribution with parameter $\lambda$.

Another example of a continuous time Markov process is the Brownian motion or Wiener process on the interval $[0, \infty)$ characterized by having independent increments over disjoint time intervals and that $X_t$ is normal$(0, \sigma^2 t)$ for $t \geq 0$.

A third example of a continuous time Markov process is the Ornstein-Uhlenbeck process, which is Gaussian process with mean zero and correlation function

$$\rho(t) = \exp(-\lambda t) \tag{120}$$

for some positive constant $\lambda$.

## 12.7 Estimation of parameters. Likelihood and least squares

Suppose that we observe a random variable or vector $X$ with a distribution that depends on a parameter $\theta$ that may be a vector. Let $\hat{\theta} = \hat{\theta}(X)$ be an estimate of $\theta$. We say that $\hat{\theta}$ is an unbiased estimate of $\theta$ if

$$\mathbf{E}(\hat{\theta}) = \theta. \tag{121}$$

Typically we observe a sample of a random variable which means that we have a sequence of independent and identically distributed random variables. We say that $\hat{\theta}$ is a consistent estimate of $\theta$ if for an arbitrary $\epsilon > 0$

$$\Pr(|\hat{\theta} - \theta| > \epsilon) \to 0 \tag{122}$$

as the number $n$ of observations goes to infinity. One can for instance show that $\hat{\theta}$ is a consistent estimate of $\theta$ if $\mathbf{E}(|\hat{\theta} - \theta|^2) \to 0$ as $n \to \infty$.

Let $X$ be a discrete or continuous random vector that we observe and that has a probability distribution depending on $\theta$. If $X$ is discrete we put $f(x, \theta) = \Pr(X = x)$ and if $X$ is continuous $f(x, \theta)$ denotes the probability density of $X$. The likelihood value corresponding to an observed value $x$ of $X$ is written

$$L(\theta) = L(\theta|x) = f(x, \theta). \tag{123}$$

In particular, if we have a sample $X = (X_1, \ldots, X_n)$ of a random variable assumed to be either discrete with probability function $\Pr(X_i = x_i) = f(x_i, \theta)$ or continuous with probability density $f(x_i, \theta)$ the corresponding likelihood function is

$$L(\theta) = L(\theta|x) = \prod_{i=1}^{n} f(x_i, \theta), \tag{124}$$

where $x = (x_1, \ldots, x_n)$.

A maximum likelihood estimate $\hat{\theta}$ of $\theta$ is a value that maximizes the likelihood function. In practice it is often more convenient to maximize the log-likelihood function

$$\ell(\theta) = \log(L(\theta)), \tag{125}$$

where log (as always in these notes) denotes the natural logarithm.

As an example, suppose that $X = (X_1, \ldots, X_n)$ is a sample of a variable that is Poisson distributed with parameter $\lambda$, that is $X_1, \ldots, X_n$ are independent and identically Poisson distributed. The log-likelihood function is

$$\ell(\lambda) = \log(\prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i!} \exp(-\lambda)) = c - n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i, \tag{126}$$

where $c$ does not depend on $\lambda$ and thus can be disregarded during the maximization. One finds that the maximum likelihood estimate of $\lambda$ is

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{127}$$

which one can show is a both unbiased and consistent estimate of $\lambda$. (In the computations in this example we have used the notation $X_i$ rather than $x_i$ which is often convenient.)

A useful complement to the maximum likelihood method to estimate parameters is the least squares method which, when applicable, is often easier to use. Suppose that

$X_1 \ldots, X_n$ are independent random variables with the same variance and with an expection that depends on a parameter $\theta$. The least squares estimate $\hat{\theta}$ is obtained by minimizing

$$Q(\theta) = \sum_{i=1}^{n} (X_i - \mathbf{E}(X_i))^2. \tag{128}$$

Let us again consider a sample $(X_1, \ldots, X_n)$ of a random variable that is Poisson distributed with parameter $\lambda$. The sum of squares (128) now becomes

$$Q(\lambda) = \sum_{i=1}^{n} (X_i - \lambda)^2, \tag{129}$$

which is minimized for $\lambda = \hat{\lambda}$ in (127). Thus the least squares and the maximum likelihood estimates coincide in this example.

## 12.8 Linear and logistic regression

Let us first consider linear regression with one explaining real variable $x$. Suppose that we observe

$$Y_i = \alpha + \beta x_i + \epsilon_i, \;\; i = 1, \ldots n, \tag{130}$$

with independent zero-mean random errors $\epsilon_i$, $i = 1, \ldots, n$, with identical variances. The least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are obtained by minimizing

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2, \tag{131}$$

which gives

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{x}, \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \tag{132}$$

where $\overline{x} = (1/n) \sum_i x_i$ and $\overline{Y} = (1/n) \sum_i Y_i$.

Let us now consider multiple linear regression with $m$ explaining variables. We assume that we have observations

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_m x_{im} + \epsilon_i, \;\; i = 1, \ldots n, \tag{133}$$

with independent zero-mean random errors $\epsilon_i$, $i = 1, \ldots, n$, with identical variances. We can write our observations on vector-matrix form as

$$Y = X\beta + \epsilon, \tag{134}$$

where

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{135}$$

It turns out that the least squares estimate of the parameter vector $\beta$ is

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{136}$$

Let us now consider logistic regression where we observe independent variables $Y_1, \ldots, Y_n$ taking values 0 or 1. We suppose that the probability $p_i = \Pr(Y_i = 1) = 1 - \Pr(Y_i = 0)$ depends on $m$ explaining variables such that

$$\log(\frac{p_i}{1 - p_i}) = \beta_1 x_{i1} + \ldots + \beta_m x_{im}, \ \ i = 1, \ldots n. \tag{137}$$

To estimate the parameters $\beta_1, \ldots, \beta_m$ we can maximize the likelihood function

$$L(\beta_1, \ldots, \beta_m) = \prod_{i=1}^{n} (p_i^{Y_i} (1 - p_i)^{1 - Y_i}). \tag{138}$$

There is no analytical expression for the maximum likelihood estimates so to maximize (138) one may use computational optimization methods such as those describe in Section 12.2 and then it is typically more convenient to maximize the log-likelihood function.

### 12.9 Confidence intervals and tests, observations from a normal distribution, the t- and chi-square distributions

Let $X$ denote observations from a distribution depending on a real-valued parameter $\theta$. We say that the interval $(L(X), U(X))$ is a confidence interval for $\theta$ with confidence degree $p$ if

$$\Pr(L(X) < \theta < U(X)) = p. \tag{139}$$

Let $X = (X_1, \ldots, X_n)$ be a sample from a normal$(\mu, \sigma^2)$ distribution. Then

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{140}$$

are unbiased and consistent estimates of $\mu$ and $\sigma^2$, respectively. To compute confidence intervals for $\mu$ and $\sigma^2$ we introduce the chi-square and $t$-distributions.

A random variable is said to be chi-square distributed with $r$ degrees of freedom if it has the same distribution as

$$\chi^2 = \sum_{i=1}^{r} Z_i^2, \tag{141}$$

where $Z_1, \ldots, Z_r$ are independent and normal$(0, 1)$. Let us note that a variable that is chi-square distributed with $r$ degrees of freedom has expectation $r$. A random variable is said to be $t$-distributed with $r$ degrees of freedom if it has the same distribution as

$$t = \frac{Z}{\sqrt{\chi^2 / r}} \tag{142}$$

where $Z$ and $\chi^2$ are independent and distributed normal$(0,1)$ and chi-squared with $r$ degrees of freedom, respectively.

Let us define quantiles for random variables with a continuous distribution function $F(x) = \Pr(X \leq x)$. A $p$th quantile $x_p$ corresponding to such a distribution satisfies $F(x_p) = p$. Let $\chi^2_p$ denote the $p$th quantile of a chi-square distribution with $n-1$ degrees of freedom. For $s^2$ defined by (140) one can then show that

$$\Pr(\chi^2_{(1-p)/2} < (n-1)s^2/\sigma^2 < \chi^2_{(1+p)/2}) = p \tag{143}$$

which gives a confidence interval for $\sigma^2$ with confidence degree $p$,

$$\Pr(\frac{(n-1)s^2}{\chi^2_{(1+p)/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{(1-p)/2}}) = p. \tag{144}$$

Similarly we let $t_p$ denote the $p$th quantile of a $t$-distribution with $n-1$ degrees of freedom. Then

$$\Pr(\overline{X} - t_{(1-p)/2}\ s/\sqrt{n} < \mu < \overline{X} + t_{(1-p)/2}\ s/\sqrt{n}) = p, \tag{145}$$

which gives a confidence interval for $\mu$ with confidence degree $p$.

Let us also briefly describe one type of test of an hypothesis $H_0 : \theta = \theta_0$. Suppose that we have a test variable $T = T(X)$ tending to take large values when the hypothesis $H_0$ is not true and that we for our observations obtain an observed value $T_{obs}$ of $T$. The strategy can then be to reject the hypothesis $H_0$ if the probability under $H_0$ to obtain a $T$-value at least as large as the observed value is small enough. More precisely we reject $H_0$ if the $p$-value

$$p = \Pr_0(T \geq T_{obs}) \tag{146}$$

is small enough. Here $\Pr_0$ denotes a probability evaluated under the probability distribution corresponding to $H_0$.

As an example let us suppose that we have a random sample $(X_1, \ldots, X_n)$ from a $N(\mu, \sigma^2)$ distribution and that we want to test the hypothesis $H_0 : \mu = \mu_0$ with the alternative hypothesis that $\mu$ is either larger or smaller than $\mu_0$. Let $X$ and $s^2$ be defined as in (140) and put

$$t_{obs} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}. \tag{147}$$

The corresponding $p$-value is then

$$p = P(|t| \geq |t_{obs}|) \tag{148}$$

evaluated with the assumption that $t$ is $t$-distributed with $n-1$ degrees of freedom.

## 12.10 The F-distribution, analysis of variance

A random variable is $F$-distributed with $(r_1, r_2)$ degrees of freedom if it has the same distribution as

$$F = \frac{\chi^2_1/r_1}{\chi^2_2/r_2}, \tag{149}$$

where $\chi_1^2$ and $\chi_2^2$ are independent chi-square distributed variables with $r_1$ and $r_2$ degrees of freedom, respectively. The $F$-distribution can be used to compare two variance estimates and in analysis of variance (ANOVA) models. Let us consider a simple ANOVA model.

Assume that $X_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ are independent normal variables with identical variance $\sigma^2$ and expectations

$$\mathbf{E}(X_{ij}) = \mu_i, \;\; i = 1, \ldots, m, \, j = 1 \ldots, n_i. \tag{150}$$

To test the hypothesis $H_0 : \mu_1 = \ldots = \mu_m$ we can use the test variable

$$F = \frac{\sum_{i=1}^{m} n_i(\overline{X_{i\cdot}} - \overline{X_{\cdot\cdot}})^2 \;/\; (m-1)}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X_{i\cdot}})^2 \;/\; (\sum_i (n_i - 1))} \tag{151}$$

where $\overline{X_{i\cdot}} = (1/n_i)\sum_j X_{ij}$ and $\overline{X_{\cdot\cdot}} = (\sum_i \sum_j X_{ij})/(\sum_i n_i)$. It turns out that under $H_0$ the test variable $F$ in (151) is $F$-distributed with $(m-1, \sum_i (n_i - 1))$degrees of freedom and we reject the hypothesis $H_0$ if $F$ is large enough.

## 12.11 Approximate statistical methods, bootstrap

In the previous sections we have seen how confidence intervals with exact confidence degree and exact $p$-values for tests can be computed for simple models with normal random variables. Otherwise such exact statistical inference is typically not possible. However, for large samples good approximate methods are often available. Let us give some examples of how such approximate methods can look.

Suppose that we have a sample $X = (X_1, \ldots, X_n)$ of a random variables with log-likelihood $\ell(\theta)$, see (125), depending on a parameter vector $\theta = (\theta_1, \ldots, \theta_d)$. Under suitable regularity conditions, see for instance Pawitan (2001), one can then show that for large $n$ the maximum likelihood estimate $\hat{\theta}$ has an approximate $d$-dimensional normal distribution, which we write

$$\hat{\theta} \xrightarrow{d} N(\theta, \mathcal{I}(\hat{\theta})^{-1}). \tag{152}$$

Here $\mathcal{I}(\hat{\theta})$ is the Fisher information matrix with matrix elements

$$\mathcal{I}_{ij}(\hat{\theta}) \;=\; -\frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta)|_{\theta=\hat{\theta}} \tag{153}$$

and we suppose that $\mathcal{I}(\hat{\theta})$ is invertible. From this we can compute confidence intervals with approximate $p$-values for the components of $\theta$ and more generally for linear combinations of these components. Let us note that the Fisher information matrix is the Hessian (see Section 12.2) of the log-likelihood function and as discussed in Section 12.2 the Hessian can be obtained by use of quasi-Newton optimization methods.

Let us now consider two hypotheses $H_0$ and $H_1$, which are nested in such a way that $H_0$ is obtained from $H_1$ by imposing $r$ linear restrictions on the parameters, for instance by putting $r$ parameters equal to zero. Let $\ell(\hat{\theta}_0)$ and $\ell(\hat{\theta}_1)$ denote the log-likelihoods corresponding to the maximum likelihood estimates obtained under $H_0$ and $H_1$. Put

$$\chi^2 = 2(\ell(\hat{\theta}_1) - \ell(\hat{\theta}_0)). \tag{154}$$

We note that as $\ell(\hat{\theta}_1)$ is obtained as a maximum under fewer restrictions than $\ell(\hat{\theta}_0)$ it follows that $\ell(\hat{\theta}_1) \geq \ell(\hat{\theta}_0)$. One can show that under the hypothesis $H_0$ the variable $\chi^2$ in (154) is approximately chi-square distributed with $r$ degrees of freedom for large samples. We can reject the hypothesis $H_0$ if the observed $\chi^2$-value is large enough, that is if the corresponding $p$-value

$$p = \Pr(\chi^2 \geq \chi^2_{obs}) \tag{155}$$

evaluated for a chi-square distribution with $r$ degrees of freedom is small enough.

One method for obtaining approximate inference that has been much used since its introduction 1979 is the bootstrap which is based on resampling from observed distributions in such a way that confidence intervals and test variables can be computed, see for instance Efron & Tibshirani (1993).

## 12.12 Random numbers, simulation

An important method to study random systems is to use simulation and this requires generation of random numbers, or more precisely pseudo-random numbers, with computers. A basic random number generator is the linear congruential generator

$$X_{n+1} = (aX_n + b) \mod m \tag{156}$$

with suitable integers $a$, $b$ and $m$ and a starting value $X_0$ called seed. This generates a sequence with approximately independent random number equidistributed on the set of integers $\{0, 1, \ldots, m-1\}$. This type of generators with some variations are used as basic random generators in computer languges such as for MATLAB. Putting $U_n = X_n/m$ gives a sequence of random numbers with an approximate uniform distribution on the unit interval $[0, 1]$.

Suppose now that we have a random number $U$ with a uniform distribution on the interval $(0, 1)$ and that we want a random number $X$ with a given distribution function $F(x) = \Pr(X \leq x)$. This can be obtained by putting

$$X = F^{-1}(U), \tag{157}$$

where $F^{-1}$ denotes the inverse of $F$. Putting

$$X = -\frac{1}{\beta}\log(1 - U) \tag{158}$$

gives for instance a random variable that is exponentially distributed with parameter $\beta$.

Sometimes one wants a random number with uniform distribution on a bounded two-dimensional set $A$. One can then use rejection sampling by first finding a rectangle $R = \{(x_1, x_2) : a_1 \leq x_1 \leq b_1, a_2 \leq x_2 \leq b_2\}$ containing $A$ as a subset. Generate then two independent random numbers $U_1$ and $U_2$ with uniform distributions on the unit interval. Put $X = (a_1 + (b_1 - a_1)U_1, a_2 + (b_2 - a_2)U_2)$. If $X \in A$ accept $X$, otherwise reject $X$ and repeat the procedure until we get a point in $A$.

## 12.13 Bayesian inference, Markov chain Monte Carlo

In Bayesian inference we have in addition to a model describing the distribution of observations $X$ given parameter $\theta$ also a random distribution for $\theta$ called the prior distribution. After obtaining observations of $X$ the distribution of $\theta$ is modified to the posterior distribution. Let us show how this goes when both $\theta$ and $X$ are discrete variables, the formulas when one or both of these variables have continuous distributions being similar. We let $\pi_i$ denote the prior probability, $\pi_i = \Pr(\theta = \theta_i)$.

From the definition of conditional probabilities for events $A$ and $B$ we have $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$. This gives the posterior distribution for $\theta$ when we observe $X = x$ as follows.

$$\Pr(\theta = \theta_i | X = x) = \frac{\Pr(X = x|\theta_i)\pi_i}{\Pr(X = x)} = \frac{\Pr(X = x|\theta_i)\pi_i}{\sum_j \Pr(X = x|\theta_j)\pi_j} \tag{159}$$

In Bayesian analysis of noisy observations of complicated high-dimensional objects such as images it is not easy to evaluate or sample from the posterior distribution. One general method that has ben much used in recent years is Markov chain Monte Carlo, abbreviated MCMC. Here you construct a Markov chain which has the distribution of interest as its stationary distribution. Useful algorithms for constructing and analyzing such Markov chains are the Gibbs sampler and the Metropolis algorithm, see for instance Gilks et al. (1996).

## 12.14 Prediction, Kalman filtering

Let us look at prediction and filtering by use of Kalman filters. We let the $d$-dimensional column vector $X_t, t = 0, 1, \ldots,$ denote the state of a system at time $t$. Assume that $X_0 \sim N(\mu_0, P_0)$ and that

$$X_t = F_t X_{t-1} + W_t, \quad t = 1, 2, \ldots, \tag{160}$$

where $F_t$ is a $d \times d$ matrix. Suppose that the dynamic $d$-dimensional noise vectors $W_t \sim N(0, Q_t)$ are independent mutually and of the initial state $X_0$. Assume further that we observe the $r$-dimensional vectors

$$Y_t = H_t X_t + V_t, \quad t = 1, 2, \ldots, \tag{161}$$

where $H_t$ is a $r \times d$ matrix and the measurement noise vectors $V_t \sim N(0, R_t)$ are independent mutually and of $(W_t)$ and the initial state $X_0$. Let $Y_{1:t} = (Y_1, \ldots, Y_t)$ denote the accumulated observations up to time $t$. We are interested in computing the optimal estimate of $X_t$ given observations up to time $t$. It turns out that given $Y_{1:t}$ the conditional distribution of $X_t$ is normal with expectation

$$\hat{X}_{t|t} = \mathbf{E}(X_t | Y_{1:t}) \tag{162}$$

and covariance matrix $P_{t|t}$. We will give a recursive algorithm for computing $\hat{X}_{t|t}$ and $P_{t|t}$ which also gives the conditional expectation and covariance matrix $\hat{X}_{t|t-1}$ and $P_{t|t-1}$ for prediction of $X_t$ from observations $Y_{1:t-1}$ up to time $t-1$. The algorithm consists of the following six equations in going from $\hat{X}_{t-1|t-1}$ and $P_{t-1|t-1}$ to $\hat{X}_{t|t}$ and $P_{t|t}$,

$$\hat{X}_{t|t-1} = F_t \hat{X}_{t-1|t-1}, \tag{163}$$

$$P_{t|t-1} = F_t P_{t-1|t-1} F_t^T + Q_t, \tag{164}$$

$$S_t = H_t P_{t|t-1} H_t^T + R_t, \tag{165}$$

$$K_t = P_{t|t-1} H_t^T S_t^{-1}, \tag{166}$$

$$\hat{X}_{t|t} = \hat{X}_{t|t-1} + K_t (Y_t - H_t \hat{X}_{t|t-1}), \tag{167}$$

$$P_{t|t} = (I - K_t H_t) P_{t|t-1}, \tag{168}$$

where $I$ denotes the unit $d \times d$-matrix.

Consider as an example motion of an object with centre at $(x_t, y_t)$ and velocity $(\dot{x}_t, \dot{y}_t)$ with a sampling interval $\Delta t$ and observation of the position but not the velocity. We can then put

$$X_t = \begin{bmatrix} x_t \\ y_t \\ \dot{x}_t \\ \dot{y}_t \end{bmatrix}, \quad F_t = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad H_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \tag{169}$$