A

Facts from Probability, Statistics, and Algebra

A.1 Introduction

It is assumed that the reader is already familiar with the basics of probability, statistics, matrix algebra, and other mathematical topics needed in this book, and so the goal of this appendix is merely to provide a quick review and cover some more advanced topics that may not be familiar.

A.2 Probability Distributions

A.2.1 Cumulative Distribution Functions

The *cumulative distribution function (CDF)* of Y is defined as

$$F_Y(y) = P\{Y \le y\}.$$

If Y has a PDF f_Y, then

$$F_Y(y) = \int_{-\infty}^{y} f_Y(u)\, du.$$

Many CDFs and PDFs can be calculated by computer software packages, for instance, `pnorm()`, `pt()`, and `pbinom()` in `R` calculate, respectively, the CDF of a normal, t, and binomial random variable. Similiarly, `dnorm()`, `dt()`, and `dbinom()` calculate the PDFs of these distributions.

D. Ruppert, D.S. Matteson, *Statistics and Data Analysis for Financial Engineering*, Springer Texts in Statistics,
DOI 10.1007/978-1-4939-2614-5

A.2.2 Quantiles and Percentiles

If the CDF $F(y)$ of a random variable Y is continuous and strictly increasing, then it has an inverse function F^{-1}. For each q between 0 and 1, $F^{-1}(q)$ is called the *q-quantile* or $100q$th percentile.

The median is the 50 % percentile or 0.5-quantile. The 25 % and 75 % percentiles (0.25- and 0.75-quantiles) are called the first and third quartiles and the median is the second quartile. The three quartiles divide the range of a continuous random variable into four groups of equal probability. Similarly, the 20 %, 40 %, 60 %, and 80 % percentiles are called quintiles and the 10 %, 20 %, ..., 90 % percentiles are called deciles.

For any CDF F, invertible or not, the *pseudo-inverse* is defined as

$$F^{-}(x) = \inf\{y : F(y) \geq x\}.$$

Here "inf" is the infinum or greatest lower bound of a set; see Appendix A.5. For any q between 0 and 1, the qth quantile will defined as $F^{-}(q)$. If F is invertible, then $F^{-1} = F^{-}$, so this definition of quantile agree with the one for invertible CDFs. F^{-} is often called the *quantile function*.

Sometimes a $(1-\alpha)$-quantile is called an α-upper quantile, to emphasize the amount of probability above the quantile. In analogy, a quantile might also be referred to as lower quantile.

Quantiles are said to "respect transformations" in the following sense. If Y is a random variable whose q-quantile equals y_q, if g is a strictly increasing function, and if $X = g(Y)$, then $g(y_q)$ is the q-quantile of X; see (A.5).

A.2.3 Symmetry and Modes

A probability density function (PDF) f is said to be *symmetric* about μ if $f(\mu - y) = f(\mu + y)$ for all y. A *mode* of a PDF is a local maximum, that is a value y such that for some $\epsilon > 0$, $f(y) > f(x)$ if $y - \epsilon < x < y$ or $y < x < y + \epsilon$. A PDF with one mode is called *unimodal*, with two modes *bimodal*, and with two or more modes *multimodal*.

A.2.4 Support of a Distribution

The support of a *discrete* distribution is the set of all y that have a positive probability. More generally, a point y is in the support of a distribution if, for every $\epsilon > 0$, the interval $(y-\epsilon, y+\epsilon)$ has positive probability. For example, the support of a normal distribution is $(-\infty, \infty)$, the support of a gamma or log-normal distribution is $[0, \infty)$, and the support of a binomial(n, p) distribution is $\{0, 1, 2, \ldots, n\}$ provided $p \neq 0, 1$.[1]

[1] It is assumed that most readers are already familiar with the normal, gamma, log-normal, and binomial distributions. However, these distributions will be discussed in some detail later.

A.3 When Do Expected Values and Variances Exist?

The expected value of a random variable could be infinite or not exist at all. Also, a random variable need not have a well-defined and finite variance. To appreciate these facts, let Y be a random variable with density f_Y. The expectation of Y is

$$\int_{-\infty}^{\infty} y f_Y(y) dy$$

provided that this integral is defined. If

$$\int_{-\infty}^{0} y f_Y(y) dy = -\infty \text{ and } \int_{0}^{\infty} y f_Y(y) dy = \infty, \tag{A.1}$$

then the expectation is, formally, $-\infty + \infty$, which is not defined, so the expectation does not exist. If integrals in (A.1) are both finite, then $E(Y)$ exists and equals the sum of these two integrals. The expectation can exist but be infinite, because if

$$\int_{-\infty}^{0} y f_Y(y) dy = -\infty \text{ and } \int_{0}^{\infty} y f_Y(y) dy < \infty,$$

then $E(Y) = -\infty$, and if

$$\int_{-\infty}^{0} y f_Y(y) dy > -\infty \text{ and } \int_{0}^{\infty} y f_Y(y) dy = \infty,$$

then $E(Y) = \infty$.

If $E(Y)$ is not defined or is infinite, then the variance that involves $E(Y)$ cannot be defined either. If $E(Y)$ is defined and finite, then the variance is also defined. The variance is finite if $E(Y^2) < \infty$; otherwise the variance is infinite.

The nonexistence of finite expected values and variances is of importance for modeling financial markets data, because, for example, the popular GARCH models discussed in Chap. 14 need not have finite expected values and variances. Also, t-distributions that, as demonstrated in Chap. 5, can provide good fits to equity returns may have nonexistent means or variances.

One could argue that any variable Y derived from financial markets will be bounded, that is, that there is a constant $M < \infty$ such that $P(|Y| \leq M) = 1$. In this case, the integrals in (A.1) are both finite, in fact at most M, and $E(Y)$ exists and is finite. Also, $E(Y^2) \leq M^2$, so the variance of Y is finite. So should we worry at all about the mathematically niceties of whether expected values and variances exist and are finite? The answer is that we should. A random variable might be bounded in absolute value by a very large constant M and yet, if M is large enough, behave much like a random variable that does not have an expected value or has an expected value that is infinite or has a finite expected value but an infinite variance. This can be seen in the simulations

of GARCH processes. Results from computer simulations are bounded by the maximum size of a number in the computer. Yet these simulations behave as if the variance were infinite.

A.4 Monotonic Functions

The function g is increasing if $g(x_1) \leq g(x_2)$ whenever $x_1 < x_2$ and strictly increasing if $g(x_1) < g(x_2)$ whenever $x_1 < x_2$. Decreasing and strictly decreasing are defined similarly, and g is (strictly) monotonic if it is either (strictly) increasing or (strictly) decreasing.

A.5 The Minimum, Maximum, Infinum, and Supremum of a Set

The minimum and maximum of a set are its smallest and largest values, if these exists. For example, if $A = \{x : \ 0 \leq x \leq 1\}$, then the minimum and maximum of A are 0 and 1. However, not all sets have a minimum or a maximum, for example, $B = \{x : \ 0 < x < 1\}$ has neither a minimum nor a maximum. Every set as an infinum (or inf) and a supremum (or sup). The inf of a set C is the largest number that is less than or equal to all elements of C. Similarly, the sup of C is the smallest number that is greater than or equal to every element of C. The set B just defined has an inf of 0 and a sup of 1. The following notation is standard: $\min(C)$ and $\max(C)$ are the minimum and maximum of C, if these exist, and $\inf(C)$ and $\sup(C)$ are the infinum and supremum.

A.6 Functions of Random Variables

Suppose that X is a random variable with PDF $f_X(x)$ and $Y = g(X)$ for g a strictly increasing function. Since g is strictly increasing, it has an inverse, which we denote by h. Then Y is also a random variable and its CDF is

$$F_Y(y) = P(Y \leq y) = P\{g(X) \leq y\} = P\{X \leq h(y)\} = F_X\{h(y)\}. \quad \text{(A.2)}$$

Differentiating (A.2), we find the PDF of Y:

$$f_Y(y) = f_X\{h(y)\}h'(y). \quad \text{(A.3)}$$

Applying a similar argument to the case, where g is strictly decreasing, one can show that whenever g is strictly monotonic, then

$$f_Y(y) = f_X\{h(y)\}|h'(y)|. \quad \text{(A.4)}$$

Also from (A.2), when g is strictly increasing, then

$$F_Y^{-1}(p) = g\{F_X^{-1}(p)\}, \tag{A.5}$$

so that the pth quantile of Y is found by applying g to the pth quantile of X. When g is strictly decreasing, then it maps the pth quantile of X to the $(1-p)$th quantile of Y.

Result A.1 *Suppose that $Y = a + bX$ for some constants a and $b \neq 0$. Let $g(x) = a + bx$, so that the inverse of g is $h(y) = (y-a)/b$ and $h'(y) = 1/b$. Then*

$$\begin{aligned} F_Y(y) &= F_X\{b^{-1}(y-a)\}, \quad b > 0, \\ &= 1 - F_X\{b^{-1}(y-a)\}, \quad b < 0, \\ f_Y(y) &= |b|^{-1} f_X\{b^{-1}(y-a)\}, \end{aligned}$$

and

$$\begin{aligned} F_Y^{-1}(p) &= a + bF_X^{-1}(p), \quad b > 0 \\ &= a + bF_X^{-1}(1-p), \quad b < 0. \end{aligned}$$

A.7 Random Samples

We say that $\{Y_1, \ldots, Y_n\}$ is a *random sample* from a probability distribution if they each have that probability distribution and are independent. In this case, we also say that they are *independent and identically distributed* or simply i.i.d. The probability distribution is often called the population and its expected value, variance, CDF, and quantiles are called the *population mean*, *population variance*, *population CDF*, and *population quantiles*. It is worth mentioning that the population is, in effect, infinite. There is a statistical theory of sampling, usually without replacement, from finite populations, but sampling of this type will not concern us here. Even in cases where the population is finite, such as, when sampling house prices, the population is usually large enough, so that it can be treated as infinite.

If $Y_1, \ldots, Y_n$ is a sample from an unknown probability distribution, then the population mean can be estimated by the *sample mean*

$$\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i, \tag{A.6}$$

and the population variance can be estimated by the *sample variance*

$$s_Y^2 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}. \tag{A.7}$$

The reason for the denominator of $n-1$ rather than n is discussed in Sect. 5.9. The *sample standard deviation* is s_Y, the square root of s_Y^2.

A.8 The Binomial Distribution

Suppose that we conduct n experiments for some fixed (nonrandom) integer n. On each experiment there are two possible outcomes called "success" and "failure"; the probability of a success is p, and the probability of a failure is $q = 1 - p$. It is assumed that p and q are the same for all n experiments. Let Y be the total number of successes, so that Y will equal 0, 1, 2, ..., or n. If the experiments are independent, then

$$P(Y = k) = \binom{n}{k} p^k q^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n,$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

The distribution of Y is called the *binomial distribution* and denoted Binomial(n, p). The expected value of Y is np and its variance is npq. The Binomial$(1, p)$ distribution is also called the Bernoulli distribution and its density is

$$P(Y = y) = p^y (1-p)^{1-y}, \ y = 0, 1. \tag{A.8}$$

Notice that p^y is equal to either p (when $y = 1$) or 1 (when $y = 0$), and similarly for $(1-p)^{1-y}$.

The functions `pbinom()`, `dbinom()`, `qbinom()`, and `rbinom()` compute binomial CDFs, pdfs, quantiles, and random numbers, respectively. For example,

```
> pbinom(3,6,0.5)
[1] 0.65625
```

shows that the probability of 3 or less heads in 6 tosses of a fair coin is 0.65625.

A.9 Some Common Continuous Distributions

A.9.1 Uniform Distributions

The uniform distribution on the interval (a, b) is denoted by Uniform(a, b) and has PDF equal to $1/(b-a)$ on (a, b) and equal to 0 outside this interval. It is easy to check that if Y is Uniform(a, b), then its expectation is

$$E(Y) = \frac{1}{b-a} \int_a^b Y \, dY = \frac{a+b}{2},$$

which is the midpoint of the interval. Also,

$$E(Y^2) = \frac{1}{b-a} \int_a^b Y^2 \, dY = \frac{Y^3|_a^b}{3(b-a)} = \frac{b^2 + ab + a^2}{3}.$$

Therefore,

$$\sigma_Y^2 = E(Y^2) - \{E(Y)\}^2 = \frac{b^2 + ab + a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

Reparameterization means replacing the parameters of a distribution by an equivalent set. The uniform distribution can be reparameterized by using $\mu = (a+b)/2$ and $\sigma = (b-a)/\sqrt{12}$ as the parameters. Then μ is a location parameter and σ is the scale parameter. Which parameterization of a distribution is used depends upon which aspects of the distribution one wishes to emphasize. The parameterization (a, b) of the uniform specifies its endpoints while the parameterization (μ, σ) gives the mean and standard deviation. One is free to move back and forth between two or more parameterizations, using whichever is most useful in a given context. The uniform distribution does not have a shape parameter since the shape of its density is always rectangular.

The functions `punif()`, `dunif()`, `qunif()`, and `runif()` compute uniform CDFs, pdfs, quantiles, and random numbers, respectively. For example,

```
> runif(3,0,5)
[1] 1.799252 4.003232 3.978002
```

are three random numbers uniformly distributed between 0 and 5.

A.9.2 Transformation by the CDF and Inverse CDF

If Y has a continuous CDF F, then $F(Y)$ has a Uniform(0,1) distribution. $F(Y)$ is often called the *probability transformation* of Y. This fact is easy to see if F is strictly increasing, since then F^{-1} exists, so that

$$P\{F(Y) \leq y\} = P\{Y \leq F^{-1}(y)\} = F\{F^{-1}(y)\} = y. \tag{A.9}$$

The result holds even if F is not strictly increasing, but the proof is slightly more complicated. It is only necessary that F be continuous.

If U is Uniform(0,1) and F is a CDF, then $Y = F^-(U)$ has F as its CDF. Here F^- is the pseudo-inverse of F. This can be proved easily when F is continuous and strictly increasing, since then $F^{-1} = F^-$ and

$$P(Y \leq y) = P\{F^{-1}(U) \leq y\} = P\{Y \leq F(y)\} = F(y).$$

In fact, the result holds for any CDF F, but it is more difficult to prove in the general case. $F^-(U)$ is often called the *quantile transformation* since F^- is the quantile function.

A.9.3 Normal Distributions

The *standard normal distribution* has the familiar bell-shaped density

$$\phi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-y^2/2\right), \quad -\infty < y < \infty.$$

The standard normal has mean 0 and variance 1. If Z is standard normal, then the distribution of $\mu + \sigma Z$ is called the *normal distribution with mean* μ *and variance* σ^2 and denoted by $N(\mu, \sigma^2)$. By Result A.1, the $N(\mu, \sigma^2)$ density is

$$\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}. \tag{A.10}$$

The parameter μ is a location parameter and σ is a scale parameter. The normal distribution does not have a shape parameter since its density is always the same bell-shaped curve.[2] The standard normal CDF is

$$\Phi(y) = \int_{-\infty}^{y} \phi(u)du.$$

Φ can be evaluated using software such as R's `pnorm` function. If Y is $N(\mu, \sigma^2)$, then since $Y = \mu + \sigma Z$, where Z is standard normal, by Result A.1,

$$F_Y(y) = \Phi\{(y-\mu)/\sigma\}. \tag{A.11}$$

Normal distribution are also called Gaussian distributions after the great German mathematician Carl Friedrich Gauss.

Normal Quantiles

The q-quantile of the $N(0,1)$ distribution is $\Phi^{-1}(q)$ and, more generally, the q-quantile of an $N(\mu, \sigma^2)$ distribution is $\mu + \sigma\Phi^{-1}(q)$. The α-upper quantile of Φ, that is, $\Phi^{-1}(1-\alpha)$, is denoted by z_α. As shown later, z_α is widely used for confidence intervals.

For example, $z_{0.1}$ and $z_{0.01}$ are 1.282 and 2.326, respectively, as can be seen in the following R output:

```
> round(qnorm(c(0.1, 0.01), lower.tail = FALSE), 3)
[1] 1.282 2.326
```

A.9.4 The Lognormal Distribution

If Z is distributed $N(\mu, \sigma^2)$, then $Y = \exp(Z)$ is said to have a Lognormal(μ, σ^2) distribution. In other words, Y is *lognormal* if its logarithm is normally

[2] In contrast, a t-density is also a bell curve, but the exact shape of the bell depends on a shape parameter, the degrees of freedom which is a tail index.

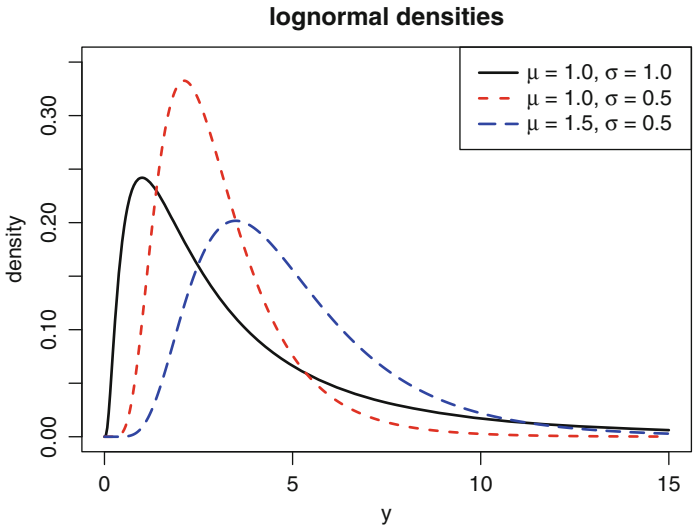

Fig. A.1. *Examples of lognormal probability densities. Here μ and σ are the log-mean and log-standard deviation, that is, the mean and standard deviation of the logarithm of the lognormal random variable.*

distributed. We will call μ the log-mean and σ the log-standard deviation. Also, σ^2 will be called the log-variance.

The median of Y is $\exp(\mu)$ and the expected value of Y is $\exp(\mu+\sigma^2/2)$.[3] The expectation is larger than the median because the lognormal distribution is right skewed, and the skewness is more extreme with larger values of σ. Skewness is discussed further in Sect. 5.4. The probability density functions of several lognormal distributions are shown in Fig. A.1.

The log-mean μ is a scale parameter and the log-standard deviation σ is a shape parameter. The lognormal distribution does not have a location parameter since its support is fixed to start at 0.

Use the functions `plnorm()`, `dlnorm()`, `qlnorm()`, and `rlnorm()` for the lognormal distribution. For example,

```
> options(digits = 3)
> dlnorm(0.5, meanlog = 1, sdlog = 2)
[1] 0.279
```

computes the lognormal density at 0.5 when the log-mean is 1 and the log-standard deviation is 2.

[3] It is important to remember that if Y is lognormal(μ, σ), then μ is the expected value of $\log(Y)$, not of Y.

A.9.5 Exponential and Double-Exponential Distributions

The *exponential distribution* with scale parameter $\theta > 0$, which we denote by Exponential(θ), has CDF

$$F(y) = 1 - e^{-y/\theta}, \quad y > 0.$$

The Exponential(θ) distribution has PDF

$$f(y) = \frac{e^{-y/\theta}}{\theta}, \tag{A.12}$$

expected value θ, and standard deviation θ. The inverse CDF is

$$F^{-1}(y) = -\theta \log(1 - y), \quad 0 < y < 1.$$

Use the functions `pexp()`, `dexp()`, `qexp()`, and `rexp()` for the exponential distribution.

The *double-exponential* or *Laplace distribution* with mean μ and scale parameter θ has PDF

$$f(y) = \frac{e^{-|y-\mu|/\theta}}{2\theta}. \tag{A.13}$$

If Y has a double-exponential distribution with mean μ, then $|Y - \mu|$ has an exponential distribution. A double-exponential distribution has a standard deviation of $\sqrt{2}\theta$. The mean μ is a location parameter and θ is a scale parameter.

A.9.6 Gamma and Inverse-Gamma Distributions

The *gamma distribution* with scale parameter $b > 0$ and shape parameter $\alpha > 0$ has density

$$\frac{y^{\alpha-1}}{\Gamma(\alpha)b^{\alpha}} \exp(-y/b),$$

where Γ is the gamma function defined in Sect. 5.5.2. The mean, variance, and skewness coefficient of this distribution are $b\alpha$, $b^2\alpha$, and $2\alpha^{-1/2}$, respectively. Figure A.2 shows gamma densities with shape parameters equal to 0.75, 3/2, and 7/2 and each with a mean equal to 1.

The gamma distribution is often parameterized using $\beta = 1/b$, so that the density is

$$\frac{\beta^{\alpha} y^{\alpha-1}}{\Gamma(\alpha)} \exp(-\beta y).$$

With this form of the parameterization, β is an *inverse-scale parameter* and the mean and variance are α/β and α/β^2. Also, β is often called the rate parameter, e.g., in `R`.

Use the functions `pgamma()`, `dgamma()`, `qgamma()`, and `rgamma()` for the gamma distribution. For example, the median of the gamma distribution with $\alpha = 2$ and $\beta = 3$ can be computed in two equivalent ways:

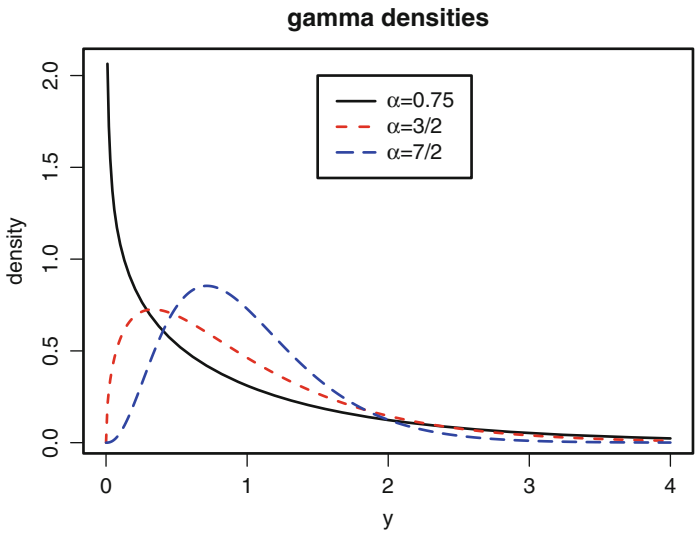

Fig. A.2. *Examples of gamma probability densities with differing shape parameters. In each case, the scale parameter has been chosen so that the expectation is 1.*

```
> qgamma(0.5, shape = 2, rate = 3)
[1] 0.559
> qgamma(0.5, shape = 2, scale = 1/3)
[1] 0.559
```

If X has a gamma distribution with inverse-scale parameter β and shape parameter α, then we say that $1/X$ has an *inverse-gamma distribution* with scale parameter β and shape parameter α. The mean of this distribution is $\beta/(\alpha-1)$ provided $\alpha > 1$ and the variance is $\beta^2/\{(\alpha-1)^2(\alpha-2)\}$ provided that $\alpha > 2$.

A.9.7 Beta Distributions

The beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$ has density

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1. \tag{A.14}$$

The mean and variance are $\alpha/(\alpha+\beta)$ and $(\alpha\beta)/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$, and if $\alpha > 1$ and $\beta > 1$, then the mode is $(\alpha-1)/(\alpha+\beta-2)$.

Figure A.3 shows beta densities for several choices of shape parameters. A beta density is right-skewed, symmetric about 1/2, or left-skewed depending on whether $\alpha < \beta$, $\alpha = \beta$, or $\alpha > \beta$.

Use the functions `pbeta()`, `dbeta()`, `qbeta()`, and `rbeta()` for the beta distribution. For example, the code below created Fig. A.3.

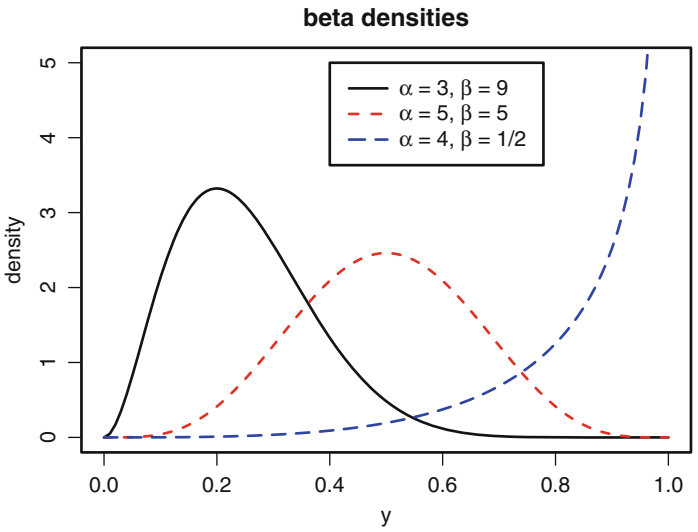

Fig. A.3. *Examples of beta probability densities with differing shape parameters.*

```
pdf("beta_densities.pdf", width = 6, height = 5)  ##  Figure A.3
par(lwd = 2)
x = seq(0, 1, 0.01)
plot(x, dbeta(x, 3, 9), type = "l", lty = 1, xlab = "y",
     ylab = "density", main = "beta densities", ylim = c(0, 5))
lines(x, dbeta(x, 5, 5), type = "l", lty = 2, col = "red")
lines(x, dbeta(x, 4, 1/2), type = "l", lty = 5, col = "blue")
legend(0.4, 5, c(
  expression(paste(alpha," = 3, ",beta," = 9")) ,
  expression(paste(alpha," = 5, ",beta," = 5")),
  expression(paste(alpha," = 4, ",beta," = 1/2"))),
       lty = c(1, 2,5 ), col = c("black", "red", "blue"), lwd = 2)
graphics.off()
```

A.9.8 Pareto Distributions

A random variable X has a Pareto distribution, named after the Swiss economics professor Vilfredo Pareto (1848–1923), if its CDF for some $a > 0$

$$F(x) = 1 - \left(\frac{c}{x}\right)^a, \quad x > c, \tag{A.15}$$

where $c > 0$ is the minimum possible value of X.

The PDF of the distribution in (A.15) is

$$f(x) = \frac{ac^a}{x^{a+1}}, \quad x > c, \tag{A.16}$$

so a Pareto distribution has polynomial tails and a is the *tail index*. It is also called the *Pareto constant*.

A.10 Sampling a Normal Distribution

A common situation is that we have a random sample from a normal distribution and we wish to have confidence intervals for the mean and variance or test hypotheses about these parameters. Then, the following distributions are very important, since they are the basis for many commonly used confidence intervals and tests.

A.10.1 Chi-Squared Distributions

Suppose that $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$. Then, the distribution of $Z_1^2 + \cdots + Z_n^2$ is called the *chi-squared distribution* with n *degrees of freedom*. This distribution has an expected value of n and a variance of $2n$. The α-upper quantile of this distribution is denoted by $\chi^2_{\alpha,n}$ and is used in tests and confidence intervals about variances; see Appendix A.10.1 for the latter. Also, as discussed in Sect. 5.11, $\chi^2_{\alpha,n}$ is used in likelihood ratio testing. As an example, $\chi^2_{0.05,10}$ is 18.31 and can be computed in two ways:

```
> qchisq(0.05, 10, lower.tail = FALSE)
[1] 18.31
> qchisq(0.95, 10)
[1] 18.31
```

So far, the degrees-of-freedom parameter has been an integer-valued, but this can be generalized. The chi-squared distribution with ν degrees of freedom is equal to the gamma distribution with scale parameter equal to 2 and shape parameter equal to $\nu/2$. Thus, since the shape parameter of a gamma distribution can be any positive value, the chi-squared distribution can be defined for any positive value of ν as the gamma distribution with scale and shape parameters equal to 2 and $\nu/2$, respectively.

A.10.2 *F*-Distributions

If U and W are independent and chi-squared-distributed with n_1 and n_2 degrees of freedom, respectively, then the distribution of

$$\frac{U/n_1}{W/n_2}$$

is called the F-distribution with n_1 and n_2 degrees of freedom. The α-upper quantile of this distribution is denoted by F_{α,n_1,n_2}. F_{α,n_1,n_2} is used as a critical value for F-tests in regression. For example, $F_{0.95,3,7}$ is 4.347:

```
> qf(0.95, 3, 7)
[1] 4.347
> qf(0.05, 3, 7, lower.tail = FALSE)
[1] 4.347
```

The degrees-of-freedom parameters of the chi-square, t-, and F-distributions are shape parameters.

A.11 Law of Large Numbers and the Central Limit Theorem for the Sample Mean

Suppose that $\overline{Y}_n$ is the mean of an i.i.d. sample $Y_1, \ldots, Y_n$. We assume that their common expected value $E(Y_1)$ exists and is finite and call it μ. The *law of large numbers* states that

$$P(\overline{Y}_n \to \mu \text{ as } n \to \infty) = 1.$$

Thus, the sample mean will be close to the population mean for large enough sample sizes. However, even more is true. The famous *central limit theorem* (CLT) states that if the common variance σ^2 of $Y_1, \ldots, Y_n$ is finite, then the probability distribution of $\overline{Y}_n$ gets closer to a normal distribution as n converges to ∞. More precisely, the CLT states that

$$P\{\sqrt{n}(\overline{Y}_n - \mu) \le y\} \to \Phi(y/\sigma) \text{ as } n \to \infty \text{ for all } y. \tag{A.17}$$

Stated differently, for large n, $\overline{Y}$ is approximately $N(\mu, \sigma^2/n)$.

Students often misremember or misunderstand the CLT. A common misconception is that a large *population* is approximately normally distributed. The CLT says nothing about the distribution of a population; it is only a statement about the distribution of a sample mean. Also, the CLT does not assume that the population is large; it is the size of the sample that is converging to infinity. Assuming that the sampling is with replacement, the population could be quite small, in fact, with only two elements.

When the variance of $Y_1, \ldots, Y_n$ is infinite, then the limit distribution of $\overline{Y}_n$ may still exist but will be a nonnormal stable distribution.

Although the CLT was first discovered for the sample mean, other estimators are now known to also have approximate normal distributions for large sample sizes. In particular, there are central limit theorems for the maximum likelihood estimators of Sect. 5.9 and the least-squares estimators discussed in Chap. 9. This is very important, since most estimators we use will be maximum likelihood estimators or least-squares estimators. So, if we have a reasonably large sample, we can assume that these estimators have an approximately normal distribution and the normal distribution can be used for testing and constructing confidence intervals.

A.12 Bivariate Distributions

Let $f_{Y_1,Y_2}(y_1, y_2)$ be the joint density of a pair of random variables (Y_1, Y_2). Then, the *marginal density* of Y_1 is obtained by "integrating out" Y_2:

$$f_{Y_1}(y_1) = \int f_{Y_1,Y_2}(y_1, y_2)\, dy_2,$$

and similarly $f_{Y_2}(y_2) = \int f_{Y_1,Y_2}(y_1, y)\, dy_1$.

The *conditional density* of Y_2 given Y_1 is

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_1}(y_1)}. \tag{A.18}$$

Equation (A.18) can be rearranged to give the joint density of Y_1 and Y_2 as the product of a marginal density and a conditional density:

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2|Y_1}(y_2|y_1) = f_{Y_2}(y_2) f_{Y_1|Y_2}(y_1|y_2). \tag{A.19}$$

The *conditional expectation* of Y_2 given Y_1 is just the expectation calculated using $f_{Y_2|Y_1}(y_2|y_1)$:

$$E(Y_2|Y_1 = y_1) = \int y_2 f_{Y_2|Y_1}(y_2|y_1) dy_2,$$

which is, of course, a function of y_1. The conditional variance of Y_2 given Y_1 is

$$\mathrm{Var}(Y_2|Y_1 = y_1) = \int \{y_2 - E(Y_2|Y_1 = y_1)\}^2 f_{Y_2|Y_1}(y_2|y_1)\, dy_2.$$

A formula that is important elsewhere in this book is

$$\begin{aligned} f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f_{Y_1}(y_1) f_{Y_2|Y_1}(y_2|y_1) \cdots \\ f_{Y_n|Y_1,\ldots,Y_{n-1}}(y_n|y_1, \ldots, y_{n-1}), \end{aligned} \tag{A.20}$$

which follows from repeated use of (A.19).

The marginal mean and variance are related to the conditional mean and variance by

$$E(Y) = E\{E(Y|X)\} \tag{A.21}$$

and

$$\mathrm{Var}(Y) = E\{\mathrm{Var}(Y|X)\} + \mathrm{Var}\{E(Y|X)\}. \tag{A.22}$$

Result (A.21) has various names, especially the *law of iterated expectations* and the *tower rule*.

Another useful formula is that if Z is a function of X, then

$$E(ZY|X) = ZE(Y|X). \tag{A.23}$$

The idea here is that, given X, Z is constant and can be factored outside the conditional expectation.

A.13 Correlation and Covariance

Expectations and variances summarize the individual behavior of random variables. If we have two random variables, X and Y, then it is convenient to have some way to summarize their joint behavior—correlation and covariance do this.

The *covariance* between two random variables X and Y is

$$\text{Cov}(X, Y) = \sigma_{XY} = E\Big[\{X - E(X)\}\{Y - E(Y)\}\Big].$$

The two notations $\text{Cov}(X, Y)$ and σ_{XY} will be used interchangeably. If (X, Y) is continuously distributed, then using (A.36), we have

$$\sigma_{XY} = \int \{x - E(X)\}\{y - E(Y)\} f_{XY}(x, y)\, dx\, dy.$$

The following are useful formulas:

$$\sigma_{XY} = E(XY) - E(X)E(Y), \tag{A.24}$$
$$\sigma_{XY} = E[\{X - E(X)\}Y], \tag{A.25}$$
$$\sigma_{XY} = E[\{Y - E(Y)\}X], \tag{A.26}$$
$$\sigma_{XY} = E(XY) \text{ if } E(X) = 0 \text{ or } E(Y) = 0. \tag{A.27}$$

The covariance between two variables measures the linear association between them, but it is also affected by their variability; all else equal, random variables with larger standard deviations have a larger covariance. Correlation is covariance after this size effect has been removed, so that correlation is a pure measure of how closely two random variables are related, or more precisely, linearly related. The *Pearson correlation coefficient* between X and Y is

$$\text{Corr}(X, Y) = \rho_{XY} = \sigma_{XY}/\sigma_X\, \sigma_Y. \tag{A.28}$$

The Pearson correlation coefficient is sometimes called simply the correlation coefficient, though there are other types of correlation coefficients; see Sect. 8.5.

Given a bivariate sample $\{(X_i, Y_i)\}_{i=1}^n$, the sample covariance, denoted by s_{XY} or $\widehat{\sigma}_{XY}$, is

$$s_{XY} = \widehat{\sigma}_{XY} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}), \tag{A.29}$$

where $\overline{X}$ and $\overline{Y}$ are the sample means. Often the factor $(n-1)^{-1}$ is replaced by n^{-1}, but this change has little effect relative to the random variation in $\widehat{\sigma}_{XY}$. The *sample correlation* is

$$\widehat{\rho}_{XY} = r_{XY} = \frac{s_{XY}}{s_X s_Y}, \tag{A.30}$$

where s_X and s_Y are the sample standard deviations.

To provide the reader with a sense of what particular values of a correlation coefficient imply about the relationship between two random variables, Fig. A.4 shows scatterplots and the sample correlation coefficients for nine bivariate random samples. A *scatterplot* is just a plot of a bivariate sample, $\{(X_i, Y_i)\}_{i=1}^n$. Each plot also contain the *linear* least-squares fit (Chap. 9) to illustrate the linear relationship between y and x. Notice that

- an absolute correlation of 0.25 or less is weak—see panels (a) and (b);
- an absolute correlation of 0.5 is only moderately strong—see (c);
- an absolute correlation of 0.9 is strong—see (d);
- an absolute correlation of 1 implies an exact linear relationship—see (e) and (h);
- a strong nonlinear relationship may or may not imply a high correlation—see (f) and (g);
- positive correlations imply an increasing relationship (as X increases, Y increases on average)—see (b)–(e) and (g);
- negative correlations imply a decreasing relationship (as X increases, Y decreases on average)—see (h) and (i).

If the correlation between two random variables is equal to 0, then we say that they are *uncorrelated*.

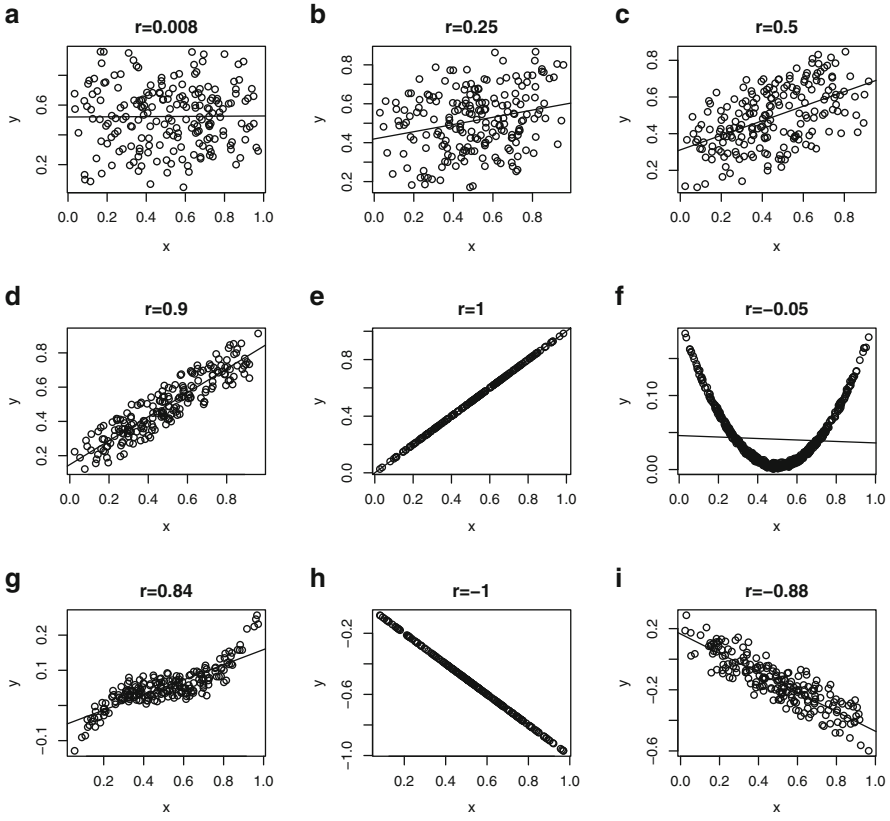

Fig. A.4. *Sample correlation coefficients for nine random samples. Each plot also contains the linear regression line of y on x.*

If X and Y are independent, then for all functions g and h,

$$E\{g(X)h(Y)\} = E\{g(X)\}E\{h(Y)\}. \tag{A.31}$$

This fact can be used to prove that if X and Y are independent, then $\sigma_{XY} = 0$, so the variables are uncorrelated. The opposite is not true. For example, if X is uniformly distributed on $[-1, 1]$ and $Y = X^2$, then a simple calculation shows that $\sigma_{XY} = 0$, but the two random variables are not independent. The key point here is that Y is related to X, in fact, completely determined by X, but the relationship is highly nonlinear and correlation measures linear association.

Another example of random variables that are uncorrelated but dependent is the bivariate t-distribution. For this distribution, the two variates are dependent even when their correlation is 0; see Sect. 7.6.

If $E(Y|X) = 0$, then Y and X are uncorrelated, since

$$E(Y) = E\{E(Y|X)\} = 0 \tag{A.32}$$

by the law of iterated expectations, and then

$$\mathrm{Cov}(Y, X) = E(YX) = E\{E(YX|X)\} = E\{XE(Y|X)\} = 0 \tag{A.33}$$

by (A.27), a second application of the law of iterated expectations, (A.23) with $Z = X$, and (A.32).

Result (A.22) has an important interpretation. If X is known and one needs to predict Y, then $E(Y|X)$ is the best predictor in that it minimizes the expected squared prediction error. If the best predictor is used, then the prediction error is $Y - E(Y|X)$ and $E\{Y - E(Y|X)\}^2$ is the expected squared prediction error. From the law of iterated expectations, that latter is

$$E\{Y - E(Y|X)\}^2 = E\Big(E\big[\{Y - E(Y|X)\}^2|X\big]\Big) = E\{\mathrm{Var}(Y|X)\}, \tag{A.34}$$

the first summand on the right-hand side of (A.22). Also, $\mathrm{Var}\{E(Y|X)\}$, the second summand there, is the variability of the best predictor and a measure of how well $E(Y|X)$ can track Y—the more $E(Y|X)$ can vary, the better it can track Y. Therefore, the sum of the tracking ability and the expected squared prediction error is the constant $\mathrm{Var}(Y)$—increasing the tracking ability decreases the expected squared prediction error.

Some insight can be gained by looking at the worst and best cases. The worst case is when X is independent of Y. Then, $E(Y|X) = E(Y)$, the tracking ability is $\mathrm{Var}\{E(Y|X)\} = 0$, and the expected squared prediction takes on its maximum value, $\mathrm{Var}(Y)$. The best case is when Y is a function of X, say $y = g(X)$ for some g. Then, $E(Y|X) = g(X) = Y$, the prediction error is 0, and the tracking ability is $\mathrm{Var}(Y)$, its maximum possible value.

A.13.1 Normal Distributions: Conditional Expectations and Variance

The calculation of conditional expectations and variances can be difficult for some probability distributions, but it is quite easy for a pair (Y_1, Y_2) that has a bivariate normal distribution.

For a bivariate normal pair, the conditional expectation of Y_2 given Y_1 equals the best linear predictor[4] of Y_2 given Y_1:

$$E(Y_2|Y_1 = y_1) = E(Y_2) + \frac{\sigma_{Y_1,Y_2}}{\sigma^2_{Y_1}}\{y_1 - E(Y_1)\}.$$

Therefore, for normal random variables, best linear prediction is the same as best prediction. Also, the conditional variance of Y_2 given Y_1 is the expected squared prediction error:

$$\mathrm{Var}(Y_2|Y_1 = y_1) = \sigma^2_{Y_2}(1 - \rho^2_{Y_1,Y_2}). \tag{A.35}$$

In general, $\mathrm{Var}(Y_2|Y_1 = y_1)$ is a function of y_1 but we see in (A.35) that for the special case of a bivariate normal distribution, $\mathrm{Var}(Y_2|Y_1 = y_1)$ is constant, that is, independent of y_1.

A.14 Multivariate Distributions

Multivariate distributions generalized the bivariate distributions of Appendix A.12. A *random vector* is a vector whose elements are random variable. A random vector of continuously distributed random variables, $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$, has a *multivariate probability density function* $f_{Y_1,\ldots,Y_d}(y_1, \ldots, y_d)$ if

$$P\{(Y_1, \ldots, Y_d) \in A\} = \int \cdots \int_A f_{Y_1,\ldots,Y_d}(y_1, \ldots, y_d)\, dy_1 \cdots dy_d$$

for all sets $A \subset \Re^p$.

The PDF of Y_j is obtained by integrating the other variates out of $f_{Y_1,\ldots,Y_d}$:

$$f_{Y_j}(y_j)$$
$$= \int_{y_1} \cdots \int_{y_{j-1}} \int_{y_{j+1}} \cdots \int_{y_d} f_{Y_1,\ldots,Y_d}(y_1, \ldots, y_d)\, dy_1 \cdots dy_{j-1} dy_{j+1} \cdots dy_d.$$

Similarly, the PDF of any subset of $(Y_1, \ldots, Y_d)$ is obtained by integrating the other variables out of $f_{Y_1,\ldots,Y_d}(y_1, \ldots, y_d)$.

The expectation of a function g of $Y_1, \ldots, Y_d$ is given by the formula

$$E\{g(Y_1, \ldots, Y_d)\} = \int_{y_1} \cdots \int_{y_d} g(y_1, \ldots, y_d) f_{Y_1,\ldots,Y_d}(y_1, \ldots, y_d)\, dy_1 \cdots dy_d. \tag{A.36}$$

[4] See Sect. 11.9.

If $Y_1, \ldots, Y_d$ are discrete, then their joint probability distribution specifies $P\{Y_1 = x_1, \ldots, Y_d = y_d\}$ for all values of $y_1, \ldots, y_d$. If $Y_1, \ldots, Y_d$ are discrete and independent, then

$$P\{Y_1 = y_1, \ldots, Y_d = y_d\} = P\{Y_1 = y_1\} \cdots P\{Y_d = y_d\}. \tag{A.37}$$

The joint CDF of $Y_1, \ldots, Y_d$, whether they are continuous or discrete, is

$$F_{Y_1,\ldots,Y_d}(x_1, \ldots, y_d) = P(Y_1 \le y_1, \ldots, Y_d \le y_d).$$

Suppose there is a sample of size n of d-dimensional random vectors, $\{\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,d}) : i = 1, \ldots, n\}$. Then the empirical CDF is

$$F_n(y_1, \ldots, y_d) = \frac{\sum_{i=1}^n I\{Y_{i,j} \le y_j, \text{ for } j = 1, \ldots, d\}}{n}. \tag{A.38}$$

A.14.1 Conditional Densities

The conditional density of $Y_1, \ldots, Y_q$ given $Y_{q+1} \ldots, Y_d$, where $1 \le q < d$, is

$$f_{Y_1,\ldots,Y_q \,|\, Y_{q+1}\ldots,Y_d}(y_1, \ldots, y_q \,|\, y_{q+1} \ldots, y_d) = \frac{f_{Y_1,\ldots,Y_d}(y_1, \ldots, y_d)}{f_{Y_{q+1}\ldots,Y_d}(y_{q+1} \ldots, y_d)}. \tag{A.39}$$

Since $Y_1, \ldots, Y_d$ can be arranged in any order that is convenient, (A.39) provides a formula for the conditional density of any subset of the variables, given the other variables. Also, (A.39) can be rearranged to give the *multiplicative formula*

$$\begin{aligned} &f_{Y_1,\ldots,Y_d}(y_1, \ldots, y_d) \\ &= f_{Y_1,\ldots,Y_q \,|\, Y_{q+1}\ldots,Y_d}(y_1, \ldots, y_q \,|\, y_{q+1} \ldots, y_d) f_{Y_{q+1}\ldots,Y_d}(y_{q+1} \ldots, y_d). \end{aligned} \tag{A.40}$$

Repeated use of (A.40) gives a formula that will be useful later for calculating likelihoods for dependent data

$$\begin{aligned} &f_{Y_1,\ldots,Y_d}(y_1, \ldots, y_d) \\ &= f_{Y_1}(y_1)\, f_{Y_2|Y_1}(y_2|y_1)\, f_{Y_3|Y_1,Y_2}(y_3|y_1, y_2) \cdots f_{Y_d|Y_1,\ldots,Y_{d-1}}(y_d|y_1, \ldots, y_{d-1}). \end{aligned} \tag{A.41}$$

If $Y_1, \ldots, Y_d$ are independent, then

$$f_{Y_1,\ldots,Y_d}(y_1, \ldots, y_d) = f_{Y_1}(y_1) \cdots f_{Y_d}(y_d). \tag{A.42}$$

A.15 Stochastic Processes

A discrete-time stochastic process is a sequence of random variables $\{Y_1, Y_2, Y_3, \ldots\}$. The distribution of Y_n is called its marginal distribution. The process is said to be Markov, or Markovian, if the conditional distribution of Y_{n+1}

given $\{Y_1, Y_2, \ldots, Y_n\}$ equals the conditional distribution of Y_{n+1} given Y_n, so Y_{n+1} depends only on the previous value of the process. The AR(1) process in Sect. 12.4 is a simple example of a Markov process. A process generated by computer simulation will be Markov if only Y_n and random numbers independent of $\{Y_1, Y_2, \ldots, Y_{n-1}\}$ are used to generate Y_{n+1}. An important example is Markov chain Monte Carlo, the topic of Sect. 20.7.

A distribution π is a stationary distribution for a Markov process if, for all n, Y_{n+1} has distribution π whenever Y_n has distribution π.

Stochastic processes can also have a continuous-time parameter. Examples are Brownian motion and geometric Brownian motion, which are used, *inter alia*, to model the log-prices and prices of equities, respectively, in continuous time.

A.16 Estimation

A.16.1 Introduction

One of the major areas of statistical inference is estimation of unknown parameters, such as a population mean, from data. An estimator is defined as any function of the observed data. The key question is which of many possible estimators should be used. If θ is an unknown parameter and $\widehat{\theta}$ is an estimator, then $E(\widehat{\theta}) - \theta$ is called the *bias* and $E\{\widehat{\theta} - \theta\}^2$ is called the *mean-squared error* (MSE). One seeks estimators that are efficient, that is, having the smallest possible value of the MSE (or of some other measure of inaccuracy). It can be shown from simple algebra that the MSE is the squared bias plus the variance, that is,

$$E\{\widehat{\theta} - \theta\}^2 = \{E(\widehat{\theta}) - \theta\}^2 + \mathrm{Var}(\widehat{\theta}), \tag{A.43}$$

so an efficient estimator will have both a small bias and a small variance. An estimator with a zero bias is called *unbiased*. However, it is not necessary to use an unbiased estimator—we only want the bias to be small, not necessarily exactly zero. One should be willing to accept a small bias if this leads to a significant reduction in variance.

The most popular methods of estimation are least squares (Sect. 9.2.1), maximum likelihood (Sects. 5.9 and 5.14), and Bayes estimation (Chap. 20).

A.16.2 Standard Errors

When a estimator is calculated from a random sample, it is a random variable, but this fact is often not appreciated by beginning students. When first exposed to statistical estimation, students tend not to think of estimators such as a sample mean as random. If we have only a single sample, then the sample mean does not *appear* random. However, if we realize that the observed

sample is only one of many possible samples that could have been drawn, and that each sample has a different sample mean, then we see that the mean is in fact random.

Since an estimator is a random variable, it has an expectation and a standard deviation. We have already seen that the difference between its expectation and the parameter is called the bias. The standard deviation of an estimator is called its *standard error*. If there are unknown parameters in the formula for this standard deviation, then they can be replaced by estimates. If $\widehat{\theta}$ is an estimator of θ, then $s_{\widehat{\theta}}$ will denote its standard error with any unknown parameters replaced by estimates.

Example A.1. The standard error of the mean

Suppose that $Y_1, \ldots, Y_n$ are i.i.d. with mean μ and variance σ^2. Then, it follows from (7.13) that the standard deviation of $\overline{Y}$ is $\sigma/\sqrt{n}$. Thus, $\sigma/\sqrt{n}$, or when σ is unknown $s_Y/\sqrt{n}$, is called the standard error of the sample mean. That is, $s_{\overline{Y}}$ is $\sigma/\sqrt{n}$ or $s_Y/\sqrt{n}$ depending on whether or not σ is known. □

A.17 Confidence Intervals

Instead of estimating an unknown parameter by a single number, it is often better to provide a range of numbers that gives a sense of the uncertainty of the estimate. Such ranges are called *interval estimates*. One type of interval estimate, the Bayesian credible interval, is introduced in Chap. 20. Another type of interval estimate is the confidence interval. A *confidence interval* is defined by the requirement that the probability that the interval will include the true parameter is a specified value called the *confidence coefficient*,, so, for example, if a large number of independent 90 % intervals are constructed, then approximately 90 % of them will contain the parameter.

A.17.1 Confidence Interval for the Mean

If $\overline{Y}$ is the mean of a sample from a normal population, then

$$\overline{Y} \pm t_{\alpha/2,n-1}\, s_{\overline{Y}} \tag{A.44}$$

is a confidence interval with $(1 - \alpha)$ confidence. This confidence interval is derived in Sect. 6.3.2. If $\alpha = 0.05$ (0.95 or 95 % confidence) and if n is reasonably large, then $t_{\alpha/2,n-1}$ is approximately 2, so $\overline{Y} \pm 2\, s_{\overline{Y}}$ is often used as an approximate 95 % confidence interval. Since $s_{\overline{Y}} = s_Y/\sqrt{n}$, the confidence can also be written as $\overline{Y} \pm 2\, s_Y/\sqrt{n}$. When n is reasonably large, say 20 or more, then $\overline{Y}$ will be approximately normally distributed by the central limit theorem, and the assumption that the population itself is normal can be dropped.

Example A.2. Confidence interval for a normal mean

Suppose we have a sample of size 25 from a normal distribution, $s_Y^2 = 2.7$, $\overline{Y} = 16.1$, and we want a 99 % confidence interval for μ. We need $t_{0.005,24}$. This quantile can be found, for example, using the `R` function `qt` and $t_{0.005,24} = 2.797$. Then, the 99 % confidence interval for μ is

$$16.1 \pm \frac{(2.797)\sqrt{2.7}}{\sqrt{25}} = 16.1 \pm 0.919 = [15.18, 17.02].$$

Since $n = 25$ is reasonably large, this interval should have approximately 99 % confidence even if the population is not normally distributed. The exception would be if the population was extremely heavily skewed or had very heavy tails; in such cases a sample size larger than 25 might be necessary for this confidence interval to have near 99 % coverage.

Just how large a sample is needed for $\overline{Y}$ to be nearly normally distributed depends on the population. If the population is symmetric and the tails are not extremely heavy, then approximate normality is often achieved with n around 10. For skewed populations, 30 observations may be needed, and even more in extreme cases. If the data appear to come from a highly skewed or heavy-tailed population, it might be better to assume a parametric model and compute the MLE as discussed in Chap. 5 and perhaps to use the bootstrap (Chap. 6) for finding the confidence interval.

The function `t.test()` computes a confidence interval for a normal mean. The output below gives a 99 % confidence interval for daily log-returns on Ford using `t.test()` and then using (A.44). The interval is (−0.000417, 0.003407)

```
> ford = read.csv("RecentFord.csv")
> returns = diff(log(ford[ , 7]))
> options(digits = 3)
> t.test(returns, conf.level = 0.99)

One Sample t-test

data:  returns
t = 2.02, df = 1256, p-value = 0.04388
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 -0.000417  0.003407
sample estimates:
mean of x
  0.00149

> n = length(returns)
> mean(returns) + c(-1, 1) *
   qt(0.995, n - 1) * sd(returns) / sqrt(n)
[1] -0.000417  0.003407
```

□

A.17.2 Confidence Intervals for the Variance and Standard Deviation

A $(1-\alpha)$ confidence interval for the variance of a normal distribution is given by

$$\left[\frac{(n-1)s_Y^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s_Y^2}{\chi^2_{1-\alpha/2,n-1}}\right], \tag{A.45}$$

where n is the sample size, s_Y^2 is the sample variance given by equation (A.7), and, as defined in Appendix A.10.1, $\chi^2_{\gamma,n-1}$ is the $(1-\gamma)$-quantile of the chi-square distribution with $n-1$ degrees of freedom.

Example A.3. Confidence interval for a normal standard deviation

Suppose we have a sample of size 25 from a normal distribution, $s_Y^2 = 2.7$, and we want a 90 % confidence interval for σ^2. The quantiles we need for constructing the interval are $\chi^2_{0.95,24} = 13.848$ and $\chi^2_{0.05,24} = 36.415$. These values can be found using software such as `qchisq()` in R. The 90 % confidence interval for σ^2 is

$$\left[\frac{(2.7)(24)}{36.415}, \frac{(2.7)(24)}{13.848}\right] = [1.78, 4.68].$$

Taking square roots of both endpoints, we get $1.33 < \sigma < 2.16$ as a 90 % confidence interval for the standard deviation.

As another example, confidence intervals for the variance and standard deviation of daily Ford returns are calculate below. The confidence interval for the standard deviation is (0.0253, 0.0273).

```
> ford = read.csv("RecentFord.csv")
> returns = diff(log(ford[,7]))
> n = length(returns)
> options(digits = 3)
> ci = (n - 1) * var(returns) / qchisq(c(0.025, 0.975), n - 1,
   lower.tail = FALSE)
> ci
[1] 0.000639 0.000748
> sqrt(ci)
[1] 0.0253 0.0273
```

□

Unfortunately, the assumption that the population is normally distributed cannot be dispensed with, even if the sample size is large. If a normal probability plot or test of normality (see Sect. 4.4) suggests that the population might be nonnormally distributed, then one might instead construct a confidence interval for σ using the bootstrap; see Chap. 6. Another possibility is to assume a nonnormal parametric model such as the t-model if the data are symmetric and heavy-tailed; see Example 5.3.

A.17.3 Confidence Intervals Based on Standard Errors

Many estimators are approximately unbiased and approximately normally distributed. Then, an approximate 95 % confidence interval is the estimator plus or minus twice its standard error; that is,

$$\widehat{\theta} \pm 2\, s_{\widehat{\theta}}$$

is an approximate 95 % confidence interval for θ.

A.18 Hypothesis Testing

A.18.1 Hypotheses, Types of Errors, and Rejection Regions

Statistical hypothesis testing uses data to decide whether a certain statement called the *null hypothesis* is true. The negation of the null hypothesis is called the *alternative hypothesis*. For example, suppose that $Y_1, \ldots, Y_n$ are i.i.d. $N(\mu, 1)$ and μ is unknown. The null hypothesis could be that μ is 1. Then, we write H_0: $\mu = 1$ and H_1: $\mu \neq 1$ to denote the null and alternative hypotheses.

There are two types of errors that we hope to avoid. If the null hypothesis is true but we reject it, then we are making a *type I error*. Conversely, if the null hypothesis is false and we accept it, then we are making a *type II error*.

The *rejection region* is the set of possible samples that lead us to reject H_0. For example, suppose that μ_0 is a hypothesized value of μ and the null hypothesis is H_0: $\mu = \mu_0$ and the alternative is H_1: $\mu \neq \mu_0$. One rejects H_0 if $|\overline{Y} - \mu_0|$ exceeds an appropriately chosen cutoff value c called a *critical value*. The rejection region is chosen to keep the probability of a type I error below a prespecified small value called the *level* and often denoted by α. Typical values of α used in practice are 0.01, 0.05, or 0.1. As α is made smaller, the rejection region must be made smaller. In the example, since we reject the null hypothesis when $|\overline{Y} - \mu_0|$ exceeds c, the critical value c gets larger as the α gets smaller. The value of c is easy to determine. Assuming that σ is known, c is $z_{\alpha/2}\, \sigma/\sqrt{n}$, where, as defined in Appendix A.9.3, $z_{\alpha/2}$ is the $\alpha/2$-upper quantile of the standard normal distribution. If σ is unknown, then σ is replaced by s_X and $z_{\alpha/2}$ is replaced by $t_{\alpha/2,n-1}$, where, as defined in Sect. 5.5.2, $t_{\alpha/2,n-1}$ is the $\alpha/2$-upper quantile of the t-distribution with $n-1$ degrees of freedom. The test using the t-quantile is called the *one-sample t-test*.

A.18.2 *p*-Values

Rather than specifying α and deciding whether to accept or reject the null hypothesis at that α, we might ask "for what values of α do we reject the null hypothesis?" The *p-value* for a sample is defined as the smallest value of α for

which the null hypothesis is rejected. Stated differently, to perform the test using a given sample, we first find the p-value of that sample, and then H_0 is rejected if we decide to use α larger than the p-value and H_0 is accepted if we use α smaller than the p-value. Thus,

- a small p-value is evidence *against* the null hypothesis

while

- a large p-value shows that the *data are consistent* with the null hypothesis.

Example A.4. Interpreting p-values

If the p-value of a sample is 0.033, then we reject H_0 if we use α equal to 0.05 or 0.1, but we accept H_0 if we use $\alpha = 0.01$. □

The p-value not only tells us whether the null hypothesis should be accepted or rejected, but it also tells us whether or not the decision to accept or reject H_0 is a close call. For example, if we are using $\alpha = 0.05$ and the p-value were 0.047, then we would reject H_0 but we would know the decision was close. If instead the p-value were 0.001, then we would know the decision was not so close.

When performing hypothesis tests, statistical software routinely calculates p-values. Doing this is much more convenient than asking the user to specify α, and then reporting whether the null hypothesis is accepted or rejected for that α.

A.18.3 Two-Sample *t*-Tests

Two-sample t-tests are used to test hypotheses about the difference between two population means. The independent-samples t-test is used when we sample independently from the two populations. Let μ_i, $\overline{Y}_i$, s_i, and n_i be the population mean, sample mean, sample standard deviation, and sample size for the ith sample, $i = 1, 2$, respectively. Let Δ_0 be a hypothesized value of $\mu_1 - \mu_2$. We assume that the two populations have the same standard deviation and estimate this parameter by the *pooled standard deviation*, which is

$$s_{\text{pool}} = \left\{ \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right\}^{1/2}. \tag{A.46}$$

The independent-samples t-statistic is

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - \Delta_0}{s_{\text{pool}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

If the hypotheses are H_0: $\mu_1 - \mu_2 = \Delta_0$ and H_1: $\mu_1 - \mu_2 \neq \Delta_0$, then H_0 is rejected if $|t| > t_{\alpha/2|n_1+n_2-2}$. If the hypotheses are H_0: $\mu_1 - \mu_2 \leq \Delta_0$ and H_1: $\mu_1 - \mu_2 > \Delta_0$, then H_0 is rejected if $t > t_{\alpha\,|n_1+n_2-2}$ and if they are H_0: $\mu_1 - \mu_2 \geq \Delta_0$ and H_1: $\mu_1 - \mu_2 < \Delta_0$, then H_0 is rejected if $t < -t_{\alpha|n_1+n_2-2}$.

Sometimes the samples are paired rather than independent. For example, suppose we wish to compare returns on small-cap versus large-cap[5] stocks and for each of n years we have the returns on a portfolio of small-cap stocks and on a portfolio of large-cap stocks. For any year, the returns on the two portfolios will be correlated, so an independent-samples test is not valid. Let $d_i = X_{i,1} - X_{i,2}$ be the difference between the observations from populations 1 and 2 for the ith pair, and let $\overline{d}$ and s_d be the sample mean and standard deviation of $d_1, \ldots, d_n$. The paired-sample t-statistics is

$$t = \frac{\overline{d} - \Delta_0}{s_d/\sqrt{n}}. \tag{A.47}$$

The rejection regions are the same as for the independent-samples t-tests except that the degrees-of-freedom parameter for the t-quantiles is $n-1$ rather than $n_1 + n_2 - 2$.

The power of a test is the probability of correctly rejecting H_0 when H_1 is true. Paired samples are often used to obtain more power. In the example of comparing small- and large-cap stocks, the returns on both portfolios will have high year-to-year variation, but the d_i will be free of this variation, so that s_d should be relatively small compared to s_1 and s_2. A small variation in the data means that $\mu_1 - \mu_2$ can be more accurately estimated and deviations of this parameter from Δ_0 are more likely to be detected.

Since $\overline{d} = \overline{Y}_1 - \overline{Y}_2$, the numerators in (A.46) and (A.47) are equal. What differs are the denominators. The denominator in (A.47) will be smaller than in (A.46) when the correlation between observations $(Y_{i,2}, Y_{i,2})$ in a pair is positive. It is the smallness of the denominator in (A.47) that gives the paired t-test increased power.

Suppose someone had a paired sample but incorrectly used the independent-samples t-test. If the correlation between $Y_{i,1}$ and $Y_{i,2}$ is zero, then the paired samples behave the same as independent samples and the effect of using the incorrect test would be small. Suppose that this correlation is positive. The result of using the incorrect test would be that if H_0 is false, then the true p-value would be overestimated and one would be less likely to reject H_0 than if the paired-sample test had been used. However, if the p-value is small, then one can be confident in rejecting H_0 because the p-value for the paired-sample test would be even smaller.[6] Unfortunately, statistical methods are often used

[5] The market capitalization of a stock is the product of the share price and the number of shares outstanding. If stocks are ranked based on market capitalization, then all stocks below some specified quantile would be small-cap stocks and all above another specified quantile would be large-cap.

[6] An exception would be the rare situation, where $Y_{i,1}$ and $Y_{i,2}$ are *negatively* correlated.

by researchers without a solid understanding of the underlying theory, and this can lead to misapplications. The hypothetical use just described of an incorrect test is often a reality, and it is sometimes necessary to evaluate whether the results that are reported can be trusted.

Confidence intervals can also be constructed for the difference between the two means and are

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{\alpha/2|n_1+n_2-2} s_{\text{pool}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{A.48}$$

for unpaired samples and

$$\overline{d} \pm t_{\alpha/2|n_1+n_2-2} s_d / \sqrt{n}. \tag{A.49}$$

for paired samples.

Example A.5. A Paired Two-sample t-test and Confidence Interval

In the next example, a 95 % confidence interval is created for the difference between the mean daily log-returns on Merck and Pfizer. Since the prices were taken over the same time intervals, the daily log-returns are highly correlated ($\widehat{\rho} = 0.547$), so a paired test and interval were used. The confidence interval was also calculated using (A.49).

```
> prices = read.csv("Stock_Bond.csv")
> prices_Merck = prices[ , 11]
> return_Merck = diff(log(prices_Merck))
> prices_Pfizer = prices[ , 13]
> return_Pfizer = diff(log(prices_Pfizer))
> cor(return_Merck,return_Pfizer)
[1] 0.547
> t.test(return_Merck, return_Pfizer, paired = TRUE)

Paired t-test

data:  return_Merck and return_Pfizer
t = -0.406, df = 4961, p-value = 0.6849
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.000584  0.000383
sample estimates:
mean of the differences
                 -1e-04

> differences = return_Merck - return_Pfizer
> n = length(differences)
> mean(differences) + c(-1,1) * qt(0.025, n - 1,
   lower.tail = FALSE) * sd(differences) / sqrt(n)
[1] -0.000584  0.000383
```

□

A.18.4 Statistical Versus Practical Significance

When we reject a null hypothesis, we often say there is a *statistically significant effect*. In this context, the word "significant" is easily misconstrued. It does *not* mean that there is an effect of practical importance. For example, suppose we were testing the null hypothesis that the means of two populations are equal versus the alternative that they are unequal. Statistical significance simply means that the two sample means are sufficiently different that this difference cannot reasonably be attributed to mere chance. Statistical significance does *not* mean that the population means are so dissimilar that their difference is of any practical importance. When large samples are used, small and unimportant effects are likely to be statistically significant.

When determining practical significance, confidence intervals are more useful than tests. In the case of the comparison between two population means, it is important to construct a confidence interval and to conclude that there is an effect of practical significance only if *all* differences in that interval are large enough to be of practical importance. How large is "large enough" is *not* a statistical question but rather must be answered by a subject-matter expert. For an example, suppose a difference between the two population means that exceeds 0.2 is considered important, at least for the purpose under consideration. If a 95 % confidence interval were $[0.23, 0.26]$, then with 95 % confidence we could conclude that there is an important difference. If instead the interval were $[0.13, 0.16]$, then we could conclude with 95 % confidence that there is no important difference. If the confidence interval were $[0.1, 0.3]$, then we could not state with 95 % confidence whether the difference is important or not.

A.19 Prediction

Suppose that Y is a random variable that is unknown at the present time, for example, a future change in an interest rate or stock price. Let $\boldsymbol{X}$ be a known random vector that is useful for predicting Y. For example, if Y is a future change in a stock price or a macroeconomic variable, $\boldsymbol{X}$ might be the vector of recent changes in that stock price or macroeconomic variable.

We want to find a function of $\boldsymbol{X}$, which we will call $\widehat{Y}(\boldsymbol{X})$, that best predicts Y. By this we mean that the mean-squared error $E[\{Y-\widehat{Y}(\boldsymbol{X})\}^2]$ is made as small as possible. The function $\widehat{Y}(\boldsymbol{X})$ that minimizes the mean-squared error will be called the best predictor of Y based on $\boldsymbol{X}$. Note that $\widehat{Y}(\boldsymbol{X})$ can be any function of $\boldsymbol{X}$, not necessarily a linear function as in Sect. 11.9.1. The *best predictor* is theoretically simple—it is the conditional expectation of Y given $\boldsymbol{X}$. That is, $E(Y|\boldsymbol{X})$ is the best predictor of Y in the sense of minimizing $E[\{Y - \widehat{Y}(\boldsymbol{X})\}^2]$ among *all* possible choices of $\widehat{Y}(\boldsymbol{X})$ that are arbitrary functions of $\boldsymbol{X}$.

If Y and $\boldsymbol{X}$ are independent, then $E(Y|\boldsymbol{X}) = E(Y)$. If $\boldsymbol{X}$ were unobserved, then $E(Y)$ would be used to predict $\boldsymbol{Y}$. Thus, when Y and $\boldsymbol{X}$ are independent, the best predictor of Y is the same as if $\boldsymbol{X}$ were unknown, because $\boldsymbol{X}$ contains no information that is useful for prediction of Y.

In practice, using $E(Y|\boldsymbol{X})$ for prediction is not trivial. The problem is that $E(Y|\boldsymbol{X})$ may be difficult to estimate whereas the best linear predictor can be estimated by linear regression as described in Chap. 9. However, the newer technique of *nonparametric regression* can be used to estimate $E(Y|\boldsymbol{X})$. Nonparametric regression is discussed in Chap. 21.

A.20 Facts About Vectors and Matrices

The norm of the vector $\boldsymbol{x} = (x_1, \ldots, x_p)^\mathsf{T}$ is $\|\boldsymbol{x}\| = (\sum_{i=1}^p x_i^2)^{1/2}$.

A square matrix $\boldsymbol{A}$ is diagonal if $A_{i,j} = 0$ for all $i \neq j$. We use the notation $\text{diag}(d_1, \ldots, d_p)$ for a $p \times p$ diagonal matrix $\boldsymbol{A}$ such that $A_{i,i} = d_i$.

A matrix $\boldsymbol{O}$ is orthogonal if $\boldsymbol{O}^\mathsf{T} = \boldsymbol{O}^{-1}$. This implies that the columns of $\boldsymbol{O}$ are mutually orthogonal (perpendicular) and that their norms are all equal to 1.

Any symmetric matrix $\boldsymbol{\Sigma}$ has an *eigenvalue-eigenvector decomposition*, eigen-decomposition for short, which is

$$\boldsymbol{\Sigma} = \boldsymbol{O}\,\text{diag}(\lambda_i)\,\boldsymbol{O}^\mathsf{T}, \tag{A.50}$$

where $\boldsymbol{O}$ is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}$ and $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\boldsymbol{\Sigma}$. Also, if all of $\lambda_1, \ldots, \lambda_p$ are nonzero, then $\boldsymbol{\Sigma}$ is nonsingular and

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{O}\,\text{diag}(1/\lambda_i)\,\boldsymbol{O}^\mathsf{T}.$$

Let $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_p$ be the columns of $\boldsymbol{O}$. Then, since $\boldsymbol{O}$ is orthogonal,

$$\boldsymbol{o}_j^\mathsf{T}\boldsymbol{o}_k = 0 \tag{A.51}$$

for any $j \neq k$. Moreover,

$$\boldsymbol{o}_j^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{o}_k = 0 \tag{A.52}$$

for $j \neq k$. To see this, let $\boldsymbol{e}_j$ be the jth unit vector, that is, the vector with a one in the jth coordinate and zeros elsewhere. Then, $\boldsymbol{o}_j^\mathsf{T}\boldsymbol{O} = \boldsymbol{e}_j^\mathsf{T}$ and $\boldsymbol{O}^\mathsf{T}\boldsymbol{o}_k = \boldsymbol{e}_k$, so that for $j \neq k$,

$$\boldsymbol{o}_j^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{o}_k = \boldsymbol{o}_j^\mathsf{T}\left\{\boldsymbol{O}\,\text{diag}(\lambda_i)\,\boldsymbol{O}^\mathsf{T}\right\}\boldsymbol{o}_k = \lambda_j\lambda_k\boldsymbol{e}_j^\mathsf{T}\boldsymbol{e}_k = 0.$$

The eigenvalue-eigenvector decomposition of a covariance matrix is used in Sect. 7.8 to find the orientation of elliptically contoured densities. This decomposition can be important even if the density is not elliptically contoured and is the basis of principal components analysis (PCA).

Example A.6. An Eigendecomposition

In the next example, a 3×3 symmetric matrix `Sigma` is created and its eigenvalues and eigenvectors are computed using the function `eigen()`. The eigenvalues are in the vector `decomp$values` and the eigenvectors are in the matrix `decomp$vectors`. It is also verified that `decomp$vectors` is an orthogonal matrix.

```
> Sigma = matrix(c(1, 3, 4, 3, 6, 2, 4, 2, 8), nrow = 3,
   byrow = TRUE)
> Sigma
     [,1] [,2] [,3]
[1,]    1    3    4
[2,]    3    6    2
[3,]    4    2    8
> decomp = eigen(Sigma)
> decomp
$values
[1] 11.59  4.79 -1.37

$vectors
       [,1]    [,2]   [,3]
[1,] -0.426 -0.0479  0.903
[2,] -0.499 -0.8203 -0.279
[3,] -0.754  0.5699 -0.326

> round(decomp$vectors %*% t(decomp$vectors), 5)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

□

A.21 Roots of Polynomials and Complex Numbers

The roots of polynomials play an important role in the study of ARMA processes. Let $p(x) = b_0 + b_1 x + \cdots b_p x^p$, with $b_p \neq 0$, be a pth-degree polynomial. The fundamental theorem of algebra states that $p(x)$ can be factored as

$$b_p(x - r_1)(x - r_2) \cdots (x - r_p),$$

where $r_1, \ldots, r_p$ are the roots of $p(x)$, that is, the solutions to $p(x) = 0$. The roots need not be distinct and they can be complex numbers. In `R`, the roots of a polynomial can be found using the function `polyroot()`.

A complex number can be written as $a + b\imath$, where $\imath = \sqrt{-1}$. The absolute value or magnitude of $a + b\imath$ is $\sqrt{a^2 + b^2}$. The complex plane is the set of all two-dimensional vectors (a, b), where (a, b) represents the complex number $a + b\imath$. The unit circle is the set of all complex number with magnitude 1.

A complex number is inside or outside the unit circle depending on whether its magnitude is less than or greater than 1.

Example A.7. Roots of a Cubic Polynomial

As an example, the roots of the cubic polynomial $1 + 2x + 3x^2 + 4x^3$ are computed below. We see that there is one real root, -0.606 and two complex roots, $-0.072 \pm 0.638i$. It is also verified that these are roots.

```
> roots = polyroot(c(1, 2, 3, 4))
> roots
[1] -0.072+0.638i -0.606-0.000i -0.072-0.638i
> fn = function(x){1 + 2 * x + 3 * x^2 + 4 * x^3}
> round(fn(roots), 5)
[1] 0+0i 0+0i 0+0i
```

□

A.22 Bibliographic Notes

Casella and Berger (2002) covers in greater detail most of the statistical theory in this chapter and elsewhere in the book. Wasserman (2004) is a modern introduction to statistical theory and is also recommended for further study. Alexander (2001) is a recent introduction to financial econometrics and has a chapter on covariance matrices; her technical appendices cover maximum likelihood estimation, confidence intervals, and hypothesis testing, including likelihood ratio tests. Evans, Hastings, and Peacock (1993) provides a concise reference for the basic facts about commonly used distributions in statistics. Johnson, Kotz, and Kemp (1993) discusses most of the common discrete distributions, including the binomial. Johnson, Kotz, and Balakrishnan (1994, 1995) contain a wealth of information and extensive references about the normal, lognormal, chi-square, exponential, uniform, t, F, Pareto, and many other continuous distributions. Together, these works by Johnson, Kotz, Kemp, and Balakrishnan are essentially an encyclopedia of statistical distributions.

References

Alexander, C. (2001) *Market Models: A Guide to Financial Data Analysis*, Wiley, Chichester.

Casella, G. and Berger, R. L. (2002) *Statistical Inference*, 2nd ed., Duxbury/Thomson Learning, Pacific Grove, CA.

Evans, M., Hastings, N., and Peacock, B. (1993) *Statistical Distributions*, 2nd ed., Wiley, New York.

Gourieroux, C., and Jasiak, J. (2001) *Financial Econometrics*, Princeton University Press, Princeton, NJ.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994) *Continuous Univariate Distributions, Vol. 1*, 2nd ed., Wiley, New York.
Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995) *Continuous Univariate Distributions, Vol. 2*, 2nd ed., Wiley, New York.
Johnson, N. L., Kotz, S., and Kemp, A. W. (1993) *Discrete Univariate Distributions*, 2nd ed., Wiley, New York.
Wasserman, L. (2004) *All of Statistics*, Springer, New York.

Index

D. Ruppert, D.S. Matteson, *Statistics and Data Analysis for Financial Engineering*, Springer Texts in Statistics,
DOI 10.1007/978-1-4939-2614-5