The Open University

M249
Handbook

Contents

This version of the Handbook is not for use in the examination.

1.1
SUP 86991 7

1 Greek alphabet

α	A	Alpha	ι	I	Iota	ρ	P	Rho
β	B	Beta	κ	K	Kappa	σ	Σ	Sigma
γ	Γ	Gamma	λ	Λ	Lambda	τ	T	Tau
δ	Δ	Delta	μ	M	Mu	υ	Υ	Upsilon
ε	E	Epsilon	ν	N	Nu	ϕ	Φ	Phi
ζ	Z	Zeta	ξ	Ξ	Xi	χ	X	Chi
η	H	Eta	o	O	Omicron	ψ	Ψ	Psi
θ	Θ	Theta	π	Π	Pi	ω	Ω	Omega

2 Notation

General notation

n	number of observations in a sample, or sample size
$x_1, x_2, \ldots, x_n$	data values in a sample
$\sum$	summation sign
$\overline{x}$	sample mean
m	sample median
s, s^2	sample standard deviation, sample variance
$f(x)$	probability density function of X
$p(x)$	probability mass function of X
p.d.f.	probability density function
p.m.f.	probability mass function
$E(X)$, μ	expectation or mean of X
$V(X)$, σ^2	variance of X
q_α	α-quantile
$\simeq$	is approximately equal to
$\sim$	is distributed as
$\approx$	has approximately the same distribution as
$N(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$M(\lambda)$	exponential distribution with parameter λ
$U(a, b)$	continuous uniform distribution on the interval $a \leq x \leq b$
$B(n, p)$	binomial distribution with parameters n and p
Poisson(μ)	Poisson distribution with parameter μ
$\widehat{\theta}$	estimate or estimator of a parameter θ
θ^-, θ^+	lower and upper confidence limits for θ
H_0, H_1	null and alternative hypotheses
p value	significance probability
$\mathrm{Cov}(x, y)$	sample covariance of observations on X and Y
$P(Y = y \mid X = x)$	conditional probability that $Y = y$ given that $X = x$
r	Pearson correlation coefficient

Medical statistics

$P(D\|E)$	probability of disease D, given exposure E
$P(D\|\text{not } E)$	probability of disease D, given no exposure E
RR	relative risk
OR	odds ratio
a, b, c, d	entries in a 2×2 table for a cohort or case-control study
n_1, n_2	numbers exposed and not exposed in a cohort study
m_1, m_2	numbers with and without disease in a case-control study
OR_i	odds ratio for exposure category i relative to the reference exposure category, or odds ratio for stratum i, or odds ratio for dose level i relative to the lowest dose
O_i	observed value for the ith cell of a contingency table
E_i	expected value for the ith cell of a contingency table
$\chi^2(\nu)$	chi-squared distribution on ν degrees of freedom
χ^2	test statistic for the chi-squared test for no association and McNemar's test
$\widehat{OR}_{MH}$	Mantel–Haenszel estimate of the common odds ratio
f, g	numbers of discordant pairs in a 1–1 matched case-control study
RCT	randomized controlled trial
α	significance level (for sample size calculation)
γ	power (for sample size calculation)
π_T, π_C	design values for the treatment group (T) and control group (C) (for sample size calculation)

Time series

X_t	time series, or the random variable representing the value at time t in a time series
x_t	observed time series, or the observed value at time t
T	period of a cyclic time series
m_t	trend component of a time series, or the level at time t
s_t	seasonal component of a time series
$s_1, \ldots, s_T$	seasonal factors
W_t	irregular (or random) component of a time series
$MA(t)$	moving average centred on t (for smoothing)
$SA(t)$	weighted moving average used for removing the seasonal component of a seasonal time series
F_j	raw seasonal factor for season j
$\widehat{x}_{n+1}$	1-step ahead forecast of X_{n+1}
α, γ, δ	smoothing parameters for exponential smoothing
e_t	1-step ahead forecast error
SSE	sum of squared errors
r_k	sample autocorrelation at lag k
ρ_k	autocorrelation at lag k
ACF	autocorrelation function
α_k	partial autocorrelation at lag k
PACF	partial autocorrelation function
Z_t	white noise
AR(p)	autoregressive model of order p
MA(q)	moving average model of order q
ARMA(p, q)	autoregressive moving average model of order (p, q)
ARIMA(p, d, q)	integrated autoregressive moving average model of order (p, d, q)
d	order of differencing

Multivariate analysis

p	dimension of a multivariate data set (number of variables)
n	number of observations in a multivariate data set
$\mathbf{X}$	data matrix, with n rows and p columns
X_j	jth column of a data matrix, containing values of the jth variable
x_{ij}	value of X_j for observation i; (i, j)th element of $\mathbf{X}$
$\mathbf{y}$	vector with jth element y_j
$\overline{x}_j$ or $\overline{X}_j$	sample mean of X_j
$\overline{\mathbf{x}}$	mean vector of $X_1, \ldots, X_p$
s_j^2	sample variance of X_j
s_{jk}	sample covariance between X_j and X_k
$\mathbf{S}$	covariance matrix of $X_1, \ldots, X_p$
Z_j	standardized (or group-standardized) variable
$\mathrm{Corr}(X_j, X_k)$	correlation coefficient between X_j and X_k
Y_1, Y_k	first and kth principal components of a data set
α_j	loading of the first principal component, or of the first discriminant function, for the jth variable
α_{kj}	loading of the kth principal component, or of the kth discriminant function, for the jth variable
$\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_k$	loadings vectors
TV	total variance
PVE	percentage variance explained
CPVE	cumulative percentage variance explained
G	number of groups
n_g	number of observations in group g
N	total number of observations in all groups, $N = n_1 + \cdots + n_G$
$\overline{x}_g$	for grouped data: mean of X for group g
$\overline{\overline{x}}$, $\overline{\overline{X}}$	grand mean of X
s_g^2	for grouped data: sample variance of X for group g
V_w, $V_w(X_j)$	within-groups variance
V_b, $V_b(X_j)$	between-groups variance
$\mathrm{Cov}_w(X_i, X_j)$	within-groups covariance of X_i and X_j
$\mathrm{Cov}_b(X_i, X_j)$	between-groups covariance of X_i and X_j
$\mathbf{W}$	within-groups covariance matrix
$\mathbf{B}$	between-groups covariance matrix
D_1, D_k	first and kth discriminant functions
$\mathrm{Sep}(D)$	separation achieved by the linear combination D
a_j	loading for a discriminant function based on group-standardized variables
PSA_j	percentage separation achieved by the jth discriminant function
CPSA_j	cumulative percentage separation achieved by the first j discriminant functions
$l_1, \ldots, l_{G-1}$	cutpoints for an allocation rule involving G groups

Bayesian statistics

$P(A)$	probability of event A
$P(A\|B)$	probability of A given B
$f(\theta)$	prior density of θ
$L(\theta)$	likelihood of θ given data
$\theta\|\text{data}$	the parameter θ, conditional on data
$f(\theta\|\text{data})$	posterior density of θ
$N(a,b)$	normal prior with mean a and variance b
$\text{Beta}(a,b)$	beta prior with parameters a and b
$\text{Gamma}(a,b)$	gamma prior with parameters a and b
$U(a,b)$	uniform prior on the interval $a \leq x \leq b$
τ	precision σ^{-2}, the reciprocal of the variance
(L,U)	equal-tailed $100(1-\alpha)\%$ interval for a parameter θ as used to specify a prior density
(l,u)	$100(1-\alpha)\%$ credible interval for a parameter θ
HPD	highest posterior density
N	number of samples drawn in a simulation
MC	Monte Carlo
MCMC	Markov chain Monte Carlo

3 Table of discrete probability distributions

Name	**Notation**	**Typical use**	**Range**	**Probability mass function** $p(x)$	**Mean**	**Variance**
Binomial	$B(n,p)$	Total number of successes in n independent Bernoulli trials	$0,1,\ldots,n$	$\binom{n}{x} p^x(1-p)^{n-x}$	np	$np(1-p)$
Poisson	$\text{Poisson}(\mu)$	Counts of independently occurring events	$0,1,\ldots$	$\dfrac{\mu^x e^{-\mu}}{x!}$	μ	μ
Discrete uniform		Equally likely events labelled 1 to n	$1,\ldots,n$	$\dfrac{1}{n}$	$\dfrac{n+1}{2}$	$\dfrac{n^2-1}{12}$

4 Table of continuous probability distributions

Name	Notation	Typical use	Range	Probability density function $f(x)$	Location	Variance
Normal	$N(\mu, \sigma^2)$	Measurements clustered symmetrically around a mean	$-\infty < x < \infty$	$\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right)$	mean $= \mu$ median $= \mu$ mode $= \mu$	σ^2
Standard normal	$N(0,1)$	Calculation of z-intervals, sample size estimation	$-\infty < x < \infty$	$\dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{1}{2}x^2\right)$	mean $= 0$ median $= 0$ mode $= 0$	1
Continuous uniform	$U(a,b)$	Equally likely values on the interval $[a,b]$; flat priors	$a \le x \le b$	$\dfrac{1}{b-a}$	mean $= \dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Exponential	$M(\lambda)$	Time intervals between successive events	$x \ge 0$	$\lambda e^{-\lambda x}$	mean $= \dfrac{1}{\lambda}$ mode $= 0$	$\dfrac{1}{\lambda^2}$
Chi-squared	$\chi^2(\nu)$	Null distribution of tests for no association, no interaction, no linear association	$x > 0$	$c\,x^{\nu/2-1}\,e^{-x/2}$	mean $= \nu$	2ν
Beta	$\mathrm{Beta}(a,b)$	Prior for a probability or proportion	$0 \le x \le 1$	$c\,x^{a-1}(1-x)^{b-1}$ where c is a constant	mean $= \dfrac{a}{a+b}$ mode $= \begin{cases} \dfrac{a-1}{a+b-2} & \text{if } a > 1 \text{ and } b > 1 \\ 0 & \text{if } 0 < a < 1 \\ 1 & \text{if } 0 < b < 1 \end{cases}$	$\dfrac{ab}{(a+b)^2(a+b+1)}$
Gamma	$\mathrm{Gamma}(a,b)$	Prior for a non-negative parameter	$x \ge 0$	$c\,x^{a-1}e^{-bx}$ where c is a constant	mean $= \dfrac{a}{b}$ mode $= \begin{cases} \dfrac{a-1}{b} & \text{if } a > 1 \\ 0 & \text{if } 0 < a \le 1 \end{cases}$	$\dfrac{a}{b^2}$

5 Outlines

5.1 Background material from the Introduction to statistical modelling

Graphical and numerical summaries

1 Useful graphical representations of statistical data include bar charts, histograms and scatterplots. **Bar charts** are generally used with **categorical** data, or **discrete numerical** data. **Histograms** are generally used with **continuous** data, by grouping the data into intervals or **bins**. **Scatterplots** are used to display the relationship between two numerical variables.

2 **Measures of location** include the mean, median and mode. If the n items in a data set are denoted $x_1, x_2, \ldots, x_n$, then the **sample mean**, which is denoted $\overline{x}$, is given by

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

3 The **median** of a sample of data with an odd number of values is the middle value of the data set when the values are placed in order of increasing size. If the sample size is even, then the median is halfway between the two middle values.

4 The **mode** of a set of categorical data is the most frequently occurring (or modal) value. The term mode is also used to describe a clear peak in a histogram or a bar chart of a set of numerical data.

5 **Measures of dispersion** describe the variation within a sample around its average value. They include the standard deviation and the variance. If the n items in a data set with sample mean $\overline{x}$ are denoted $x_1, x_2, \ldots, x_n$, then the **sample standard deviation**, denoted s, is given by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

The quantity s^2 is known as the **sample variance**.

6 The **skewness** of a sample is a measure of departure from symmetry. If the data are symmetrically distributed around the median, then the skewness is zero. If there is a relatively long tail of values to the right of the median, then the data are said to be **right-skew**, or **positively skewed**. If there is a relatively long tail of values to the left of the median, then the data are said to be **left-skew**, or **negatively skewed**.

Probability models

7 A probability model for a continuous random variable X is specified by the **probability density function** (p.d.f.) $f(x)$ of the random variable. A probability model for a discrete random variable X is specified by the **probability mass function** (p.m.f.) $p(x)$ of the random variable. Details of specific p.d.f.s and p.m.f.s are given in the tables in Sections 3 and 4 of this Handbook.

8 The **population mean** of a random variable X is denoted μ or $E(X)$; it is also called the **expectation** or **expected value** of X. The **population variance** of X is denoted σ^2 or $V(X)$; it is equal to $E(X-\mu)^2$. The **population standard deviation** is σ.

9 The **α-quantile** of a continuous random variable X is the value q_α such that

$$\alpha = P(X \leq q_\alpha).$$

The **population median** of X is the 0.5-quantile. The **lower quartile** of X is the 0.25-quantile, and the **upper quartile** of X is the 0.75-quantile.

10 The **central limit theorem** states that if n independent random observations are taken from a population with mean μ and variance σ^2, then for large n the distribution of their mean $\widehat{\mu}$ (also called the **sampling distribution** of the mean) is approximately normal with mean μ and variance σ^2/n. The standard deviation of the sampling distribution, which is equal to $\sigma/\sqrt{n}$, is called the **standard error** of $\widehat{\mu}$.

Confidence intervals

11 A $100(1-\alpha)\%$ **confidence interval** (μ^-, μ^+) for a population mean μ, calculated from a sample of size n with sample mean $\overline{x}$, may be used to represent the uncertainty in the estimate $\overline{x}$ of μ. The confidence interval may be interpreted in two ways — using the **repeated experiments** interpretation (based on a large number of repetitions of the experiment with samples of size n), and using the **plausible range** interpretation (based on the probability of observing a sample mean as extreme as $\overline{x}$, if μ were to take values outside the confidence interval). These interpretations are equivalent.

12 Given a random sample of size n from a population with mean μ, an approximate **$100(1-\alpha)\%$ confidence interval for μ** is given by the **z-interval**

$$(\mu^-, \mu^+) = \left(\widehat{\mu} - z\frac{s}{\sqrt{n}},\ \widehat{\mu} + z\frac{s}{\sqrt{n}}\right),$$

where $\widehat{\mu}$ is the sample mean, s is the sample standard deviation, and z is $q_{1-\alpha/2}$, the $(1-\alpha/2)$-quantile of the standard normal distribution.

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

13 An approximate **$100(1-\alpha)\%$ confidence interval for a parameter θ** is given by the z-interval

$$(\theta^-, \theta^+) = (\widehat{\theta} - z\widehat{\sigma},\ \widehat{\theta} + z\widehat{\sigma}),$$

where $\widehat{\theta}$ is the sample estimate of θ, $\widehat{\sigma}$ is the estimated standard error of the estimator $\widehat{\theta}$, and z is $q_{1-\alpha/2}$, the $(1-\alpha/2)$-quantile of the standard normal distribution.

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

14 When θ is a binomial proportion p, $\widehat{\theta}$ is its sample estimate $\widehat{p}$ and the standard error of $\widehat{p}$ may be estimated by

$$\widehat{\sigma} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Significance tests

15 A **significance test** may be used to evaluate the strength of evidence against a **null hypothesis** H_0 of the form

$$H_0\colon \theta = \theta_0.$$

The corresponding **alternative hypothesis** H_1 is

$$H_1\colon \theta \neq \theta_0.$$

16 The strength of evidence against H_0 is quantified by the **significance probability** or ***p* value**. The procedure for carrying out a significance test is as follows.

- ◇ Determine the null hypothesis H_0 and the alternative hypothesis H_1.
- ◇ Choose a suitable test statistic and determine the null distribution of the test statistic.
- ◇ Calculate the observed value of the test statistic and identify the values that are at least as extreme as the observed value in relation to H_0.
- ◇ Calculate the significance probability p.
- ◇ Interpret the significance probability and report the results.

17 The following table provides a rough guide for interpreting p values.

Significance probability p	Rough interpretation
$p > 0.10$	little evidence against H_0
$0.10 \geq p > 0.05$	weak evidence against H_0
$0.05 \geq p > 0.01$	moderate evidence against H_0
$p \leq 0.01$	strong evidence against H_0

Correlation and association

18 Two random variables are said to be **related**, or **associated**, if knowing something about the value of one variable tells you something about the value of the other.

19 A measure of the strength of a linear association is provided by the (Pearson) correlation coefficient. This is based on the sample covariance. For observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) on two random variables X and Y with sample means $\overline{x}$ and $\overline{y}$ and sample standard deviations s_x and s_y, the **sample covariance** is

$$\mathrm{Cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}),$$

and the **correlation coefficient** is

$$r = \frac{\mathrm{Cov}(x, y)}{s_x s_y}.$$

20 **Conditional probabilities** are probabilities of the form 'probability that $Y = y$, given that $X = x$', and are written

$$P(Y = y | X = x).$$

21 Two discrete random variables X and Y are **independent** if, for all values of x and y,

$$P(Y = y | X = x) = P(Y = y).$$

If X and Y are not independent, they are said to be **dependent**, or **related**, or **associated**.

5.2 Medical statistics

Cohort and case-control studies

1 A **cohort study** of the association between an exposure E and a disease D typically includes one group with exposure E (the **exposed group**) and one group without exposure E (the **control group**). The groups are followed over time and the occurrences of disease D in each group are identified.

2 A **case-control study** of the association between an exposure E and a disease D typically includes a group of **cases** with the disease D and a group of **controls** without the disease D, who are otherwise comparable to the cases. The past exposures of the cases and controls are determined and the occurrences of exposure E are identified.

3 The **risk** of disease D, given exposure E, is $P(D|E)$. The **relative risk** is

$$RR = \frac{P(D|E)}{P(D|\text{not } E)}.$$

4 The **odds** of disease D, given exposure E, is

$$OD(D|E) = \frac{P(D|E)}{P(\text{not } D|E)}.$$

The **odds ratio** is

$$OR = \frac{P(D|E) \times P(\text{not } D|\text{not } E)}{P(\text{not } D|E) \times P(D|\text{not } E)}.$$

5 Data from a cohort study may be presented in a table as follows.

	Disease outcome		
Exposure category	D	not D	Total
E	a	b	n_1
not E	c	d	n_2

The sample estimate of the relative risk RR from a cohort study is

$$\widehat{RR} = \frac{a/n_1}{c/n_2}.$$

An approximate $100(1-\alpha)\%$ confidence interval for the relative risk RR is

$$(RR^-, RR^+) = (\widehat{RR} \times \exp(-z\widehat{\sigma}),\ \widehat{RR} \times \exp(z\widehat{\sigma})),$$

where z is the $(1-\alpha/2)$-quantile of the standard normal distribution, and

$$\widehat{\sigma} = \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}}.$$

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

6 Data from a case-control study may be presented in a table as follows.

	Disease outcome	
Exposure category	D (cases)	not D (controls)
E	a	b
not E	c	d
Total	m_1	m_2

The sample estimate of the odds ratio OR from a case-control study or a cohort study is

$$\widehat{OR} = \frac{a \times d}{b \times c}.$$

An approximate $100(1-\alpha)\%$ confidence interval for the odds ratio OR is

$$(OR^-, OR^+) = (\widehat{OR} \times \exp(-z\widehat{\sigma}),\ \widehat{OR} \times \exp(z\widehat{\sigma})),$$

where z is the $(1-\alpha/2)$-quantile of the standard normal distribution, and

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

$$\widehat{\sigma} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

7 In studies with more than one exposure category, one category is chosen as the **reference** exposure category and calculations are undertaken relative to this reference category.

8 When data are arranged in an $r \times c$ table, an approximate **test for no association** between the variables uses the **chi-squared test statistic**

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

where the sum is taken over all $r \times c$ cells of the table, O_i is the observed frequency for the ith cell, and E_i is the expected frequency for the ith cell. The expected frequency for a cell is given by

$$\text{expected frequency} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}.$$

When the null hypothesis of no association is true,

$$\chi^2 \approx \chi^2((r-1)(c-1)).$$

Table 3 of the statistical tables contains quantiles for chi-squared distributions.

The approximation is adequate provided that all the expected frequencies are at least 5. When this is not the case, **Fisher's exact test** can be used.

Bias, confounding and causation

9 A study is **biased** if some aspects of the design, sampling, data collection or analysis method produce results that systematically overestimate or underestimate the strength of association. In particular, bias may arise from **selection bias**, **information bias** or **confounding**.

10 Confounding may arise if both the exposure E and the disease D are associated with a third variable C, known as a **confounder**. Confounding bias may be explored by **stratifying** the data according to the **levels** of the confounder.

11 Data from stratum i of a cohort study or case-control study stratified according to the levels of a variable C may be presented in a table as follows.

Exposure category	Disease/Cases	No disease/Controls
Exposed	a_i	b_i
Not exposed	c_i	d_i

If the underlying stratum-specific odds ratios are the same for all strata, then their common value OR is estimated by the **Mantel–Haenszel odds ratio**:

$$\widehat{OR}_{MH} = \frac{\sum a_i d_i / N_i}{\sum b_i c_i / N_i},$$

where $N_i = a_i + b_i + c_i + d_i$, and the summations are over all the strata.

12 In a **matched case-control study**, the controls in each **matched case-control set** are selected so that they match the case with respect to the confounding variables.

13 The case-control pairs from a **1–1 matched case-control study** may be presented in a table as follows.

		Controls: Exposed	Controls: Not exposed
Cases	Exposed	e	f
	Not exposed	g	h

The Mantel–Haenszel estimate of the odds ratio is

$$\widehat{OR}_{MH} = \frac{f}{g}.$$

An approximate $100(1-\alpha)\%$ confidence interval for the odds ratio is

$$(OR^-, OR^+) = (\widehat{OR}_{MH} \times \exp(-z\widehat{\sigma}),\ \widehat{OR} \times \exp(z\widehat{\sigma})),$$

where z is the $(1-\alpha/2)$-quantile of the standard normal distribution, and

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

$$\widehat{\sigma} = \sqrt{\frac{1}{f} + \frac{1}{g}}.$$

14 **McNemar's test** for no association in a 1–1 matched case-control study is based on the test statistic

$$\chi^2 = \frac{(|f-g|-1)^2}{f+g}.$$

Under the null hypothesis of no association, $\chi^2 \approx \chi^2(1)$.

Table 3 of the statistical tables contains quantiles for $\chi^2(1)$.

15 The presence of an interaction between a stratifying variable C and the association between an exposure E and a disease outcome D may be investigated using a significance **test of homogeneity**.

If there are k strata, the null hypothesis is $OR_1 = OR_2 = \cdots = OR_k$, where OR_i is the odds ratio for stratum i. **Tarone's test for homogeneity** is based on a test statistic whose distribution is approximately $\chi^2(k-1)$ under the null hypothesis.

Table 3 of the statistical tables contains quantiles for chi-squared distributions.

16 Association does not imply not causation. **Bradford Hill's criteria** for causation may help in assessing whether an association is causal.

17 A **dose** is a quantified exposure. A **dose-response relationship** exists between an exposure E and a disease D if the risk (or odds) of disease varies according to the dose of that exposure.

18 The presence of a dose-response relationship may be investigated using the **chi-squared test for no linear trend**. The null hypothesis for this significance test is that the log odds of disease does not increase or decrease linearly with the dose. Under the null hypothesis, the distribution of the test statistic is approximately $\chi^2(1)$.

Table 3 of the statistical tables contains quantiles for $\chi^2(1)$.

Randomized controlled trials and the medical literature

19 A **randomized controlled trial** is a cohort study in which participants are randomly allocated to treatment and control groups. **Stratified randomization**, in which participants are **randomized by blocks**, may be used to improve **balance** in the characteristics of the patients allocated to the different groups. Bias is further reduced by using **concealment** procedures such as **double blinding** or **single blinding**.

20 The **flow chart** of the trial documents the numbers of participants included and excluded at each stage of the trial. The recommended method of analysis of randomized controlled trials is by **intention to treat**. In an intention-to-treat analysis, the groups analysed are as close as possible to those randomized. An alternative method of analysis is **per protocol**. In a per-protocol analysis, only participants who complete the treatment to which they were randomized are included.

21 Pharmaceutical drugs are evaluated in **clinical trials**. The evaluation progresses through four phases. **Phase III studies** are always randomized controlled trials. An independent **Data Monitoring Committee** reviews the data and can halt a trial on ethical grounds.

22 The **sample size** required for a randomized controlled trial to compare the effect of treatment on a disease D is derived within the framework of **fixed-level testing**. The null and alternative hypotheses may be written as

$$H_0 : p_T = p_C, \quad H_1 : p_T \neq p_C,$$

where p_T is the probability of disease in the treatment group, and p_C is the probability of disease in the control group.

23 A **Type I error** is said to occur if the null hypothesis H_0 is rejected when it is true. A **Type II error** is said to occur if the null hypothesis H_0 is not rejected when it is false.

The **significance level** of the test, α, is the probability of a Type I error. The **power** of the test, γ, is the probability of avoiding a Type II error.

24 To calculate the sample size for a trial with two groups of equal size, the **design values** π_T and π_C, the significance level α and the power γ must be specified. The sample size n for each trial group is given approximately by

$$n = \frac{2(q_{1-\alpha/2} + q_\gamma)^2 \pi_0 (1 - \pi_0)}{(\pi_T - \pi_C)^2},$$

where $q_{1-\alpha/2}$ and q_γ denote, respectively, the $(1 - \alpha/2)$-quantile and the γ-quantile of the standard normal distribution, and $\pi_0 = (\pi_T + \pi_C)/2$.

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

25 The power γ available in a trial with two groups each of size n is obtained from q_γ, the γ-quantile of the standard normal distribution, which is given by the expression

$$q_\gamma = |\pi_T - \pi_C| \sqrt{\frac{n}{2\pi_0(1 - \pi_0)}} - q_{1-\alpha/2}.$$

The notation in this expression is the same as that used in **24**.

26 Evidence from all available studies, or all available studies of a particular type, may be reviewed together as part of a **systematic review**. The selection of studies in such a review is particularly important in order to avoid **publication bias**. Sometimes a quantitative assessment of the strength of evidence from several studies may be possible by combining their results in a **meta-analysis**.

27 In a meta-analysis, the results of several studies are combined to obtain a **pooled odds ratio** and confidence interval, for example using the **Mantel–Haenszel odds ratio** (see **11**). The presence of heterogeneity between studies may be investigated using **Tarone's test for homogeneity** (see **15**). A **forest plot** is used to display the results of a meta-analysis.

28 Medical papers often contain statistical analyses. A typical medical paper includes the following sections: **Abstract**, **Introduction**, **Methods**, **Results**, **Discussion**.

5.3 Time series

Decomposition models

1 A **time series** is a collection of observations X_t on some random variable X at equally-spaced times $1, 2, \dots, t, t+1, \dots$. A **time plot** is a graph of the observed values x_t against t.

2 A **cycle** is a regular pattern that repeats at fixed intervals. The time interval between cycles is the **period**. A cycle whose period is determined by the natural clock is **seasonal**. A seasonal cycle with period one year is **annual**. Seasonality may be displayed using a **seasonal plot**.

3 The **additive decomposition model** for a time series X_t is

$$X_t = m_t + s_t + W_t, \quad t = 1, 2, \dots,$$

where m_t is the **trend component**, s_t is the **seasonal component** of period T, and W_t is the **irregular** (or random) **component**, sometimes also described as **noise**. The seasonal component satisfies

$$s_t = s_{t+T} \quad \text{for all } t,$$
$$s_1 + \dots + s_T = 0.$$

The distinct values $s_1, \dots, s_T$ are the **seasonal factors**.

The irregular component W_t has mean 0 and variance σ^2:

$$E(W_t) = 0, \quad V(W_t) = \sigma^2.$$

4 The **multiplicative decomposition model** for a time series X_t which takes only positive values is

$$X_t = m_t \times s_t \times W_t.$$

The seasonal component s_t satisfies

$$s_t = s_{t+T} \quad \text{for all } t,$$
$$s_1 \times s_2 \times \dots \times s_T = 1.$$

5 The **simple moving average** of order $2q+1$ centred on t is given by the transformation

$$MA(t) = \frac{1}{2q+1}(X_{t-q} + \dots + X_t + \dots + X_{t+q}).$$

6 A **weighted moving average** of order $2q+1$ has the form

$$MA(t) = a_{-q}X_{t-q} + \dots + a_{-1}X_{t-1} + a_0X_t + a_1X_{t+1} + \dots + a_qX_{t+q},$$

where the **weights** a_j, $j = -q, -q+1, \dots, q$, add up to 1.

7 Simple and weighted moving averages may be used for **smoothing** a time series. The **order** of the moving average should be chosen so as to avoid both **over-smoothing** and **under-smoothing** the time series.

8 For a seasonal time series X_t, which may be described by an additive model, and for which the seasonal period is T (an even number), the seasonal component s_t may be estimated as follows.

First, the series is smoothed using a suitable weighted moving average $SA(t)$. Then the series of differences $y_t = x_t - SA(t)$ is obtained, and the **raw seasonal factors** F_j, $j = 1, \ldots, T$, are calculated by averaging the values y_t for each season. Finally, the seasonal factors are estimated by

$$\widehat{s}_j = F_j - \overline{F}, \quad j = 1, \ldots, T,$$

where $\overline{F}$ is the average of the raw seasonal factors.

9 A time series is **seasonally adjusted** if its seasonal component has been estimated and removed, leaving only a trend component and an irregular component.

Forecasting

10 **Forecasting** is the process of predicting future values of a time series based on the past values of the time series. A forecast $\widehat{x}_{n+1}$ of X_{n+1} based on $x_n, x_{n-1}, x_{n-2}, \ldots$ is called a **1-step ahead forecast** of X_{n+1}.

11 If a time series X_t is described by an additive model with constant level and no seasonality, then 1-step ahead forecasts may be obtained by **simple exponential smoothing** using the formula

$$\widehat{x}_{n+1} = \alpha x_n + (1 - \alpha)\widehat{x}_n,$$

where x_n is the observed value at time n, $\widehat{x}_n$ and $\widehat{x}_{n+1}$ are the 1-step ahead forecasts of X_n and X_{n+1}, and α is a **smoothing parameter**, $0 \leq \alpha \leq 1$. The method requires an **initial value** $\widehat{x}_1$.

12 The **1-step ahead forecast error** is the difference between the observed value and the 1-step ahead forecast of X_t: $e_t = x_t - \widehat{x}_t$. The **sum of squared errors**, or SSE, is given by

$$SSE = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (x_t - \widehat{x}_t)^2 .$$

13 If a time series X_t is described by an additive model with a linear trend component and no seasonality, then 1-step ahead forecasts may be obtained by **Holt's exponential smoothing**. There are two smoothing parameters: α for the level and γ for the slope.

If in addition the time series has a seasonal component, forecasts may be obtained using **Holt–Winters exponential smoothing**. There are three smoothing parameters: α for the level, γ for the slope and δ for the seasonal component.

For all exponential smoothing methods, **optimal values** of the smoothing parameters are obtained by minimizing the SSE.

14 Suppose that X_t is a time series with n observed values $x_1, x_2, \ldots, x_n$. The time series **lagged by *k* places** is the time series with X_{t-k} in position k. The first k positions of the lagged series comprise missing values.

15 The **sample autocorrelation at lag *k*** is a correlation coefficient r_k calculated between a time series and a copy of itself, lagged by k places. It is calculated using the $n - k$ pairs of points (x_1, x_{k+1}), (x_2, x_{k+2}), $\ldots$, (x_{n-k}, x_n).

16 The population autocorrelations ρ_k, $k = 1, 2, \ldots$, define the **autocorrelation function**, or **ACF**. Under the null hypothesis $\rho_k = 0$, the distribution of the sample autocorrelation calculated from a time series with n time points is approximately normal with mean 0 and variance $1/n$.

17 The sample autocorrelations may be displayed on a **correlogram** or **sample ACF plot**. **Significance bounds** are horizontal lines plotted at positions $\pm 1.96/\sqrt{n}$ on the correlogram.

18 For a fixed number k of lags, the null hypothesis

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_k = 0$$

may be tested using a **portmanteau test** such as the **Ljung–Box test**.

19 A **$100(1-\alpha)\%$ prediction interval** for X_{n+1}, given observed values up to and including x_n, is an interval with probability $1 - \alpha$ of containing X_{n+1}.

20 Suppose that a 1-step ahead forecast $\widehat{x}_{n+1}$ for X_{n+1} has been obtained, together with the SSE, the sum of squared forecast errors at times $1, 2, \ldots, n$. An approximate $100(1 - \alpha)\%$ prediction interval for X_{n+1} is given by

$$\left(\widehat{x}_{n+1} - z\sqrt{\frac{SSE}{n}},\ \widehat{x}_{n+1} + z\sqrt{\frac{SSE}{n}}\right),$$

where z is the $(1 - \alpha/2)$-quantile of the standard normal distribution. The assumptions required are that the forecast errors are normally distributed with mean zero and constant variance, and that the autocorrelations between the forecast errors are zero at lags $k \geq 1$.

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

21 A time series Z_t is said to be **white noise** if Z_t is normally distributed with mean zero and constant variance σ^2, and the autocorrelations at all lags $k \geq 1$ are zero.

ARIMA models

22 A time series X_t is **stationary in mean** if it has constant mean, $E(X_t) = \mu$. It is **stationary in variance** if it has constant variance, $V(X_t) = \sigma^2$. It is **stationary in correlation** if for all k, ρ_k, the autocorrelation between X_t and X_{t-k}, depends only on the lag k. The time series is **stationary** if it is stationary in mean, in variance and in correlation.

23 The **partial autocorrelation** at lag k, α_k, is a measure of the direct dependence between X_t and X_{t-k}. The partial autocorrelations α_k, $k = 0, 1, 2, \ldots$, define the **partial autocorrelation function**, or **PACF**. The **partial correlogram**, or **sample PACF plot**, is a bar chart of the sample PACF.

24 Let X_t be a stationary time series with mean μ. The **autoregressive model of order p**, or **AR(p)** model, has the form

$$X_t - \mu = \beta_1(X_{t-1} - \mu) + \beta_2(X_{t-2} - \mu) + \cdots + \beta_p(X_{t-p} - \mu) + Z_t,$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are parameters to be estimated, and Z_t is white noise with mean 0 and variance σ^2.

25 The ACF for an AR(1) model is given by $\rho_k = \beta_1^k$ for $k \geq 0$. The ACF for an AR(p) model tails off to zero in magnitude, either exponentially or in a damped sinusoidal pattern, as the lag increases.

The PACF for an AR(p) model satisfies $\alpha_p = \beta_p$, and $\alpha_k = 0$ for lags $k > p$.

26 Let X_t be a stationary time series with mean μ. The **moving average model of order q**, or **MA(q)** model, has the form

$$X_t - \mu = Z_t - \theta_1 Z_{t-1} - \cdots - \theta_q Z_{t-q},$$

where $\theta_1, \theta_2, \ldots, \theta_q$ are parameters to be estimated, and Z_t is white noise with mean 0 and variance σ^2.

27 The ACF for an MA(q) model satisfies

$$\rho_q = \frac{-\theta_q}{1+\theta_1^2+\cdots+\theta_q^2},$$

and $\rho_k = 0$ for $k > q$.

The PACF for an MA(q) model tails off to zero in magnitude, either exponentially or in a damped sinusoidal pattern, as the lag increases.

28 Let X_t be a stationary time series with mean zero. The **autoregressive moving average model of order (p, q)**, or **ARMA(p, q)** model, has the form

$$X_t - \mu = \beta_1(X_{t-1} - \mu) + \cdots + \beta_p(X_{t-p} - \mu) + Z_t - \theta_1 Z_{t-1} - \cdots - \theta_q Z_{t-q}.$$

An **integrated moving average model of order (p, d, q)**, or **ARIMA(p, d, q)** model, is an ARMA(p, q) model applied to a time series after differencing of order d.

29 The key features of ARMA models are summarized in the table below.

Model	Notation	ACF	PACF
White noise	ARMA$(0, 0)$	Zero at lags > 0	Zero at lags > 0
Autoregressive	ARMA$(p, 0)$	Tails off to zero	Zero after lag p
Moving average	ARMA$(0, q)$	Zero after lag q	Tails off to zero
Mixed	ARMA(p, q)	Tails off to zero	Tails off to zero

30 The **principle of parsimony** in selecting an ARIMA model is to keep the value of $p + q$ to a minimum.

31 The steps involved in selecting an ARIMA model for a non-seasonal time series are as follows.

- ◇ Check than an additive model is appropriate. If it is not appropriate, then transform the series to obtain a series that can be represented by an additive model.
- ◇ Identify the order of differencing, d, required to obtain stationarity.
- ◇ Identify those ARIMA(p, d, q) models that are consistent with the correlogram and partial correlogram for the stationary series.
- ◇ Choose the model(s) with the lowest value of $p + q$.

32 After fitting an ARIMA model, its adequacy should be checked, as follows.

- ◇ Check the fit of the model by plotting the time series and the 1-step ahead forecasts on a multiple time plot.
- ◇ Verify that the distribution of the forecast errors is approximately normal with mean zero and constant variance.
- ◇ Use the correlogram for the forecast errors and the Ljung–Box test (see **18**) to check that the forecast errors are uncorrelated.

5.4 Multivariate analysis

Describing and displaying multivariate data

1 A multivariate data set comprises observations on two or more random variables. A **bivariate** data set has two variables. The number of variables, p, is the **dimension** of the data set. An **observation** is the set of p measurements made on one sampled unit. The variables $X_1, \ldots, X_p$ form the columns of the $n \times p$ **data matrix X**, where n is the number of observations.

2 Multivariate data may be displayed using **two-dimensional scatterplots**, **three-dimensional scatterplots**, **matrix scatterplots** and **profile plots**.

3 The **mean vector** for a data set with n observations and p variables $X_1, \ldots, X_p$ is $\mathbf{\overline{x}} = (\overline{x}_1, \ldots, \overline{x}_p)$, where $\overline{x}_j$ is the sample mean of X_j,

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}.$$

4 The **sample covariance** between X_j and X_k is

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k).$$

The covariance between a variable X_j and itself is the **sample variance** of X_j, that is, $s_{jj} = s_j^2$.

5 The **variance-covariance matrix**, or **covariance matrix**, of $X_1, \ldots, X_p$ is a square matrix $\mathbf{S}$ with p rows and p columns. Element (j, k) of $\mathbf{S}$ is s_{jk}, the sample covariance between X_j and X_k. The diagonal element (j, j) is s_j^2, the sample variance of X_j.

6 In **standardization**, each variable X_j is transformed separately in such a way that the transformed variable Z_j has mean 0 and variance 1. For observation i, the value x_{ij} of X_j is transformed to obtain the value z_{ij} of Z_j, as follows:

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j},$$

where $\overline{x}_j$ is the sample mean and s_j is the sample standard deviation of X_j.

The numbers z_{ij} do not have any units associated with them, so the standardized variable Z_j is **scale-free**.

7 The **correlation matrix** of $X_1, \ldots, X_p$ is the covariance matrix of the standardized variables $Z_1, \ldots, Z_p$. Element (j, k) is the correlation coefficient between X_j and X_k, denoted $\mathrm{Corr}(X_j, X_k)$. The diagonal elements of the correlation matrix are all equal to 1.

Reducing dimension

8 Two approximations Y_1 and Y_2 to a multivariate data set are **equivalent** if constants $c_1 \neq 0$ and c_2 can be found such that $Y_2 = c_1 Y_1 + c_2$.

9 For a data set of dimension p with variables $X_1, \ldots, X_p$, the **(first) principal component** of the data, denoted Y, is the linear combination

$$Y = \sum_{j=1}^{p} \alpha_j (X_j - \overline{X}_j),$$

where the **loadings vector** $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ is chosen so that the variance of Y is maximized, subject to the constraint

$$\sum_{j=1}^{p} \alpha_j^2 = 1.$$

10 For a data set with p variables $X_1, \ldots, X_p$, the variance of the linear combination

$$Y = \sum_{j=1}^{p} \alpha_j (X_j - \overline{X}_j)$$

can be calculated from the variances and covariances of the original variables using the formula

$$V(Y) = \sum_{j=1}^{p} \alpha_j^2 \, V(X_j) + 2 \sum_{j,k:k>j} \alpha_j \alpha_k \,\mathrm{Cov}(X_j, X_k).$$

11 For a multivariate data set with p variables $X_1, \dots, X_p$, the **total variance**, **TV**, is

$$\text{TV} = \sum_{j=1}^{p} V(X_j).$$

The **percentage variance explained**, **PVE**, by a linear combination Y is

$$\text{PVE} = \frac{V(Y)}{\text{TV}} \times 100\%.$$

12 For a data set of dimension p with variables $X_1, \dots, X_p$, the ***k*th principal component** of the data, denoted Y_k, is the linear combination

$$Y_k = \sum_{j=1}^{p} \alpha_{kj}(X_j - \overline{X}_j),$$

where the **loadings vector** $\boldsymbol{\alpha}_k = (\alpha_{k1}, \dots, \alpha_{kp})$ is chosen so that the variance of Y_k is maximized, subject to the following constraints:

$$\sum_{j=1}^{p} \alpha_{kj}^2 = 1,$$

Y_k is uncorrelated with $Y_1, \dots, Y_{k-1}$.

13 In some circumstances, it is preferable, or even essential, to calculate principal components using standardized data. In this case, the kth principal component has the form

$$Y_k = \sum_{j=1}^{p} \alpha_{kj} Z_j.$$

14 The **cumulative percentage variance explained**, **CPVE**, by the first k principal components is given by

$$\text{CPVE} = \frac{V(Y_1) + \cdots + V(Y_k)}{\text{TV}} \times 100\%.$$

15 **Kaiser's criterion** for choosing the number of principal components is to retain components with variance greater than the average of the variances of the original variables.

In a **scree plot**, the **elbow** is the point at which the plot flattens out. The point preceding the elbow indicates the last component to be retained.

Discrimination

16 Suppose that a multivariate data set comprises observations on G groups, that n_g is the size of group g, and that $\overline{x}_g$ is the mean of a variable X in group g. Let N denote the total number of observations in the G groups: $N = n_1 + \cdots + n_G$. The **grand mean** of X is denoted $\overline{\overline{x}}$ and is given by

$$\overline{\overline{x}} = \frac{1}{N} \sum_{g=1}^{G} n_g \overline{x}_g.$$

17 Suppose that the variance of X in group g is s_g^2. The **between-groups variance** of X, denoted V_b, and the **within-groups variance** of X, denoted V_w, are given by

$$V_b = \frac{1}{N-G}\sum_{g=1}^{G} n_g(\overline{x}_g - \overline{\overline{x}})^2,$$

$$V_w = \frac{1}{N-G}\sum_{g=1}^{G}(n_g - 1)s_g^2.$$

The **separation** achieved by a variable X is given by the ratio of the between-groups variance to the within-groups variance of X:

$$\text{separation} = \frac{V_b}{V_w}.$$

18 The **within-groups covariance** for a pair of variables X_i and X_j, which is denoted $\mathrm{Cov}_w(X_i, X_j)$, is the weighted average of the covariances of X_i and X_j calculated for each of the groups separately. The **between-groups covariance** of variables X_i and X_j, which is denoted $\mathrm{Cov}_b(X_i, X_j)$, is the covariance between the group means for X_i and X_j. The **within-groups covariance matrix W** has (i, j)th element $\mathrm{Cov}_w(X_i, X_j)$. The **between-groups covariance matrix B** has (i, j)th element $\mathrm{Cov}_b(X_i, X_j)$.

19 For a linear combination D of variables of the form

$$D = \sum_{j=1}^{p} \alpha_j (X_j - \overline{\overline{X}}_j),$$

the between-groups covariance of D, denoted $V_b(D)$, and the within-groups variance of D, denoted $V_w(D)$, are given by

$$V_b(D) = \sum_{j=1}^{p} \alpha_j^2\, V_b(X_j) + 2 \sum_{j,k:k>j} \alpha_j\alpha_k\, \mathrm{Cov}_b(X_j, X_k),$$

$$V_w(D) = \sum_{j=1}^{p} \alpha_j^2\, V_w(X_j) + 2 \sum_{j,k:k>j} \alpha_j\alpha_k\, \mathrm{Cov}_w(X_j, X_k).$$

The **separation** achieved by D, denoted $\mathrm{Sep}(D)$, is the ratio of the between-groups variance of D to the within-groups variance of D:

$$\mathrm{Sep}(D) = \frac{V_b(D)}{V_w(D)}.$$

20 In **canonical discrimination**, the **(first) discriminant function** D is the linear combination

$$D = \sum_{j=1}^{p} \alpha_j (X_j - \overline{\overline{X}}_j)$$

for which the separation is maximized, subject to a constraint on the loadings $\alpha_1, \ldots, \alpha_p$. Commonly used constraints are

$$\sum_{j=1}^{p} \alpha_j^2 = 1$$

and

$$V_w(D) = 1.$$

21 In canonical discrimination, the standardized version Z_j of a variable X_j is defined so that Z_j has mean 0 and within-groups variance 1, using the formula

$$Z_j = \frac{X_j - \overline{\overline{X}}_j}{\sqrt{V_w(X_j)}}.$$

The variable Z_j is called the **group-standardized** variable.

22 The discriminant function

$$D = \sum_{j=1}^{p} \alpha_j (X_j - \overline{\overline{X}}_j)$$

may be written in terms of the group-standardized variables as follows:

$$D = \sum_{j=1}^{p} a_j Z_j,$$

where the loadings a_j are given by

$$a_j = \alpha_j \sqrt{V_w(X_j)}.$$

The separation achieved by the discriminant function D is the same whether D is based on unstandardized or group-standardized variables.

23 The ***k*th discriminant function** D_k is the linear combination

$$D_k = \sum_{j=1}^{p} \alpha_{kj} (X_j - \overline{\overline{X}}_j)$$

that maximizes the separation, subject to the within-groups covariance between D_k and $D_{k-1}, \ldots, D_1$ being zero, and subject to a constraint on the loadings α_{kj} (see **20**). The kth discriminant function may also be written in terms of group-standardized variables as follows:

$$D_k = \sum_{j=1}^{p} a_{kj} Z_j,$$

with $a_{kj} = \alpha_{kj} \sqrt{V_w(X_j)}$.

24 The **total separation** is the sum of the separations achieved by all p discriminant functions:

$$\text{total separation} = \text{Sep}(D_1) + \cdots + \text{Sep}(D_p).$$

The **percentage separation achieved** by the discriminant function D_j, denoted PSA_j, is

$$\text{PSA}_j = \frac{\text{Sep}(D_j)}{\text{total separation}} \times 100\%.$$

The **cumulative percentage separation achieved** by $D_1, \ldots, D_j$, denoted CPSA_j, is

$$\text{CPSA}_j = \text{PSA}_1 + \cdots + \text{PSA}_j.$$

25 An **allocation rule** for G groups based on the discriminant function is defined by $G-1$ **cut-off points** or **cutpoints** $l_1, \ldots, l_{G-1}$ such that $l_1 < l_2 < \cdots < l_{G-1}$. The allocation rule is of the following form:

$$\begin{cases} \text{if } d \le l_1 & \text{allocate to group } 1, \\ \text{if } l_1 < d \le l_2 & \text{allocate to group } 2, \\ \quad\vdots & \quad\vdots \\ \text{if } l_{G-2} < d \le l_{G-1} & \text{allocate to group } G-1, \\ \text{otherwise} & \text{allocate to group } G. \end{cases}$$

26 In choosing the cutpoints, three factors must be considered.

- ◇ For each group g, the **probability density function** of the values of the discriminant function for an observation randomly selected from all those known to be in group g.
- ◇ For each group g, the **prior probability** that an observation randomly chosen belongs to group g.
- ◇ For each pair of groups, the **cost** of wrongly allocating an observation to one group when it actually belongs to the other group.

27 In practice, it is often assumed that the distribution of values of D for group g is normal with mean μ_g, and that the distributions for the groups have common variance. If the groups are numbered so that $\mu_1 < \mu_2 < \cdots < \mu_G$, then with the above assumption and under the assumptions of equal prior probabilities and equal costs, the cutpoints are given by

$$l_g = \tfrac{1}{2}(\mu_g + \mu_{g+1}), \quad g = 1, \ldots, G-1.$$

28 The **misclassification rate** is the percentage of observations that are misclassified:

$$\text{misclassification rate} = \frac{\text{number misclassified}}{\text{number in sample}} \times 100\%.$$

Information on the way in which observations are misclassified is conveyed in a **confusion matrix**. When there are G groups, the confusion matrix has G rows and G columns, and element (i,j) is the percentage of observations in group i that were allocated to group j.

5.5 Bayesian statistics

The Bayesian approach

1 The probability of an event may sometimes be estimated using the observed or hypothetical **relative frequency** of the event. If this is not possible, **subjective estimates** may be required. These represent the opinions and beliefs of the person making the estimate.

2 For two events A and B, the conditional probabilities $P(A|B)$ and $P(B|A)$ are related by **Bayes' theorem**:

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

where the probability $P(B)$ may be obtained using the formula

$$P(B) = P(B|A)\,P(A) + P(B|\text{not } A)\,P(\text{not } A).$$

3 Bayes' theorem provides a way of updating probabilities. In the absence of additional information, a **prior probability** is determined. Once additional information becomes available, the probability is revised to obtain the **posterior probability**. In **sequential updating**, this procedure is repeated several times.

4 In Bayesian inference about a parameter θ, prior beliefs about θ are represented by a **prior distribution** with probability density function $f(\theta)$, called the **prior density**. A prior is said to be **weak** or **strong** according to how peaked it is, greater uncertainty about θ corresponding to flatter priors.

5 The information about a parameter θ that is contained in observed data $x_1, x_2, \ldots, x_n$ on a random variable X is represented by the **likelihood function** $L(\theta)$.

6 Bayesian inference is based on the **posterior distribution** for θ, given the observed data, with **posterior density** $f(\theta|\text{data})$. This is obtained from the prior density $f(\theta)$ and the likelihood $L(\theta)$ using the expression

$$f(\theta|\text{data}) \propto L(\theta) \times f(\theta),$$

or, in words,

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

The process of obtaining the posterior distribution and using it for inference is called **prior to posterior analysis**.

Prior to posterior analyses

7 Standard distributions are often used to represent prior beliefs about a parameter θ. The **normal prior** $N(a, b)$ may be used to represent beliefs about θ that are symmetric about a single most likely value.

- ◇ Mode = median = mean = a.
- ◇ Variance = b.
- ◇ All values of θ in the range $-\infty < \theta < \infty$ are possible, but only those in the range $a \pm 3\sqrt{b}$ are likely.

8 The **uniform prior** $U(a, b)$, with parameters a and b, may be used to represent the belief that the value of θ lies between a and b when it is not known which values in the interval $[a, b]$ are more likely than others.

9 The uniform prior $U(a, b)$ is **noninformative** if the interval $[a, b]$ necessarily includes all values in the range of θ. **Improper** uniform priors may be used to represent lack of prior information about θ and its range.

10 The **beta prior** with parameters $a > 0$ and $b > 0$, which is denoted $\text{Beta}(a, b)$, may be used to represent beliefs about a proportion θ, $0 \leq \theta \leq 1$.

- ◇ When $a > 1$ and $b > 1$, the beta density has a single mode, given by

$$\text{mode} = \frac{a-1}{a+b-2}.$$

- ◇ When $a < 1$, the beta density has a mode at 0. When $b < 1$, it has a mode at 1. When $a < 1$ and $b < 1$, the density has two modes — at 0 and 1.
- ◇ The mean and variance of $\text{Beta}(a, b)$ are given by

$$\text{mean} = \frac{a}{a+b}, \quad \text{variance} = \frac{ab}{(a+b)^2(a+b+1)}.$$

- ◇ The larger the value of $a + b$ is, the stronger are the beliefs represented by the beta prior.
- ◇ The $\text{Beta}(1, 1)$ distribution is the same as the uniform distribution $U(0, 1)$.

11 The **gamma prior** with parameters $a > 0$ and $b > 0$, which is denoted $\text{Gamma}(a, b)$, may be used to represent beliefs about a parameter θ which takes only non-negative values. The parameter a is the **shape parameter**.

- ◇ When $a > 1$, the prior has a single mode given by

$$\text{mode} = \frac{a-1}{b}.$$

When $0 < a \leq 1$, there is a single mode at 0.

- ◇ The mean and variance of $\text{Gamma}(a, b)$ are given by

$$\text{mean} = \frac{a}{b}, \quad \text{variance} = \frac{a}{b^2}.$$

12 Three steps are involved in specifying a prior $f(\theta)$.

- ◇ Assess the location of $f(\theta)$.
- ◇ Assess the spread of $f(\theta)$.
- ◇ Calculate the values of a and b that give the assessed location and spread.

13 Assessing the location of a prior for a parameter θ is most readily based on the mode or median. The spread of the prior may be assessed using an **equal-tailed 100(1 − α)% interval** (L, U), where

$$P(\theta \leq L) = P(\theta > U) = \tfrac{1}{2}\alpha.$$

14 The mean a and variance b of a normal prior may be chosen as follows:

$$a = \text{assessed mode or median,}$$

$$b = \left(\frac{U-L}{2z}\right)^2,$$

where L and U are the assessed values of the $\alpha/2$-quantile and the $(1-\alpha/2)$-quantile of θ, respectively, and z is the $(1-\alpha/2)$-quantile of $N(0,1)$.

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

15 For some likelihoods, a prior can be used which produces a posterior distribution of the same form as the prior distribution. Such a prior is called a **conjugate prior**. When a conjugate prior is used, the prior to posterior Bayesian analysis is called a **conjugate analysis**. Some standard conjugate analyses are summarized in the table below; x is an observation on a random variable X, and $\overline{x}$ represents the mean of a sample of n observations on X.

Name	Prior	Data	Posterior
beta/binomial	$\theta \sim \text{Beta}(a,b)$	$X \sim B(n,\theta)$	$\text{Beta}(a+x, b+n-x)$
gamma/Poisson	$\mu \sim \text{Gamma}(a,b)$	$X \sim \text{Poisson}(\mu)$	$\text{Gamma}(a+n\overline{x}, b+n)$
normal/normal	$\mu \sim N(a,b)$	$X \sim N(\mu,\sigma^2)$ where σ^2 is known	$N\left(\frac{\sigma^2 a + nb\overline{x}}{\sigma^2+nb}, \frac{\sigma^2 b}{\sigma^2+nb}\right)$

16 Prior to posterior Bayesian analyses may be undertaken using noninformative or improper uniform priors. Some standard analyses are summarized in the table below; x is an observation on a random variable X, and $\overline{x}$ represents the mean of a sample of n observations on X.

Name	Prior	Data	Posterior
uniform/binomial	$\theta \sim U(0,1)$	$X \sim B(n,\theta)$	$\text{Beta}(1+x, 1+n-x)$
uniform/Poisson	$\mu \sim$ improper uniform on $[0,\infty)$	$X \sim \text{Poisson}(\mu)$	$\text{Gamma}(n\overline{x}, n)$
uniform/normal	$\mu \sim$ improper uniform on $(-\infty,\infty)$	$X \sim N(\mu,\sigma^2)$ where σ^2 is known	$N\left(\overline{x}, \frac{\sigma^2}{n}\right)$

17 A plot of the posterior distribution for a parameter θ is always helpful. The location of the posterior distribution may be summarized conveniently by the **posterior mode** or the **posterior median**. The spread of the posterior distribution may be summarized by the **posterior variance**. Probabilities calculated from posterior distributions may also be of interest.

18 An interval (l,u) is a **$100(1-\alpha)$% credible interval** for a parameter θ if the posterior probability that $l \le \theta \le u$, given the data, is equal to $1-\alpha$:

$$P(l \le \theta \le u|\text{data}) = 1-\alpha.$$

The probability $1-\alpha$ is the **credibility level** of the interval.

19 A **Highest Posterior Density** (**HPD**) credible interval for a posterior distribution with a single mode contains the most likely values of θ. An **equal-tailed** credible interval satisfies

$$P(\theta < l|\text{data}) = P(\theta > u|\text{data}) = \tfrac{1}{2}\alpha.$$

Bayesian inference via simulation

20 When a conjugate analysis does not seem appropriate, or when the mathematics involved in using a conjugate analysis is complicated, **simulation** can be used to obtain information about the posterior distribution. Simulation is particularly useful in non-conjugate Bayesian analyses or when functions of parameters are of interest.

21 **Stochastic simulation**, or **Monte Carlo** (MC) **simulation**, involves mimicking the properties of a distribution by 'randomly' sampling values from the distribution.

22 The **Monte Carlo standard error** of a mean obtained by simulation, or the **MC error**, relates to the variability of the simulation, and may be reduced by increasing N, the number of values sampled in the simulation. The **5% rule of thumb** states that N should be large enough to ensure that the Monte Carlo standard error of the mean is less than 5% of the sample standard deviation.

23 To make inferences about a parameter ϕ which is some function $g(\theta)$ of a parameter θ that can readily be simulated, proceed as follows.

- ◇ Simulate N values of θ, denoted $\theta_1, \ldots, \theta_N$.
- ◇ Apply the function g to each of the simulated values, to give values $\phi_1 = g(\theta_1)$, $\ldots$, $\phi_n = g(\theta_n)$.
- ◇ Use these values to make inferences about ϕ.

24 For a Bayesian analysis involving more than one unknown parameter, interest lies in the joint distribution and in the marginal distributions of the parameters.

- ◇ The **joint distribution** $f(\theta, \phi)$ of two unknown parameters θ and ϕ describes how the two parameters vary together, and may be represented by a scatterplot of simulated pairs of values (θ_1, ϕ_1), $\ldots$, (θ_N, ϕ_N).
- ◇ The **marginal distributions** are the distributions of θ and ϕ considered separately, and may be estimated using histograms of the simulated values $\theta_1, \ldots, \theta_N$ and $\phi_1, \ldots, \phi_N$, respectively.
- ◇ The mean of the marginal distribution of a parameter can be estimated by the sample mean of the simulated values of the parameter; quantiles of the distribution can be estimated using sample quantiles.

Markov chain Monte Carlo

25 A **Markov chain** is a sequence of random variables $X_1, X_2, \ldots$ for which the distribution of X_{k+1} depends only on the value of X_k and not on any earlier values in the chain. A realization of a Markov chain may be represented using a **trace plot**, that is, a plot in which the values of the Markov chain are plotted against the iteration number. Under suitable conditions, the values in a realization of a Markov chain will eventually settle down, or **converge**, to an **equilibrium distribution**.

26 **Markov chain Monte Carlo** (**MCMC**) is a technique for obtaining a posterior distribution of interest as the equilibrium distribution of a Markov chain. It is particularly useful when conjugate analyses are not available.

27 **Convergence** of a Markov chain can be assessed graphically by running the Markov chain several times from different initial values and checking that the realizations eventually overlap. The period before they overlap is the **burn-in** period. Inferences can be based on all samples obtained after the burn-in period.

28 Samples obtained using MCMC are dependent. However, the MC error can still be estimated and the 5% rule of thumb used to estimate the sample size to be used.

6 Statistical tables

Table 1 Probabilities for the standard normal distribution, $P(Z \leq z)$

z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Example: If $Z \sim N(0,1)$, then $P(Z \leq 0.62) = 0.7324$.

Table 2 Quantiles for the standard normal distribution, $P(Z \leq q_\alpha) = \alpha$

α	q_α	α	q_α	α	q_α	α	q_α
0.50	0.00000	0.67	0.4399	0.84	0.9945	0.955	1.695
0.51	0.02507	0.68	0.4677	0.85	1.036	0.960	1.751
0.52	0.05015	0.69	0.4959	0.86	1.080	0.965	1.812
0.53	0.07527	0.70	0.5244	0.87	1.126	0.970	1.881
0.54	0.1004	0.71	0.5534	0.88	1.175	0.975	1.960
0.55	0.1257	0.72	0.5828	0.89	1.227	0.980	2.054
0.56	0.1510	0.73	0.6128	0.90	1.282	0.985	2.170
0.57	0.1764	0.74	0.6433	0.905	1.311	0.990	2.326
0.58	0.2019	0.75	0.6745	0.910	1.341	0.991	2.366
0.59	0.2275	0.76	0.7063	0.915	1.372	0.992	2.409
0.60	0.2533	0.77	0.7388	0.920	1.405	0.993	2.457
0.61	0.2793	0.78	0.7722	0.925	1.440	0.994	2.512
0.62	0.3055	0.79	0.8064	0.930	1.476	0.995	2.576
0.63	0.3319	0.80	0.8416	0.935	1.514	0.996	2.652
0.64	0.3585	0.81	0.8779	0.940	1.555	0.997	2.748
0.65	0.3853	0.82	0.9154	0.945	1.598	0.998	2.878
0.66	0.4125	0.83	0.9542	0.950	1.645	0.999	3.090

Example: If $Z \sim N(0,1)$, then $P(Z \leq 1.645) = 0.950$, so $q_{0.95} = 1.645$.

Table 3 Quantiles for χ^2-distributions

df	0.1	0.3	0.5	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.999
1	0.016	0.148	0.455	0.708	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.83
2	0.211	0.713	1.39	1.83	2.41	3.22	4.61	5.99	7.38	9.21	10.60	13.82
3	0.584	1.42	2.37	2.95	3.66	4.64	6.25	7.81	9.35	11.34	12.84	16.27
4	1.06	2.19	3.36	4.04	4.88	5.99	7.78	9.49	11.14	13.28	14.86	18.47
5	1.61	3.00	4.35	5.13	6.06	7.29	9.24	11.07	12.83	15.09	16.75	20.52
6	2.20	3.83	5.35	6.21	7.23	8.56	10.64	12.59	14.45	16.81	18.55	22.46
7	2.83	4.67	6.35	7.28	8.38	9.80	12.02	14.07	16.01	18.48	20.28	24.32
8	3.49	5.53	7.34	8.35	9.52	11.03	13.36	15.51	17.53	20.09	21.95	26.12
9	4.17	6.39	8.34	9.41	10.66	12.24	14.68	16.92	19.02	21.67	23.59	27.88
10	4.87	7.27	9.34	10.47	11.78	13.44	15.99	18.31	20.48	23.21	25.19	29.59
11	5.58	8.15	10.34	11.53	12.90	14.63	17.28	19.68	21.92	24.72	26.76	31.26
12	6.30	9.03	11.34	12.58	14.01	15.81	18.55	21.03	23.34	26.22	28.30	32.91
13	7.04	9.93	12.34	13.64	15.12	16.98	19.81	22.36	24.74	27.69	29.82	34.53
14	7.79	10.82	13.34	14.69	16.22	18.15	21.06	23.68	26.12	29.14	31.32	36.12
15	8.55	11.72	14.34	15.73	17.32	19.31	22.31	25.00	27.49	30.58	32.80	37.70
16	9.31	12.62	15.34	16.78	18.42	20.47	23.54	26.30	28.85	32.00	34.27	39.25
17	10.09	13.53	16.34	17.82	19.51	21.61	24.77	27.59	30.19	33.41	35.72	40.79
18	10.86	14.44	17.34	18.87	20.60	22.76	25.99	28.87	31.53	34.81	37.16	42.31
19	11.65	15.35	18.34	19.91	21.69	23.90	27.20	30.14	32.85	36.19	38.58	43.82
20	12.44	16.27	19.34	20.95	22.77	25.04	28.41	31.41	34.17	37.57	40.00	45.31
21	13.24	17.18	20.34	21.99	23.86	26.17	29.62	32.67	35.48	38.93	41.40	46.80
22	14.04	18.10	21.34	23.03	24.94	27.30	30.81	33.92	36.78	40.29	42.80	48.27
23	14.85	19.02	22.34	24.07	26.02	28.43	32.01	35.17	38.08	41.64	44.18	49.73
24	15.66	19.94	23.34	25.11	27.10	29.55	33.20	36.42	39.36	42.98	45.56	51.18
25	16.47	20.87	24.34	26.14	28.17	30.68	34.38	37.65	40.65	44.31	46.93	52.62
26	17.29	21.79	25.34	27.18	29.25	31.79	35.56	38.89	41.92	45.64	48.29	54.05
27	18.11	22.72	26.34	28.21	30.32	32.91	36.74	40.11	43.19	46.96	49.64	55.48
28	18.94	23.65	27.34	29.25	31.39	34.03	37.92	41.34	44.46	48.28	50.99	56.89
29	19.77	24.58	28.34	30.28	32.46	35.14	39.09	42.56	45.72	49.59	52.34	58.30
30	20.60	25.51	29.34	31.32	33.53	36.25	40.26	43.77	46.98	50.89	53.67	59.70
31	21.43	26.44	30.34	32.35	34.60	37.36	41.42	44.99	48.23	52.19	55.00	61.10
32	22.27	27.37	31.34	33.38	35.66	38.47	42.58	46.19	49.48	53.49	56.33	62.49
33	23.11	28.31	32.34	34.41	36.73	39.57	43.75	47.40	50.73	54.78	57.65	63.87
34	23.95	29.24	33.34	35.44	37.80	40.68	44.90	48.60	51.97	56.06	58.96	65.25
35	24.80	30.18	34.34	36.47	38.86	41.78	46.06	49.80	53.20	57.34	60.27	66.62
36	25.64	31.12	35.34	37.50	39.92	42.88	47.21	51.00	54.44	58.62	61.58	67.99
37	26.49	32.05	36.34	38.53	40.98	43.98	48.36	52.19	55.67	59.89	62.88	69.35
38	27.34	32.99	37.34	39.56	42.05	45.08	49.51	53.38	56.90	61.16	64.18	70.70
39	28.20	33.93	38.34	40.59	43.11	46.17	50.66	54.57	58.12	62.43	65.48	72.05
40	29.05	34.87	39.34	41.62	44.16	47.27	51.81	55.76	59.34	63.69	66.77	73.40

Example: If $X \sim \chi^2(4)$, the chi-squared distribution on 4 degrees of freedom (df), then $P(X \leq 7.78) = 0.9$, so $q_{0.9} = 7.78$.

Printed in the United Kingdom.