The probability density function of a continuous random variable Y (or the probability mass function if Y is discrete) is referred to simply as a **probability distribution** and denoted by

$$f(y;\boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ represents the parameters of the distribution.

We use dot $(\cdot)$ subscripts for summation and bars $(^{-})$ for means, thus

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\, y\cdot .$$

The expected value and variance of a random variable Y are denoted by $\mathrm{E}(Y)$ and $\mathrm{var}(Y)$ respectively. Suppose random variables $Y_1, ..., Y_N$ are independent with $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \sigma_i^2$ for $i = 1, ..., n$. Let the random variable W be a **linear combination** of the Y_i's

$$W = a_1Y_1 + a_2Y_2 + ... + a_nY_n, \tag{1.1}$$

where the a_i's are constants. Then the expected value of W is

$$\mathrm{E}(W) = a_1\mu_1 + a_2\mu_2 + ... + a_n\mu_n \tag{1.2}$$

and its variance is

$$\mathrm{var}(W) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + ... + a_n^2\sigma_n^2. \tag{1.3}$$

1.4 Distributions related to the Normal distribution

The sampling distributions of many of the estimators and test statistics used in this book depend on the Normal distribution. They do so either directly because they are derived from Normally distributed random variables, or asymptotically, via the Central Limit Theorem for large samples. In this section we give definitions and notation for these distributions and summarize the relationships between them. The exercises at the end of the chapter provide practice in using these results which are employed extensively in subsequent chapters.

1.4.1 Normal distributions

1. If the random variable Y has the Normal distribution with mean μ and variance σ^2, its probability density function is

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma^2}\right)^2\right].$$

   We denote this by $Y \sim N(\mu, \sigma^2)$.

2. The Normal distribution with $\mu = 0$ and $\sigma^2 = 1$, $Y \sim N(0,1)$, is called the **standard Normal distribution**.

3. Let $Y_1, ..., Y_n$ denote Normally distributed random variables with $Y_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, ..., n$ and let the covariance of Y_i and Y_j be denoted by

$$\text{cov}(Y_i, Y_j) = \rho_{ij}\sigma_i\sigma_j \ ,$$

where ρ_{ij} is the correlation coefficient for Y_i and Y_j. Then the joint distribution of the Y_i's is the **multivariate Normal distribution** with mean vector $\boldsymbol{\mu} = [\mu_1, ..., \mu_n]^T$ and variance-covariance matrix $\mathbf{V}$ with diagonal elements σ_i^2 and non-diagonal elements $\rho_{ij}\sigma_i\sigma_j$ for $i \neq j$. We write this as $\mathbf{y} \sim \mathbf{N}(\boldsymbol{\mu}, \mathbf{V})$, where $\mathbf{y} = [Y_1, ..., Y_n]^T$.

4. Suppose the random variables $Y_1, ..., Y_n$ are independent and Normally distributed with the distributions $Y_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, ..., n$. If

$$W = a_1Y_1 + a_2Y_2 + ... + a_nY_n,$$

where the a_i's are constants. Then W is also Normally distributed, so that

$$W = \sum_{i=1}^{n} a_iY_i \sim N\left(\sum_{i=1}^{n} a_i\mu_i, \ \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$$

by equations (1.2) and (1.3).

1.4.2 Chi-squared distribution

1. The **central chi-squared distribution** with n degrees of freedom is defined as the sum of squares of n independent random variables $Z_1, ..., Z_n$ each with the standard Normal distribution. It is denoted by

$$X^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n).$$

In matrix notation, if $\mathbf{z} = [Z_1, ..., Z_n]^T$ then $\mathbf{z}^T\mathbf{z} = \sum_{i=1}^{n} Z_i^2$ so that $X^2 = \mathbf{z}^T\mathbf{z} \sim \chi^2(n)$.

2. If X^2 has the distribution $\chi^2(n)$, then its expected value is $\text{E}(X^2) = n$ and its variance is $\text{var}(X^2) = 2n$.

3. If $Y_1, ..., Y_n$ are independent Normally distributed random variables each with the distribution $Y_i \sim N(\mu_i, \sigma_i^2)$ then

$$X^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2(n) \qquad (1.4)$$

because each of the variables $Z_i = (Y_i - \mu_i)/\sigma_i$ has the standard Normal distribution $N(0, 1)$.

4. Let $Z_1, ..., Z_n$ be independent random variables each with the distribution $N(0, 1)$ and let $Y_i = Z_i + \mu_i$, where at least one of the μ_i's is non-zero. Then the distribution of

$$\sum Y_i^2 = \sum (Z_i + \mu_i)^2 = \sum Z_i^2 + 2\sum Z_i\mu_i + \sum \mu_i^2$$

has larger mean $n+\lambda$ and larger variance $2n+4\lambda$ than $\chi^2(n)$ where $\lambda = \sum \mu_i^2$. This is called the **non-central chi-squared distribution** with n degrees of freedom and **non-centrality parameter** λ. It is denoted by $\chi^2(n,\lambda)$.

5. Suppose that the Y_i's are not necessarily independent and the vector $\mathbf{y} = [Y_1,\ldots,Y_n]^T$ has the multivariate normal distribution $\mathbf{y} \sim \mathbf{N}(\boldsymbol{\mu},\mathbf{V})$ where the variance-covariance matrix $\mathbf{V}$ is non-singular and its inverse is $\mathbf{V}^{-1}$. Then
$$X^2 = (\mathbf{y}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \sim \chi^2(n). \tag{1.5}$$
6. More generally if $\mathbf{y} \sim \mathbf{N}(\boldsymbol{\mu},\mathbf{V})$ then the random variable $\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}$ has the non-central chi-squared distribution $\chi^2(n,\lambda)$ where $\lambda = \boldsymbol{\mu}^T\mathbf{V}^{-1}\boldsymbol{\mu}$.
7. If $X_1^2,\ldots,X_m^2$ are m independent random variables with the chi-squared distributions $X_i^2 \sim \chi^2(n_i,\lambda_i)$, which may or may not be central, then their sum also has a chi-squared distribution with $\sum n_i$ degrees of freedom and non-centrality parameter $\sum \lambda_i$, i.e.,
$$\sum_{i=1}^m X_i^2 \sim \chi^2\left(\sum_{i=1}^m n_i, \sum_{i=1}^m \lambda_i\right).$$
This is called the **reproductive property** of the chi-squared distribution.
8. Let $\mathbf{y} \sim \mathbf{N}(\boldsymbol{\mu},\mathbf{V})$, where $\mathbf{y}$ has n elements but the Y_i's are not independent so that $\mathbf{V}$ is singular with rank $k<n$ and the inverse of $\mathbf{V}$ is not uniquely defined. Let $\mathbf{V}^-$denote a generalized inverse of $\mathbf{V}$. Then the random variable $\mathbf{y}^T\mathbf{V}^-\mathbf{y}$ has the non-central chi-squared distribution with k degrees of freedom and non-centrality parameter $\lambda = \boldsymbol{\mu}^T\mathbf{V}^-\boldsymbol{\mu}$.

   For further details about properties of the chi-squared distribution see Rao (1973, Chapter 3).

1.4.3 t-distribution

The **t-distribution** with n degrees of freedom is defined as the ratio of two independent random variables. The numerator has the standard Normal distribution and the denominator is the square root of a central chi-squared random variable divided by its degrees of freedom; that is,
$$T = \frac{Z}{(X^2/n)^{1/2}} \tag{1.6}$$
where $Z \sim N(0,1)$, $X^2 \sim \chi^2(n)$ and Z and X^2 are independent. This is denoted by $T \sim t(n)$.

1.4.4 F-distribution

1. The **central F-distribution** with n and m degrees of freedom is defined as the ratio of two independent central chi-squared random variables each

divided by its degrees of freedom,

$$F = \frac{X_1^2}{n} / \frac{X_2^2}{m} \tag{1.7}$$

where $X_1^2 \sim \chi^2(n), X_2^2 \sim \chi^2(m)$ and X_1^2 and X_2^2 are independent. This is denoted by $F \sim F(n, m)$.

2. The relationship between the t-distribution and the F-distribution can be derived by squaring the terms in equation (1.6) and using definition (1.7) to obtain

$$T^2 = \frac{Z^2}{1} / \frac{X^2}{n} \sim F(1, n), \tag{1.8}$$

that is, the square of a random variable with the t-distribution, $t(n)$, has the F-distribution, $F(1, n)$.

3. The **non-central F-distribution** is defined as the ratio of two independent random variables, each divided by its degrees of freedom, where the numerator has a non-central chi-squared distribution and the denominator has a central chi-squared distribution, i.e.,

$$F = \frac{X_1^2}{n} / \frac{X_2^2}{m}$$

where $X_1^2 \sim \chi^2(n, \lambda)$ with $\lambda = \boldsymbol{\mu}^T \mathbf{V}^{-1} \boldsymbol{\mu}$, $X_2^2 \sim \chi^2(m)$ and X_1^2 and X_2^2 are independent. The mean of a non-central F-distribution is larger than the mean of central F-distribution with the same degrees of freedom.

1.5 Quadratic forms

1. A **quadratic form** is a polynomial expression in which each term has degree 2. Thus $y_1^2 + y_2^2$ and $2y_1^2 + y_2^2 + 3y_1 y_2$ are quadratic forms in y_1 and y_2 but $y_1^2 + y_2^2 + 2y_1$ or $y_1^2 + 3y_2^2 + 2$ are not.

2. Let $\mathbf{A}$ be a symmetric matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

where $a_{ij} = a_{ji}$, then the expression $\mathbf{y}^T \mathbf{A} \mathbf{y} = \sum_i \sum_j a_{ij} y_i y_j$ is a quadratic form in the y_i's. The expression $(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ is a quadratic form in the terms $(y_i - \mu_i)$ but not in the y_i's.

3. The quadratic form $\mathbf{y}^T \mathbf{A} \mathbf{y}$ and the matrix $\mathbf{A}$ are said to be **positive definite** if $\mathbf{y}^T \mathbf{A} \mathbf{y} > 0$ whenever the elements of $\mathbf{y}$ are not all zero. A necessary and sufficient condition for positive definiteness is that all the determinants

$$|A_1| = a_{11}, |A_2| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, |A_3| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \ldots, \text{ and}$$

$|A_n| = \det \mathbf{A}$ are all positive.

4. The rank of the matrix $\mathbf{A}$ is also called the degrees of freedom of the quadratic form $Q = \mathbf{y}^T \mathbf{A} \mathbf{y}$.
5. Suppose $Y_1, ..., Y_n$ are independent random variables each with the Normal distribution $N(0, \sigma^2)$. Let $Q = \sum_{i=1}^{n} Y_i^2$ and let $Q_1, ..., Q_k$ be quadratic forms in the Y_i's such that

$$Q = Q_1 + ... + Q_k$$

where Q_i has m_i degrees of freedom $(i = 1, \ldots, k)$. Then

$Q_1, ..., Q_k$ are independent random variables and

$Q_1/\sigma^2 \sim \chi^2(m_1)$, $Q_2/\sigma^2 \sim \chi^2(m_2), \cdots$ and $Q_k/\sigma^2 \sim \chi^2(m_k)$,

if and only if,

$$m_1 + m_2 + ... + m_k = n.$$

This is Cochran's theorem; for a proof see, for example, Hogg and Craig (1995). A similar result holds for non-central distributions; see Chapter 3 of Rao (1973).

6. A consequence of Cochran's theorem is that the difference of two independent random variables, $X_1^2 \sim \chi^2(m)$ and $X_2^2 \sim \chi^2(k)$, also has a chi-squared distribution

$$X^2 = X_1^2 - X_2^2 \sim \chi^2(m - k)$$

provided that $X^2 \geq 0$ and $m > k$.

1.6 Estimation

1.6.1 Maximum likelihood estimation

Let $\mathbf{y} = [Y_1, ..., Y_n]^T$ denote a random vector and let the joint probability density function of the Y_i's be

$$f(\mathbf{y}; \boldsymbol{\theta})$$

which depends on the vector of parameters $\boldsymbol{\theta} = [\theta_1, ..., \theta_p]^T$.

The **likelihood function** $L(\boldsymbol{\theta}; \mathbf{y})$ is algebraically the same as the joint probability density function $f(\mathbf{y}; \boldsymbol{\theta})$ but the change in notation reflects a shift of emphasis from the random variables $\mathbf{y}$, with $\boldsymbol{\theta}$ fixed, to the parameters $\boldsymbol{\theta}$ with $\mathbf{y}$ fixed. Since L is defined in terms of the random vector $\mathbf{y}$, it is itself a random variable. Let Ω denote the set of all possible values of the parameter vector $\boldsymbol{\theta}$; Ω is called the **parameter space**. The **maximum likelihood estimator** of $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ which maximizes the likelihood function, that is

$$L(\widehat{\boldsymbol{\theta}}; \mathbf{y}) \geq L(\boldsymbol{\theta}; \mathbf{y}) \qquad \text{for all } \boldsymbol{\theta} \text{ in } \Omega.$$

Equivalently, $\widehat{\boldsymbol{\theta}}$ is the value which maximizes the **log-likelihood function**

$l(\boldsymbol{\theta}; \mathbf{y}) = \log L(\boldsymbol{\theta}; \mathbf{y})$, since the logarithmic function is monotonic. Thus

$$l(\widehat{\boldsymbol{\theta}}; \mathbf{y}) \geq l(\boldsymbol{\theta}; \mathbf{y}) \qquad \text{for all } \boldsymbol{\theta} \text{ in } \Omega.$$

Often it is easier to work with the log-likelihood function than with the likelihood function itself.

Usually the estimator $\widehat{\boldsymbol{\theta}}$ is obtained by differentiating the log-likelihood function with respect to each element θ_j of $\boldsymbol{\theta}$ and solving the simultaneous equations

$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_j} = 0 \qquad \text{for } j = 1, ..., p. \tag{1.9}$$

It is necessary to check that the solutions do correspond to maxima of $l(\boldsymbol{\theta}; \mathbf{y})$ by verifying that the matrix of second derivatives

$$\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_j \partial \theta_k}$$

evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ is negative definite. For example, if $\boldsymbol{\theta}$ has only one element θ this means it is necessary to check that

$$\left[\frac{\partial^2 l(\theta, \ \mathbf{y})}{\partial \theta^2} \right]_{\theta = \widehat{\theta}} < 0.$$

It is also necessary to check if there are any values of $\boldsymbol{\theta}$ at the edges of the parameter space Ω that give local maxima of $l(\boldsymbol{\theta}; \mathbf{y})$. When all local maxima have been identified, the value of $\widehat{\boldsymbol{\theta}}$ corresponding to the largest one is the maximum likelihood estimator. (For most of the models considered in this book there is only one maximum and it corresponds to the solution of the equations $\partial l / \partial \theta_j = 0, j = 1, ..., p$.)

An important property of maximum likelihood estimators is that if $g(\boldsymbol{\theta})$ is any function of the parameters $\boldsymbol{\theta}$, then the maximum likelihood estimator of $g(\boldsymbol{\theta})$ is $g(\widehat{\boldsymbol{\theta}})$. This follows from the definition of $\widehat{\boldsymbol{\theta}}$. It is sometimes called the **invariance property** of maximum likelihood estimators. A consequence is that we can work with a function of the parameters that is convenient for maximum likelihood estimation and then use the invariance property to obtain maximum likelihood estimates for the required parameters.

In principle, it is not necessary to be able to find the derivatives of the likelihood or log-likelihood functions or to solve equation (1.9) if $\widehat{\boldsymbol{\theta}}$ can be found numerically. In practice, numerical approximations are very important for generalized linear models.

Other properties of maximum likelihood estimators include consistency, sufficiency, asymptotic efficiency and asymptotic normality. These are discussed in books such as Cox and Hinkley (1974) or Kalbfleisch (1985, Chapters 1 and 2).

1.6.2 Example: Poisson distribution

Let $Y_1, ..., Y_n$ be independent random variables each with the Poisson distribution

$$f(y_i;\theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}, \qquad y_i = 0, 1, 2, ...$$

with the same parameter θ. Their joint distribution is

$$\begin{aligned} f(y_1, \ldots, y_n;\theta) &= \prod_{i=1}^{n} f(y_i;\theta) = \frac{\theta^{y_1} e^{-\theta}}{y_1!} \times \frac{\theta^{y_2} e^{-\theta}}{y_2!} \times \cdots \times \frac{\theta^{y_n} e^{-\theta}}{y_n!} \\ &= \frac{\theta^{\Sigma y_i}\ e^{-n\theta}}{y_1!y_2!...y_n!}. \end{aligned}$$

This is also the likelihood function $L(\theta; y_1, ..., y_n)$. It is easier to use the log-likelihood function

$$l(\theta; y_1, ..., y_n) = \log L(\theta; y_1, ..., y_n) = (\sum y_i)\log\theta - n\theta - \sum(\log y_i!).$$

To find the maximum likelihood estimate $\widehat{\theta}$, use

$$\frac{dl}{d\theta} = \frac{1}{\theta}\sum y_i - n.$$

Equate this to zero to obtain the solution

$$\widehat{\theta} = \sum y_i/n = \overline{y}.$$

Since $d^2l/d\theta^2 = -\sum y_i/\theta^2 < 0$, l has its maximum value when $\theta = \widehat{\theta}$, confirming that $\overline{y}$ is the maximum likelihood estimate.

1.6.3 Least Squares Estimation

Let $Y_1, ..., Y_n$ be independent random variables with expected values $\mu_1, ..., \mu_n$ respectively. Suppose that the μ_i's are functions of the parameter vector that we want to estimate, $\boldsymbol{\beta} = [\beta_1, ..., \beta_p]^T,\ p < n$. Thus

$$E(Y_i) = \mu_i(\boldsymbol{\beta}).$$

The simplest form of the **method of least squares** consists of finding the estimator $\widehat{\boldsymbol{\beta}}$ that minimizes the sum of squares of the differences between Y_i's and their expected values

$$S = \sum \left[Y_i - \mu_i(\boldsymbol{\beta})\right]^2.$$

Usually $\widehat{\boldsymbol{\beta}}$ is obtained by differentiating S with respect to each element β_j of $\boldsymbol{\beta}$ and solving the simultaneous equations

$$\frac{\partial S}{\partial \beta_j} = 0, \qquad j = 1, ..., p.$$

Of course it is necessary to check that the solutions correspond to minima (i.e., the matrix of second derivatives is positive definite) and to identify the global minimum from among these solutions and any local minima at the boundary of the parameter space.

Now suppose that the Y_i's have variances σ_i^2 that are not all equal. Then it may be desirable to minimize the weighted sum of squared differences

$$S = \sum w_i \left[Y_i - \mu_i\left(\boldsymbol{\beta}\right)\right]^2$$

where the weights are $w_i = (\sigma_i^2)^{-1}$. In this way, the observations which are less reliable (that is, the Y_i's with the larger variances) will have less influence on the estimates.

More generally, let $\mathbf{y} = [Y_1, ..., Y_n]^T$ denote a random vector with mean vector $\boldsymbol{\mu} = [\mu_1, ..., \mu_n]^T$ and variance-covariance matrix $\mathbf{V}$. Then the **weighted least squares estimator** is obtained by minimizing

$$S = (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}).$$

1.6.4 Comments on estimation.

1. An important distinction between the methods of maximum likelihood and least squares is that the method of least squares can be used without making assumptions about the distributions of the response variables Y_i beyond specifying their expected values and possibly their variance-covariance structure. In contrast, to obtain maximum likelihood estimators we need to specify the joint probability distribution of the Y_i's.
2. For many situations maximum likelihood and least squares estimators are identical.
3. Often numerical methods rather than calculus may be needed to obtain parameter estimates that maximize the likelihood or log-likelihood function or minimize the sum of squares. The following example illustrates this approach.

1.6.5 Example: Tropical cyclones

Table 1.2 shows the number of tropical cyclones in Northeastern Australia for the seasons 1956-7 (season 1) to 1968-9 (season 13), a period of fairly consistent conditions for the definition and tracking of cyclones (Dobson and Stewart, 1974).

Table 1.2 *Numbers of tropical cyclones in 13 successive seasons.*

Season:	1	2	3	4	5	6	7	8	9	10	11	12	13
No. of cyclones	6	5	4	6	6	3	12	7	4	2	6	7	4

Let Y_i denote the number of cyclones in season i, where $i = 1, \ldots, 13$. Suppose the Y_i's are independent random variables with the Poisson distribution

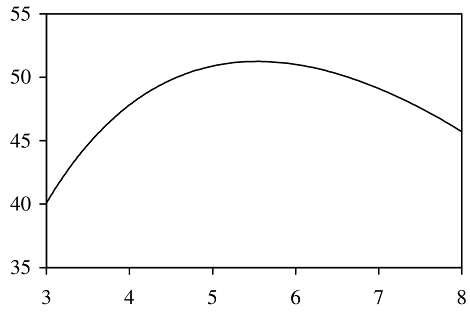

Figure 1.1 *Graph showing the location of the maximum likelihood estimate for the data in Table 1.2 on tropical cyclones.*

with parameter θ. From Example 1.6.2 $\widehat{\theta} = \overline{y} = 72/13 = 5.538$. An alternative approach would be to find numerically the value of θ that maximizes the log-likelihood function. The component of the log-likelihood function due to y_i is

$$l_i = y_i \log\theta - \theta - \log y_i!.$$

The log-likelihood function is the sum of these terms

$$l = \sum_{i=1}^{13} l_i = \sum_{i=1}^{13} \left(y_i \log\theta - \theta - \log y_i!\right).$$

Only the first two terms in the brackets involve θ and so are relevant to the optimization calculation, because the term $\sum_1^{13} \log y_i!$ is a constant. To plot the log-likelihood function (without the constant term) against θ, for various values of θ, calculate $(y_i \log\theta - \theta)$ for each y_i and add the results to obtain $l^* = \sum (y_i \log\theta - \theta)$. Figure 1.1 shows l^* plotted against θ.

Clearly the maximum value is between $\theta = 5$ and $\theta = 6$. This can provide a starting point for an iterative procedure for obtaining $\widehat{\theta}$. The results of a simple bisection calculation are shown in Table 1.3. The function l^* is first calculated for approximations $\theta^{(1)} = 5$ and $\theta^{(2)} = 6$. Then subsequent approximations $\theta^{(k)}$ for $k = 3, 4, ...$ are the average values of the two previous estimates of θ with the largest values of l^*(for example, $\theta^{(6)} = \frac{1}{2}(\theta^{(5)} + \theta^{(3)})$). After 7 steps this process gives $\widehat{\theta} \simeq 5.54$ which is correct to 2 decimal places.

1.7 Exercises

1.1 Let Y_1 and Y_2 be independent random variables with

$Y_1 \sim N(1,3)$ and $Y_2 \sim N(2,5)$. If $W_1 = Y_1 + 2Y_2$ and $W_2 = 4Y_1 - Y_2$ what is the joint distribution of W_1 and W_2?

1.2 Let Y_1 and Y_2 be independent random variables with

$Y_1 \sim N(0,1)$ and $Y_2 \sim N(3,4)$.

Table 1.3 *Successive approximations to the maximum likelihood estimate of the mean number of cyclones per season.*

k	$\theta^{(k)}$	l^*
1	5	50.878
2	6	51.007
3	5.5	51.242
4	5.75	51.192
5	5.625	51.235
6	5.5625	51.243
7	5.5313	51.24354
8	5.5469	51.24352
9	5.5391	51.24360
10	5.5352	51.24359

(a) What is the distribution of Y_1^2?

(b) If $\mathbf{y} = \begin{bmatrix} Y_1 \\ (Y_2 - 3)/2 \end{bmatrix}$, obtain an expression for $\mathbf{y}^T\mathbf{y}$. What is its distribution?

(c) If $\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ and its distribution is $\mathbf{y} \sim \mathbf{N}(\boldsymbol{\mu}, \mathbf{V})$, obtain an expression for $\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}$. What is its distribution?

1.3 Let the joint distribution of Y_1 and Y_2 be $\mathbf{N}(\boldsymbol{\mu}, \mathbf{V})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} 4 & 1 \\ 1 & 9 \end{pmatrix}.$$

(a) Obtain an expression for $(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})$. What is its distribution?

(b) Obtain an expression for $\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}$. What is its distribution?

1.4 Let $Y_1, ..., Y_n$ be independent random variables each with the distribution $N(\mu, \sigma^2)$. Let

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

(a) What is the distribution of $\overline{Y}$?

(b) Show that $S^2 = \dfrac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i - \mu)^2 - n(\overline{Y} - \mu)^2\right]$.

(c) From (b) it follows that $\sum(Y_i - \mu)^2/\sigma^2 = (n-1)S^2/\sigma^2 + \left[(\overline{Y} - \mu)^2 n/\sigma^2\right]$. How does this allow you to deduce that $\overline{Y}$ and S^2 are independent?

(d) What is the distribution of $(n-1)S^2/\sigma^2$?

(e) What is the distribution of $\dfrac{\overline{Y} - \mu}{S/\sqrt{n}}$?

Table 1.4 *Progeny of light brown apple moths.*

Progeny group	Females	Males
1	18	11
2	31	22
3	34	27
4	33	29
5	27	24
6	33	29
7	28	25
8	23	26
9	33	38
10	12	14
11	19	23
12	25	31
13	14	20
14	4	6
15	22	34
16	7	12

1.5 This exercise is a continuation of the example in Section 1.6.2 in which $Y_1, ..., Y_n$ are independent Poisson random variables with the parameter θ.

(a) Show that $\mathrm{E}(Y_i) = \theta$ for $i = 1, ..., n$.

(b) Suppose $\theta = e^{\beta}$. Find the maximum likelihood estimator of β.

(c) Minimize $S = \sum \left(Y_i - e^{\beta}\right)^2$ to obtain a least squares estimator of β.

1.6 The data below are the numbers of females and males in the progeny of 16 female light brown apple moths in Muswellbrook, New South Wales, Australia (from Lewis, 1987).

(a) Calculate the proportion of females in each of the 16 groups of progeny.

(b) Let Y_i denote the number of females and n_i the number of progeny in each group ($i = 1, ..., 16$). Suppose the Y_i's are independent random variables each with the binomial distribution

$$f(y_i; \theta) = \binom{n_i}{y_i} \theta^{y_i} (1 - \theta)^{n_i - y_i}.$$

Find the maximum likelihood estimator of θ using calculus and evaluate it for these data.

(c) Use a numerical method to estimate $\widehat{\theta}$ and compare the answer with the one from (b).