J.G. Kalbfleisch

Probability and Statistical Inference

Volume 2: Statistical Inference

Second Edition

With 27 Illustrations

Springer-Verlag
New York Berlin Heidelberg Tokyo

Springer Texts in Statistics

Advisors:
Stephen Fienberg Ingram Olkin

J.G. Kalbfleisch
University of Waterloo
Department of Statistics and Actuarial Science
Waterloo, Ontario, N2L 3G1
Canada

AMS Classification: 62-01

Library of Congress Cataloging in Publication Data
Kalbfleisch, J.G.
Probability and statistical inference.
(Springer texts in statistics)
Includes indexes.
Contents: v. 1. Probability—v. 2. Statistical inference.
1. Probabilities. 2. Mathematical statistics.
I. Title. II. Series.
QA273.K27 1985 519.5'4 85-12580

The first edition was published in two volumes,

Probability and Statistical Inference I (Universitext)
Probability and Statistical Inference II (Universitext).

Typeset by H. Charlesworth & Co. Ltd., Huddersfield, England.
Printed and bound by R.R. Donnelley and Sons, Harrisonburg, Virginia.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-96183-6 Springer-Verlag New York Berlin Heidelberg Tokyo
ISBN 3-540-96183-6 Springer-Verlag Berlin Heidelberg New York Tokyo

Preface

This book is in two volumes, and is intended as a text for introductory courses in probability and statistics at the second or third year university level. It emphasizes applications and logical principles rather than mathematical theory. A good background in freshman calculus is sufficient for most of the material presented. Several starred sections have been included as supplementary material. Nearly 900 problems and exercises of varying difficulty are given, and Appendix A contains answers to about one-third of them.

The first volume (Chapters 1–8) deals with probability models and with mathematical methods for describing and manipulating them. It is similar in content and organization to the 1979 edition. Some sections have been rewritten and expanded—for example, the discussions of independent random variables and conditional probability. Many new exercises have been added.

In the second volume (Chapters 9–16), probability models are used as the basis for the analysis and interpretation of data. This material has been revised extensively. Chapters 9 and 10 describe the use of the likelihood function in estimation problems, as in the 1979 edition. Chapter 11 then discusses frequency properties of estimation procedures, and introduces coverage probability and confidence intervals. Chapter 12 describes tests of significance, with applications primarily to frequency data. The likelihood ratio statistic is used to unify the material on testing, and connect it with earlier material on estimation. Chapters 13 and 14 present methods for analyzing data under the assumption of normality, with emphasis on the importance of correctly modelling the experimental situation. Chapter 15 considers sufficient statistics and conditional tests, and Chapter 16 presents some additional topics in statistical inference.

The content of volume two is unusual for an introductory text. The importance of the probability model is emphasized, and general techniques are presented for derivi ıg suitable estimates, intervals, and tests from the likelihood function.

The intention is to avoid the appearance of a recipe book with many special formulas set out for type problems. A wide variety of applications can be treated using the methods presented, particularly if students have access to computing facilities.

I have omitted much of the standard material on optimality criteria for estimators and tests, which is better left for later courses in mathematical statistics. Also, I have avoided using decision-theoretic language. For instance, I discuss the calculation and interpretation of the observed significance level, rather than presenting the formal theory of hypothesis testing. In most statistical applications, the aim is to learn from the data at hand, not to minimize error frequencies in a long sequence of decisions.

I wish to thank my colleagues and students at the University of Waterloo for their helpful comments on the 1979 edition, and on earlier drafts of this edition. Special thanks are due to Professor Jock MacKay for his many excellent suggestions, and to Ms. Lynda Hohner for superb technical typing. Finally, I wish to express my appreciation to my wife Rebecca, and children Jane, David, and Brian, for their encouragement and support.

I am grateful to the Biometrika trustees for permission to reproduce material from Table 8 of *Biometrika Tables for Statisticians*, Vol. 1 (3rd edition, 1966); to John Wiley and Sons Inc. for permission to reproduce portions of Table II from *Statistical Tables and Formulas* by D. Hald (1952); and to the Literary Executor of the late Sir Ronald Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to Longman Group Ltd., London, for permission to reprint Tables I, III, and V from their book *Statistical Tables for Biological, Agricultural, and Medical Research* (6th edition, 1974).

J.G. Kalbfleisch

Contents of Volume 2

Contents of Volume 1

CHAPTER 9

Likelihood Methods

The first volume dealt with probability models, and with mathematical methods for handling and describing them. Several of the simplest discrete and continuous probability models were considered in detail. This volume is concerned with applications of probability models in problems of data analysis and interpretation.

One important use of probability models is to provide simple mathematical descriptions of large bodies of data. For instance, we might describe a set of 1000 blood pressure measurements as being like a sample of 1000 independent values from a normal distribution whose mean μ and variance σ^2 are estimated from the data. This model gives a concise description of the data, and from it we can easily calculate the approximate proportion of blood pressure measurements which lie in any particular range. The accuracy of such calculations will, of course, depend upon how well the normal distribution model fits the data.

We shall be concerned primarily with applications of probability models in problems of statistical inference, where it is desired to draw general conclusions based on a limited amount of data. For instance, tests might be run to determine the length of life of an aircraft component prior to failure from metal fatigue. Such tests are typically very expensive and time consuming, and hence only a few specimens can be examined. Based on the small amount of data obtained, one would attempt to draw conclusions about similar components which had not been tested. The link between the observed sample and the remaining components is provided by the probability model. The data are used to check the adequacy of the model and to estimate any unknown parameters which it involves. General statements concerning this type of component are then based on the model.

A limited amount of data can be misleading, and therefore any general

conclusions drawn will be subject to uncertainty. Measurement of the extent of this uncertainty is an important part of the problem. An estimate is of little value unless we know how accurate it is likely to be.

In statistical inference problems, we usually start with a set of data, and with some information about the way in which the data were collected. We then attempt to formulate a probability model for the experiment which gave rise to the data. Examination of the data, and of other similar data sets, can be very useful at this stage. It is important to treat the data set in context, and to take full advantage of what is already known from other similar applications.

Usually the probability model will involve one or more unknown parameters which must be estimated from the data. We have already encountered this problem on several occasions, and have used the observed sample mean as an estimate of the mean of a Poisson or exponential distribution. Intuitively, this is a reasonable thing to do, but intuition may fail us in more complicated situations.

Section 9.1 introduces the method of maximum likelihood, which provides a routine procedure for obtaining estimates of unknown parameters. Section 2 considers the problem of estimating an unknown parameter θ on the basis of data from two independent experiments. Section 3 shows how the relative likelihood function may be used to rank possible values of θ according to their plausibilities.

Section 4 describes likelihood methods when the probability model is continuous. The special case of censoring in lifetime experiments is considered in Section 5. Section 6 discusses the invariance property of likelihood methods, and Section 7 describes a normal approximation to the log relative likelihood function. The use of Newton's method in finding maximum likelihood estimates and likelihood intervals is illustrated in Section 8.

In this chapter it is assumed that the probability model involves only one unknown parameter. Likelihood methods for the estimation of two or more unknown parameters are described in Chapter 10. Some theoretical properties of these estimation procedures are considered in Chapter 11.

Chapter 12 introduces tests of significance, which are used to investigate whether various hypotheses of interest are consistent with the data. Several applications of significance tests to frequency data are given.

Traditionally, the normal distribution has played a very important role in statistical applications. Chapters 13 and 14 develop estimation procedures and significance tests for a variety of situations where measurements are assumed to be independent and normally distributed. Finally, Chapters 15 and 16 deal with some more advanced topics in statistical inference.

9.1 The Method of Maximum Likelihood

Suppose that a probability model has been formulated for an experiment, and that it involves a single unknown parameter θ. The experiment is performed and some data are obtained. We wish to use the data to estimate the value of θ. More generally, we wish to determine which of the possible values of θ are plausible or likely in the light of the observations.

The observed data can be regarded as an event E in the sample space for the probability model. The probability of event E can be determined from the model, and in general it will be a function of the unknown parameter, $P(E; \theta)$. The *maximum likelihood estimate* (*MLE*) of θ is the value of θ which maximizes $P(E; \theta)$. The MLE of θ is usually denoted by $\hat{\theta}$. It is the parameter value which best explains the data E in the sense that it maximizes the probability of E under the model.

EXAMPLE 9.1.1. Suppose that we wish to estimate θ, the proportion of people with tuberculosis in a large homogeneous population. To do this, we randomly select n individuals for testing, and find that x of them have the disease. Since the population is large and homogeneous, we assume that the n individuals tested are independent, and that each has probability θ of having tuberculosis. The probability of the observed event (data) is then

$$P(E; \theta) = P(x \text{ out of } n \text{ have tuberculosis})$$
$$= \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{9.1.1}$$

where $0 \le \theta \le 1$. The maximum likelihood estimate $\hat{\theta}$ is the value of θ which maximizes (9.1.1). We shall show later that (9.1.1) is maximized for $\theta = \frac{x}{n}$, and so the MLE of θ is $\hat{\theta} = \frac{x}{n}$. To maximize the probability of the data we estimate θ, the proportion of diseased persons in the population, by $\frac{x}{n}$, the proportion of diseased persons in the sample.

The Likelihood and Log Likelihood Functions

Note that the constant factor $\binom{n}{x}$ will have no effect on the maximization of (9.1.1) over θ. To simplify expressions, we shall generally omit such constants and consider only the part of $P(E; \theta)$ which involves θ.

The *likelihood function* of θ is defined as follows:

$$L(\theta) = c \cdot P(E; \theta). \tag{9.1.2}$$

Here c is any positive constant with respect to θ; that is, c is not a function of θ, although it may be a function of the data. We choose c to obtain a simple expression for $L(\theta)$, and subsequent results will not depend upon the specific choice made.

Usually $P(E;\theta)$ and $L(\theta)$ are products of terms, and it will be more convenient to work with logarithms. The *log likelihood function* is the natural logarithm of L:

$$l(\theta) = \log L(\theta). \tag{9.1.3}$$

Note that, by (9.1.2),

$$l(\theta) = c' + \log P(E;\theta)$$

where $c' = \log c$ is not a function of θ.

The maximum likelihood estimate $\hat{\theta}$ is the value of θ which maximizes $P(E;\theta)$. The value of θ which maximizes $P(E;\theta)$ will also maximize $L(\theta)$ and $l(\theta)$. Thus the MLE $\hat{\theta}$ is the value of θ which maximizes the likelihood function and the log likelihood function. Usually it is easiest to work with the log likelihood function.

EXAMPLE 9.1.1 (continued). The likelihood function of θ is any constant c times the expression for $P(E;\theta)$ in (9.1.1), where c may depend on n and x but not on θ. Since the aim in choosing c is to simplify the expression, a natural choice is $c = 1\Big/\binom{n}{x}$, and then

$$L(\theta) = \theta^x(1-\theta)^{n-x} \qquad \text{for } 0 \le \theta \le 1.$$

The log likelihood function is now

$$l(\theta) = x\log\theta + (n-x)\log(1-\theta) \qquad \text{for } 0 \le \theta \le 1.$$

The MLE $\hat{\theta}$ is the value of θ which maximizes $l(\theta)$.

The Score and Information Functions

To evaluate $\hat{\theta}$, we need to locate the maximum of $l(\theta)$ over all possible values of θ. This can usually be done by differentiating $l(\theta)$ with respect to θ, setting the derivative equal to zero, and solving for θ. It is possible that this procedure might yield a relative minimum or point of inflexion instead of the maximum desired. Thus it is necessary to verify that a maximum has been found, perhaps by checking that the second derivative is negative.

The *score function* $S(\theta)$ is defined to be the first derivative of the log likelihood function with respect to θ:

$$S(\theta) = l'(\theta) = \frac{dl(\theta)}{d\theta}; \tag{9.1.4}$$

The *information function* $\mathcal{I}(\theta)$ is minus the second derivative of the log likelihood function with respect to θ:

$$\mathcal{I}(\theta) = -l''(\theta) = -S'(\theta) = -\frac{d^2l(\theta)}{d\theta^2}. \tag{9.1.5}$$

Note that neither $S(\theta)$ nor $\mathcal{I}(\theta)$ depends on the choice of c in (9.1.2).

The set Ω of possible values of θ is called the *parameter space*. Usually Ω is an interval of real values, such as $[0, 1]$ in the example above, and the first and second derivatives of $l(\theta)$ with respect to θ exist at all interior points of Ω. Then, if $\hat{\theta}$ is an interior point of Ω, the first derivative will be zero and the second derivative will be negative at $\theta = \hat{\theta}$. Thus under these conditions we have

$$S(\hat{\theta}) = 0; \qquad \mathcal{I}(\hat{\theta}) > 0. \tag{9.1.6}$$

To find $\hat{\theta}$, we determine the roots of the *maximum likelihood equation* $S(\theta) = 0$. We then verify, by checking the sign of $\mathcal{I}(\theta)$ or otherwise, that a relative maximum has been found.

In some simple examples, the maximum likelihood equation $S(\theta) = 0$ can be solved algebraically to yield a formula for $\hat{\theta}$. In more complicated situations, it will be necessary to solve this equation numerically (see Section 9.8).

Situations do arise in which $\hat{\theta}$ cannot be found by solving the maximum likelihood equation $S(\theta) = 0$. For instance $S(\hat{\theta})$ need not be zero if the overall maximum of $l(\theta)$ occurs on a boundary of the parameter space Ω (see Examples 9.1.1 and 9.1.2). The same is true if θ is restricted to a discrete set of possible values such as the integers (see Problems 9.1.7 and 9.1.11).

EXAMPLE 9.1.1 (continued). For this example, the score and information functions are

$$S(\theta) = \frac{dl(\theta)}{d\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta} \qquad \text{for } 0 < \theta < 1;$$

$$\mathcal{I}(\theta) = -\frac{dS(\theta)}{d\theta} = \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2} \qquad \text{for } 0 < \theta < 1.$$

For $1 \le x \le n-1$, the maximum likelihood equation $S(\theta) = 0$ has a unique solution $\theta = \frac{x}{n}$. Since $\mathcal{I}(\theta) > 0$ at $\theta = \frac{x}{n}$, the likelihood function has a relative maximum at $\theta = \frac{x}{n}$. Furthermore, since $L(\theta) = 0$ for $\theta = 0$ and for $\theta = 1$, we have found the overall maximum, and thus $\hat{\theta} = \frac{x}{n}$.

If $x = 0$, the equation $S(\theta) = 0$ has no solution, and the maximum occurs on a boundary of the parameter space $\Omega = [0, 1]$. In this case we have

$$P(E;\theta) = (1-\theta)^n \qquad \text{for } 0 \le \theta \le 1,$$

which is clearly largest when $\theta = 0$. Thus $\hat{\theta} = 0$ when $x = 0$. Similarly we find that $\hat{\theta} = 1$ when $x = n$, and the formula $\hat{\theta} = \dfrac{x}{n}$ holds for all x.

EXAMPLE 9.1.2. Some laboratory tests are run on samples of river water in order to determine whether the water is safe for swimming. Of particular interest is the concentration of coliform bacteria in the water. The number of coliform bacteria is determined for each of n unit-volume samples of river water, giving n observed counts $x_1, x_2, \ldots, x_n$. The problem is to estimate μ, the average number of coliform bacteria per unit volume in the river.

We assume that the bacteria are distributed randomly and uniformly throughout the river water, so that the assumptions of a Poisson process (Section 4.4) are satisfied. Then the probability of observing x_i bacteria in a sample of unit volume is given by a Poisson distribution:

$$f(x_i) = \mu^{x_i} e^{-\mu}/x_i! \qquad \text{for } x_i = 0, 1, 2, \ldots .$$

Since disjoint volumes are independent, the probability of the n observed counts $x_1, x_2, \ldots, x_n$ is

$$P(E; \mu) = f(x_1) f(x_2) \ldots f(x_n)$$
$$= \prod_{i=1}^{n} \frac{\mu^{x_i} e^{-\mu}}{x_i!} = \frac{\mu^{\Sigma x_i} e^{-n\mu}}{x_1!\, x_2! \ldots x_n!}.$$

The likelihood function is $c \cdot P(E; \mu)$ where c is any constant not depending upon μ. We choose c to simplify the expression for $L(\mu)$, and a natural choice here is

$$c = 1/x_1!\, x_2! \ldots x_n!.$$

For this choice of c, the likelihood function is

$$L(\mu) = \mu^{\Sigma x_i} e^{-n\mu} \qquad \text{for } 0 \le \mu < \infty,$$

and the log-likelihood function is

$$l(\mu) = \Sigma x_i \log \mu - n\mu.$$

The score and information functions are

$$S(\mu) = \frac{1}{\mu} \Sigma x_i - n; \qquad \mathcal{I}(\mu) = \Sigma x_i / \mu^2.$$

These functions will be the same no matter how the constant c is chosen.

If $\Sigma x_i > 0$, the maximum likelihood equation $S(\mu) = 0$ has a unique solution $\mu = \dfrac{1}{n} \Sigma x_i = \bar{x}$. Since $\mathcal{I}(\mu) > 0$ at $\mu = \bar{x}$, we have found a relative maximum. Furthermore, since $L(0) = 0$ and $L(\mu) \to 0$ as $\mu \to \infty$, this must be the overall maximum.

If $\Sigma x_i = 0$, the equation $S(\mu) = 0$ has no solution, and the maximum occurs on the boundary of the parameter space: $\hat{\mu} = 0$.

In both cases, the MLE is $\hat{\mu} = \bar{x}$. To maximize the probability of the data $x_1, x_2, \ldots, x_n$, we estimate the population mean μ by the sample mean $\bar{x}$.

EXAMPLE 9.1.3. It is usually not possible to count the number of bacteria in a sample of river water; one can only determine whether or not *any* are present. n test tubes each containing a volume v of river water are incubated and tested. A negative test shows that there were no bacteria present, while a positive test tube shows that at least one bacterium was present. If y tubes out of the n tested give negative results, what is the maximum likelihood estimate of μ?

SOLUTION. The probability that there are x bacteria in a volume v of river water is given by a Poisson distribution with mean μv:

$$f(x) = (\mu v)^x e^{-\mu v}/x!; \qquad x = 0, 1, 2, \ldots .$$

The probability of a negative reaction (no bacteria) is

$$p = f(0) = e^{-\mu v};$$

the probability of a positive reaction (at least one bacterium) is

$$1 - p = 1 - e^{-\mu v}.$$

Since disjoint volumes are independent, the n test tubes constitute independent trials. The probability of observing y negative reactions out of n is therefore

$$P(E; \mu) = \binom{n}{y} p^y (1-p)^{n-y}$$

where $p = e^{-v\mu}$ and $0 \le \mu < \infty$.

The likelihood function is $c \cdot P(E; \mu)$ where c does not depend upon μ and is chosen to give a simple expression for $L(\mu)$. Taking $c = 1 \Big/ \binom{n}{y}$, we have

$$L(\mu) = p^y (1-p)^{n-y}$$

where $p = e^{-v\mu}$ and $0 \le \mu < \infty$.

Since $p = e^{-v\mu}$, it follows that $\mu = -\dfrac{1}{v} \log p$. From Example 9.1.1, the function $p^y(1-p)^{n-y}$ is maximized for $\hat{p} = \dfrac{y}{n}$. The corresponding value of μ is

$$\hat{\mu} = -\frac{1}{v} \log \hat{p} = -\frac{1}{v} \log \frac{y}{n} = \frac{\log n - \log y}{v}.$$

Here we have used the invariance property of likelihood (see Section 9.6).

For instance, suppose that 40 test tubes each containing 10 ml of river water are incubated. If 28 give negative tests and 12 give positive tests, then

$$\hat{\mu} = \frac{\log 40 - \log 28}{10} = 0.0357.$$

The concentration of bacteria in the river is estimated to be 0.0357 per ml.

The greater the concentration of bacteria in the river, the more probable it is that all n test tubes will give positive results. Hence the larger the value of μ, the more probable the observation $y = 0$. If we observe $y = 0$, the MLE of μ will be $+\infty$. In this case, it does not make much practical sense to give merely a single estimate of μ. What we require is an indication of the range of μ-values which are plausible in the light of the data, rather than a single "most plausible" value. This can be obtained by examining the relative likelihood function (see Section 3). □

Likelihoods Based on Frequency Tables

Data from n independent repetitions of an experiment are often summarized in a frequency table:

Event or class	A_1	A_2	...	A_k	Total
Observed frequency	f_1	f_2		f_k	n
Expected frequency	np_1	np_2	...	np_k	n

The sample space S for a single repetition of the experiment is partitioned into k mutually exclusive classes or events, $S = A_1 \cup A_2 \cup \cdots \cup A_k$. Then f_j is the number of times that A_j occurs in n repetitions ($\Sigma f_j = n$). Let p_j be the probability of event A_j in any one repetition ($\Sigma p_j = 1$). The p_j's can be determined from the probability model. If the model involves an unknown parameter θ, the p_j's will generally be functions of θ.

The probability of observing a particular frequency table is given by the multinomial distribution

$$P(E; \theta) = \binom{n}{f_1 f_2 \ldots f_k} p_1^{f_1} p_2^{f_2} \ldots p_k^{f_k}.$$

The likelihood function is

$$L(\theta) = c p_1^{f_1} p_2^{f_2} \ldots p_k^{f_k} \tag{9.1.7}$$

where we have absorbed the multinomial coefficient into the constant c. The MLE $\hat{\theta}$ is the value of θ which maximizes (9.1.7). Using $\hat{\theta}$, we can compute estimated expected frequencies $n\hat{p}_j$ for comparison with the observed frequencies f_j.

EXAMPLE 9.1.4. On each of 200 consecutive working days, ten items were randomly selected from a production line and tested for imperfections, with the following results:

Number of defective items	0	1	2	3	≥4	Total
Frequency observed	133	52	12	3	0	200

The number of defective items out of 10 is thought to have a binomial distribution. Find the MLE of θ, the probability that an item is defective, and compute estimated expected frequencies under the binomial distribution model.

SOLUTION. According to a binomial distribution model, the probability of observing j defectives out of 10 is

$$p_j = \binom{10}{j} \theta^j (1-\theta)^{10-j}; \qquad j = 0, 1, 2, \ldots, 10.$$

The probability of observing 4 or more defectives is $p_{4+} = 1 - p_0 - p_1 - p_2 - p_3$. By (9.1.7), the likelihood function of θ is

$$L(\theta) = c p_0^{133} p_1^{52} p_2^{12} p_3^3 p_{4+}^0 \qquad \text{for } 0 \le \theta \le 1,$$

where c is any convenient positive constant. Taking

$$c = 1 \Big/ \binom{10}{0}^{133} \binom{10}{1}^{52} \binom{10}{2}^{12} \binom{10}{3}^3,$$

we obtain

$$\begin{aligned} L(\theta) &= [(1-\theta)^{10}]^{133} [\theta(1-\theta)^9]^{52} [\theta^2(1-\theta)^8]^{12} [\theta^3(1-\theta)^7]^3 \\ &= \theta^{85}(1-\theta)^{1915}. \end{aligned}$$

This likelihood function is of the form considered in Example 9.1.1, with $x = 85$ and $n = 2000$. Hence $\hat{\theta} = \dfrac{85}{2000} = 0.0425$.

The estimated probability for class $j = 0$ is

$$\hat{p}_0 = \binom{10}{0} \hat{\theta}^0 (1-\hat{\theta})^{10-0} = (1 - 0.0425)^{10} = 0.6477$$

and the estimated expected frequency for this class is

$$n\hat{p}_0 = 200(0.6477) = 129.54.$$

Similarly we can compute $n\hat{p}_j$ for $j = 1, 2, 3$ and then find the estimated expected frequency for the last class by subtraction from 200. The results are as follows:

Number of defectives	0	1	2	3	≥4	Total
Observed frequency	133	52	12	3	0	200
Expected frequency	129.54	57.50	11.48	1.36	0.12	200

The agreement between observed and estimated expected frequencies appears to be reasonably good.

Since the items are chosen at random, we would not observe exactly the same results if we repeated the experiment. According to the model, the f_j's

are observed values of random variables. Some differences between the observed and expected frequencies will occur owing to chance variation in the f_j's. A test of significance (Chapter 12) may be used to verify that the differences found here are not too great to be accounted for by chance variation, and hence the binomial model is satisfactory. □

PROBLEMS FOR SECTION 9.1

1.† Suppose that diseased trees are distributed randomly and uniformly throughout a large forest with an average of λ per acre. The numbers of diseased trees observed in ten four-acre plots were 0, 1, 3, 0, 0, 2, 2, 0, 1, 1. Find the maximum likelihood estimate of λ.

2. Suppose that the n counts in Example 9.1.2 were summarized in a frequency table as follows:

Number of bacteria	0	1	2 ...	Total
Frequency observed	f_0	f_1	f_2 ...	n

The number of bacteria in a sample is assumed to have a Poisson distribution with mean μ. Find the likelihood function and maximum likelihood estimate of μ based on the frequency table, and show that they agree with the results obtained in Example 9.1.2.

3. Consider the following two experiments whose purpose is to estimate θ, the fraction of a large population having blood type A.

(i) Individuals are selected at random until 10 with blood type A are obtained. The total number of people examined is found to be 100.
(ii) 100 individuals are selected at random, and it is found that 10 of them have blood type A.

Show that the two experiments lead to proportional likelihood functions, and hence the same MLE for θ.

4.† According to genetic theory, blood types MM, NM, and NN should occur in a very large population with relative frequencies θ^2, $2\theta(1-\theta)$, and $(1-\theta)^2$, where θ is the (unknown) gene frequency.

(a) Suppose that, in a random sample of size n from the population, there are x_1, x_2, and x_3 of the three types. Find an expression for $\hat{\theta}$.
(b) The observed frequencies in a sample of size 100 were 32, 46, and 22, respectively. Compute $\hat{\theta}$ and the estimated expected frequencies for the three blood types under the model.

5. A brick-shaped die (Example 1.3.2) is rolled n times, and the ith face comes up x_i times ($i = 1, 2, \ldots, 6$), where $\Sigma x_i = n$.

(a) Show that $\hat{\theta} = (3t - 2n)/12n$, where $t = x_1 + x_2 + x_3 + x_4$.
(b) Suppose that the observed frequencies are 11, 15, 13, 15, 22, 24. Compute estimated expected frequencies under the model.

6. A sample of n items is examined from each large batch of a mass-produced article. The number of good items in a sample has a binomial distribution with parameters n and p. The batch is accepted if all n items are good, and is rejected otherwise. Out of m batches, x are accepted and $m - x$ are rejected. Find the maximum likelihood estimate of p.

7.† "The enemy" has an unknown number N of tanks, which he has obligingly numbered 1, 2, ..., N. Spies have reported sighting 8 tanks with numbers 137, 24, 86, 33, 92, 129, 17, 111. Assume that sightings are independent, and that each of the N tanks has probability $1/N$ of being observed at each sighting. Show that $\hat{N} = 137$.

8. Blood samples from nk people are analyzed to obtain information about θ, the fraction of the population infected with a certain disease. In order to save time, the nk samples are mixed together k at a time to give n pooled samples. The analysis of a pooled sample will be negative if the k individuals are free from the disease, and positive otherwise. Out of the n pooled samples, x give negative results and $n - x$ give positive results. Find an expression for $\hat{\theta}$.

9.† Specimens of a new high-impact plastic are tested by repeatedly striking them with a hammer until they fracture. If the specimen has a constant probability θ of surviving a blow, independently of the number of previous blows received, the number of blows required to fracture a specimen will have a geometric distribution,

$$f(x) = \theta^{x-1}(1-\theta) \qquad \text{for } x = 1, 2, 3, \ldots .$$

The results of tests on 200 specimens were as follows:

Number of blows required	1	2	3	≥4	Total
Number of specimens	112	36	22	30	200

Find the maximum likelihood estimate of θ, and compute estimated expected frequencies.

10. The n progeny in a breeding experiment are of three types, there being x_i of the ith type ($i = 1, 2, 3$). According to a genetic model, the proportions of the three types should be $(2+p)/4$, $(1-p)/2$, and $p/4$, and progeny are independent of one another.

(a) Show that $\hat{p}$ is a root of the quadratic equation

$$np^2 + (2x_2 + x_3 - x_1)p - 2x_3 = 0.$$

(b) Suppose that $x_1 = 58$, $x_2 = 33$, and $x_3 = 9$. Find $\hat{p}$, and compute estimated expected frequencies under the model.

11. An urn contains r red balls and b black balls, where r is known but b is unknown. Of n balls chosen at random without replacement, x were red and y were black ($x + y = n$).

(a) Show that $L(b)$ is proportional to $b^{(y)}/(r+b)^{(n)}$.

(b) Show that $\dfrac{L(b+1)}{L(b)} = \dfrac{(b+1)(r+b-n+1)}{(b-y+1)(r+b+1)}$.

(c) By considering the conditions under which $L(b+1)/L(b)$ exceeds one, show that $\hat{b}$ is the smallest integer which exceeds $\dfrac{nr}{x} - (r+1)$. When is $\hat{b}$ not unique?

12. For a certain mass-produced article, the proportion of defectives is θ. It is customary to inspect a sample of 3 items from each large batch. Records are kept only for those samples which contain at least one defective item.

(a) Show that the conditional probability that a sample contains i defectives, given that it contains at least one defective, is

$$\binom{3}{i}\theta^i(1-\theta)^{3-i}/[1-(1-\theta)^3] \qquad (i=1,2,3).$$

(b) Suppose that x_i samples out of n recorded contain i defectives ($i=1, 2, 3$; $\Sigma x_i = n$). Show that $\hat{\theta}$ is the smaller root of the quadratic equation

$$t\theta^2 - 3t\theta + 3(t-n) = 0$$

where $t = x_1 + 2x_2 + 3x_3$.

13.† Leaves of a plant are examined for insects. The number of insects on a leaf is thought to have a Poisson distribution with mean μ, except that many leaves have no insects because they are unsuitable for feeding and not merely because of the chance variation allowed by the Poisson law. The empty leaves are therefore not counted.

(a) Find the conditional probability that a leaf contains i insects, given that it contains at least one.

(b) Suppose that x_i leaves are observed with i insects ($i=1, 2, 3, \ldots$), where $\Sigma x_i = n$. Show that the MLE of μ satisfies the equation

$$\hat{\mu} = \bar{x}(1 - e^{-\hat{\mu}})$$

where $\bar{x} = \Sigma i x_i/n$.

(c) Determine $\hat{\mu}$ numerically for the case $\bar{x} = 3.2$.

14. In Problem 9.1.12, suppose that samples of size $k > 3$ are examined, and that x_i of those recorded contain i defectives ($i = 1, 2, \ldots, k$; $\Sigma x_i = n$).

(a) Show that the MLE of θ satisfies the equation

$$\bar{x}[1-(1-\theta)^k] - k\theta = 0$$

where $\bar{x} = \Sigma i x_i/n$.

(b) Use the binomial theorem to show that, if $\hat{\theta}$ is small, then

$$\hat{\theta} \approx 2(\bar{x}-1)/(k-1)\bar{x}.$$

(c) Solve for $\hat{\theta}$ in the case $k = 5$, $\bar{x} = 1.12$.

9.2. Combining Independent Experiments

Suppose that two independent experiments both give information about the same parameter θ. Experiment 1 gives data E_1 with probability $P(E_1; \theta)$. The likelihood function of θ based on the first experiment is

$$L_1(\theta) = c_1 \cdot P(E_1; \theta)$$

where c_1 is any positive constant. Similarly, experiment 2 gives data E_2, and the likelihood function is

$$L_2(\theta) = c_2 \cdot P(E_2; \theta).$$

We wish to obtain the likelihood function of θ based on both sets of data, E_1 and E_2.

As in Section 3.2, we consider the two experiments as components of a single composite experiment. The sample space for the composite experiment is a Cartesian product, and the data from the composite experiment correspond to the event E_1E_2 in this space. The probability of the data is $P(E_1E_2; \theta)$, and the likelihood function based on both experiments is

$$L(\theta) = c \cdot P(E_1E_2; \theta)$$

where c is any positive constant.

Since the experiments are independent, we have

$$P(E_1E_2; \theta) = P(E_1; \theta) \cdot P(E_2; \theta).$$

It follows that

$$L(\theta) = c' \cdot L_1(\theta) \cdot L_2(\theta)$$

where $c' = c/c_1c_2$ is any positive constant.

In the examples of the last section we chose the proportionality constant c in (9.1.2) to simplify the expression for $L(\theta)$. We noted that $S(\theta)$, $\mathcal{I}(\theta)$, and $\hat{\theta}$ are unaffected by the choice of c. In the same spirit, we can take $c' = 1$ above. For this choice of c' we have

$$L(\theta) = L_1(\theta) \cdot L_2(\theta), \tag{9.2.1}$$

and taking natural logarithms on both sides gives

$$l(\theta) = l_1(\theta) + l_2(\theta). \tag{9.2.2}$$

To combine information about θ from two or more independent experiments, we multiply the likelihood functions, or add the log likelihood functions.

It follows from (9.2.2) and (9.1.4) that

$$S(\theta) = S_1(\theta) + S_2(\theta). \tag{9.2.3}$$

The score function for the composite experiment is the sum of the score functions for the independent components. Similarly, (9.2.3) and (9.1.5) give

$$\mathcal{I}(\theta) = \mathcal{I}_1(\theta) + \mathcal{I}_2(\theta). \tag{9.2.4}$$

Let $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}$ be the MLE's of θ based on just the first experiment, on just the second experiment, and on both experiments, respectively. Thus $\hat{\theta}_1$ maximizes $l_1(\theta)$, $\hat{\theta}_2$ maximizes $l_2(\theta)$, and $\hat{\theta}$ maximizes $l(\theta)$. Except in special cases, it is not possible to compute $\hat{\theta}$ from just $\hat{\theta}_1$ and $\hat{\theta}_2$. One has to add log likelihoods using (9.2.2) and then remaximize to get $\hat{\theta}$.

If $\hat{\theta}_1 = \hat{\theta}_2$, then both terms on the right hand side of (9.2.2) attain their maxima at the same value of θ, and hence $\hat{\theta} = \hat{\theta}_1 = \hat{\theta}_2$. Otherwise, the overall maximum will usually lie between $\hat{\theta}_1$ and $\hat{\theta}_2$.

If the estimates $\hat{\theta}_1, \hat{\theta}_2$ were quite different, it would usually be unwise to combine results from the two experiments to obtain a single overall estimate. Instead the results from the two experiments should be reported separately, and an explanation of the difference should be sought. For further discussion see Example 9.3.2 and Section 12.3.

EXAMPLE 9.2.1. Suppose that, in Example 9.1.1, m additional people are randomly selected, and y of them are found to have tuberculosis. Find the MLE of θ based on both sets of data.

SOLUTION. For the first experiment, the log likelihood function is

$$l_1(\theta) = x \log \theta + (n - x) \log(1 - \theta), \tag{9.2.5}$$

and the maximum likelihood estimate is $\hat{\theta}_1 = \dfrac{x}{n}$. For the second experiment, we similarly obtain

$$l_2(\theta) = y \log \theta + (m - y) \log(1 - \theta),$$

and $\hat{\theta}_2 = \dfrac{y}{m}$. Because the population is large, the two samples will be very nearly independent, and hence by (9.2.2), the log likelihood function based on both samples is

$$\begin{aligned} l(\theta) &= l_1(\theta) + l_2(\theta) \\ &= (x + y) \log \theta + (n + m - x - y) \log(1 - \theta). \end{aligned} \tag{9.2.6}$$

This is of the same form as (9.2.5), and the overall MLE is

$$\hat{\theta} = \frac{x + y}{n + m}.$$

Since $x = n\hat{\theta}_1$ and $y = m\hat{\theta}_2$, we have

$$\hat{\theta} = \frac{n}{n+m}\hat{\theta}_1 + \frac{m}{n+m}\hat{\theta}_2,$$

which is a weighted average of $\hat{\theta}_1$ and $\hat{\theta}_2$. For instance, if 90 individuals are examined in the first sample ($n = 90$), and only 10 in the second ($m = 10$), we have

$$\hat{\theta} = 0.9\hat{\theta}_1 + 0.1\hat{\theta}_2.$$

The overall MLE lies between $\hat{\theta}_1$ and $\hat{\theta}_2$, and is closer to $\hat{\theta}_1$, the MLE from the larger sample, than to $\hat{\theta}_2$.

Note that the log likelihood function (9.2.6) is the same as would be obtained if we considered a single sample of $n + m$ individuals, $x + y$ of whom were found to have tuberculosis. The division of the results into two separate experiments is irrelevant in so far as estimation of θ is concerned. □

EXAMPLE 9.2.2. In performing the experiment described in Example 9.1.3, it is necessary to specify the volume v of river water which is to be placed in each test tube. If v is made too large, then all of the test tubes will contain bacteria and give a positive reaction. If v is too small, we may get only negative reactions. In either case, the experiment will be rather uninformative about μ, the concentration of bacteria in the river.

One way to guard against this difficulty is to prepare two (or more) different types of test tubes containing different volumes of river water. Suppose that 40 test tubes containing 10 ml of river water were tested, and 28 gave negative results. Also, 40 test tubes containing 1 ml of river water were tested, and 37 gave negative results. What is the maximum likelihood estimate of μ?

SOLUTION. From Example 9.1.3, the likelihood function based on the 40 tubes containing 10 ml is

$$L_1(\mu) = p_1^{28}(1 - p_1)^{12}$$

where $p_1 = e^{-10\mu}$, and the MLE of μ is $\hat{\mu}_1 = 0.0357$. The log likelihood function is

$$l_1(\mu) = 28 \log p_1 + 12 \log(1 - p_1).$$

Similarly, from the 40 tubes containing 1 ml we obtain

$$l_2(\mu) = 37 \log p_2 + 3 \log(1 - p_2)$$

where $p_2 = e^{-\mu}$, and the MLE of μ is

$$\hat{\mu}_2 = \frac{\log n - \log y}{v} = \frac{\log 40 - \log 37}{1} = 0.0780.$$

By (9.2.2), the log likelihood function based on all 80 tubes is

$$l(\mu) = l_1(\mu) + l_2(\mu),$$

and the overall MLE $\hat{\mu}$ is chosen to maximize this function.

For the first sample we have

$$\begin{aligned} S_1(\mu) &= \frac{d}{d\mu} l_1(\mu) = \frac{dl_1}{dp_1} \cdot \frac{dp_1}{d\mu} \\ &= -10p_1 \left[\frac{28}{p_1} - \frac{12}{1 - p_1} \right] = \frac{120}{1 - p_1} - 400; \end{aligned}$$

$$\mathscr{I}_1(\mu) = -\frac{d}{d\mu}S_1(\mu) = -\frac{120}{(1-p_1)^2}\frac{dp_1}{d\mu} = \frac{1200p_1}{(1-p_1)^2}.$$

Similarly, for the second sample we obtain

$$S_2(\mu) = \frac{3}{1-p_2} - 40; \qquad \mathscr{I}_2(\mu) = \frac{3p_2}{(1-p_2)^2}.$$

Thus, by (9.2.3) and (9.2.4), the combined results are

$$S(\mu) = \frac{120}{1-p_1} + \frac{3}{1-p_2} - 440; \qquad \mathscr{I}(\mu) = \frac{1200p_1}{(1-p_1)^2} + \frac{3p_2}{(1-p_2)^2}$$

where $p_1 = e^{-10\mu}$ and $p_2 = e^{-\mu}$.

The score function is more complicated than in previous examples, and it is not possible to solve the maximum likelihood equation $S(\mu) = 0$ algebraically. However, $\hat{\mu}$ can easily be found numerically. For instance, we could evaluate $S(\mu)$ for various values of μ, and hence find by trial and error the approximate value of μ for which $S(\mu) = 0$. Alternatively, an iterative root-finding procedure such as Newton's method can be used (see Section 9.8). For this example we find that $\hat{\mu} = 0.04005$, correct to five decimal places. □

PROBLEMS FOR SECTION 9.2

1.†(a) In a population in which the frequency of the gene for color blindness is θ, genetic theory indicates that the probability that a male is color-blind is θ, and the probability that a female is color-blind is θ^2. A random sample of M males is found to include m color-blind, and a random sample of N females includes n color-blind. Find the likelihood function of θ based on both samples, and show that $\hat{\theta}$ can be obtained as a root of a quadratic equation.

(b) One hundred males and 100 females were examined. Eleven males and two females were found to be color-blind. Find the MLE of θ based on the data for males, the MLE of θ based on the data for females, and the overall MLE of θ based on all of the data.

2 (a) If deaths from a rare non-contagious disease occur randomly and uniformly throughout the population, the number of deaths in a region of population P should have a Poisson distribution with mean λP. Suppose that the numbers of deaths observed in n regions with populations $P_1, P_2, \ldots, P_n$ are $y_1, y_2, \ldots, y_n$. Derive an expression for the MLE of λ.

(b) The table on the following page shows the number of male deaths from cancer of the liver during 1964–8 for Ontario regions. Find $\hat{\lambda}$ for these data, and compute the estimated expected number of deaths for each region. Do the data appear to be consistent with the assumptions in (a)?

3. (a) Suppose that $\hat{\theta}$ is a weighted average of $\hat{\theta}_1$ and $\hat{\theta}_2$; that is,

$$\hat{\theta} = a_1\hat{\theta}_1 + a_2\hat{\theta}_2$$

	Region	Population	Deaths
1.	Eastern Ontario	423,447	37
2.	Lake Ontario	175,685	11
3.	Central Ontario	1,245,379	72
4.	Niagara	413,465	40
5.	Lake Erie	216,476	12
6.	Lake St. Clair	242,810	14
7.	Mid-Western Ontario	213,591	16
8.	Georgian Bay	166,045	9
9.	Northeastern Ontario	265,880	15
10.	Lakehead–NW Ontario	116,371	12

where a_1 and a_2 are positive real numbers with $a_1 + a_2 = 1$. Show that $\hat{\theta}$ must lie between $\hat{\theta}_1$ and $\hat{\theta}_2$.

(b) Suppose that

$$\hat{\theta} = a_1\hat{\theta}_1 + a_2\hat{\theta}_2 + \cdots + a_n\hat{\theta}_n$$

where the a_i's are positive and $\Sigma a_i = 1$. Show that $\hat{\theta}$ must lie between the smallest and the largest of the $\hat{\theta}_i$'s.

9.3. Relative Likelihood

As in Section 9.1 we suppose that the data (observed event) E from an experiment has probability $P(E; \theta)$ which depends upon an unknown parameter θ. The maximum likelihood estimate $\hat{\theta}$ is the value of θ which maximizes $P(E; \theta)$. It is the "most likely" or "most plausible" value of θ in the sense that it maximizes the probability of what has been observed.

The relative plausibilities of other θ-values may be examined by comparing them with $\hat{\theta}$. Values of θ such that $P(E; \theta)$ is nearly as large as $P(E; \hat{\theta})$ are fairly plausible in that they explain the data almost as well as $\hat{\theta}$ does. Values of θ for which $P(E; \theta)$ is much less than $P(E; \hat{\theta})$ are implausible because they make what has been observed much less probable than $\hat{\theta}$ does.

The *relative likelihood function* (RLF) of θ is defined as the ratio of the likelihood function $L(\theta)$ to its maximum $L(\hat{\theta})$:

$$R(\theta) = L(\theta)/L(\hat{\theta}). \tag{9.3.1}$$

Since $L(\theta) = c \cdot P(E; \theta)$ where c does not depend upon θ, it follows that

$$R(\theta) = \frac{c \cdot P(E; \theta)}{c \cdot P(E; \hat{\theta})} = \frac{P(E; \theta)}{P(E; \hat{\theta})}.$$

The multiplicative constant c in (9.1.2) cancels out of the expression for $R(\theta)$. Thus $R(\theta)$, like $\hat{\theta}$, $S(\theta)$, and $\mathscr{I}(\theta)$, is not affected by the choice of c in (9.1.2). Note that since $L(\theta) \le L(\hat{\theta})$ for all possible θ-values, it follows that $0 \le R(\theta) \le 1$.

The *log relative likelihood function* is the natural logarithm of the relative likelihood function:

$$r(\theta) = \log R(\theta) = \log L(\theta) - \log L(\hat{\theta}).$$

It follows that

$$r(\theta) = l(\theta) - l(\hat{\theta}) \tag{9.3.2}$$

where $l(\theta)$ is the log likelihood function. Since $0 \le R(\theta) \le 1$, we have $-\infty \le r(\theta) \le 0$ for all possible parameter values.

Let θ_1 denote some particular parameter value. Then

$$R(\theta_1) = \frac{L(\theta_1)}{L(\hat{\theta})} = \frac{P(E;\theta_1)}{P(E;\hat{\theta})}$$

$$= \frac{\text{Probability of the data } E \text{ when } \theta = \theta_1}{\text{Maximum probability of } E \text{ for any value of } \theta}.$$

If $R(\theta_1) = 0.1$, say, then θ_1 is rather an implausible parameter value because the data are ten times more probable when $\theta = \hat{\theta}$ than they are when $\theta = \theta_1$. However if $R(\theta_1) = 0.5$, say, then θ_1 is a fairly plausible parameter value because it gives the data 50% of the maximum possible probability under the model. The relative likelihood function ranks all possible parameter values according to their plausibilities in the light of the data.

Usually $\hat{\theta}$ exists and is unique, and the definition (9.3.1) applies. More generally, $R(\theta)$ may be defined as the ratio of $L(\theta)$ to its supremum over all parameter values,

$$R(\theta) = L(\theta)/\sup_{\theta} L(\theta).$$

Since $L(\theta) = c \cdot P(E;\theta)$ where $P(E;\theta) \le 1$, the supremum is finite. The relative likelihood function exists and may be used to rank parameter values according to their plausibilities even when $\hat{\theta}$ does not exist.

Likelihood Regions and Intervals

The set of θ-values for which $R(\theta) \ge p$ is called a $100p\%$ *likelihood region* for θ. Usually the $100p\%$ likelihood region will consist of an interval of real values, and then it is called a $100p\%$ *likelihood interval* (LI) for θ.

Usually we consider 50%, 10%, and 1% likelihood intervals or regions. Values inside the 10% LI will be referred to as "plausible", and values outside this interval as "implausible". Similarly, we shall refer to values inside the 50% LI as "very plausible", and values outside the 1% LI as "very implausible". Of course, the choice of division points at 50%, 10%, and 1% is rather arbitrary and should not be taken too seriously.

The 14.7% and 3.6% likelihood intervals are sometimes calculated. These correspond approximately to 95% and 99% confidence intervals (see Section 11.4).

Likelihood regions or intervals may be determined from a graph of $R(\theta)$ or its logarithm $r(\theta)$, and usually it is more convenient to work with $r(\theta)$. Since $\log 0.5 = -0.69$, we have $r(\theta) \ge -0.69$ for the 50% likelihood interval. Similarly, $r(\theta) \ge -2.30$ for the 10% LI, and $r(\theta) \ge -4.61$ for the 1% LI. Alternatively, the endpoints of the $100p\%$ LI can be found as roots of the equation $r(\theta) - \log p = 0$. Usually it is necessary to solve this equation numerically (see Section 9.8).

When $r(\theta)$ has a simple form as in the examples of this section, the information about θ can be summarized adequately by reporting $\hat{\theta}$ and two or three likelihood intervals. Given these results, it is possible to reconstruct a graph of $r(\theta)$ to a reasonable approximation. Such a summary might not be appropriate if, for instance, $r(\theta)$ had several relative maxima and minima in the neighborhood of $\hat{\theta}$. In this case it would be better to present a graph of $r(\theta)$ than to attempt a summary.

EXAMPLE 9.3.1 (continuation of Example 9.1.1). Suppose that, out of 100 people examined, three are found to have tuberculosis. On the basis of this observation, which values of θ are plausible? Compare with the results that would be obtained if 200 people were examined and six were found to have tuberculosis.

SOLUTION. From Example 9.1.1, the log likelihood function is

$$l(\theta) = 3 \log \theta + 97 \log (1 - \theta),$$

and the maximum likelihood estimate is $\hat{\theta} = 0.03$. The maximum of the log likelihood is

$$l(\hat{\theta}) = 3 \log (0.03) + 97 \log (0.97) = -13.47.$$

The log relative likelihood function is thus

$$r(\theta) = l(\theta) - l(\hat{\theta}) = 3 \log \theta + 97 \log (1 - \theta) + 13.47.$$

A graph of this function is shown in Figure 9.3.1 (solid line). From the graph we find that $r(\theta) \ge -2.30$ for $0.006 \le \theta \le 0.081$, and this is the 10% LI for θ. Values of θ inside this interval are fairly plausible in light of the data. Similarly, the 50% LI is $0.014 \le \theta \le 0.054$. Values within this interval are quite plausible, because they give the data at least 50% of the maximum probability which is possible under the model.

If we observed 6 diseased out of 200, we would have

$$l(\theta) = 6 \log \theta + 194 \log (1 - \theta),$$

and $\hat{\theta} = 0.03$ as before. The maximum of the log likelihood is now

$$l(\hat{\theta}) = -26.95.$$

Figure 9.3.1 shows the corresponding log relative likelihood function with a broken line. Both functions attain their maximum at $\hat{\theta} = 0.03$. However the log RLF based on the sample of 200 people is more sharply peaked than the log

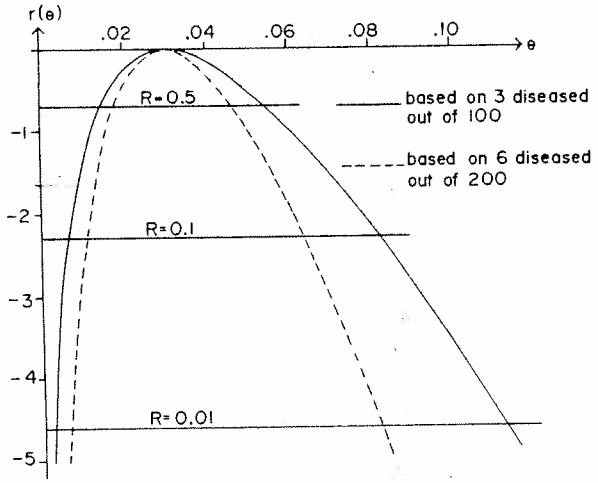

Figure 9.3.1. Log relative likelihood functions from Example 9.3.1.

RLF based on the sample of 100 people. As a result, the larger sample gives shorter likelihood intervals for θ. For instance, the 10% LI is (0.011, 0.063) for the sample of 200, as opposed to (0.006, 0.081) for the sample of 100.

In general, increasing the number of independent observations will produce a more sharply peaked likelihood function and thus shorter likelihood intervals for θ. With more observations there will be a shorter range of plausible values for θ, and so θ can be more precisely estimated. As a rough guide, the length of the $100p\%$ likelihood interval is inversely proportional to the square root of the number of independent observations. Thus about 4 times as many observations are needed to produce an interval only half as wide. □

EXAMPLE 9.3.2. In Example 9.2.2, we considered data from two experiments with test tubes containing river water:

Observation 1: $y = 28$ negative reactions out of $n = 40$ test tubes each containing $v = 10$ ml.

Observation 2: $y = 37$ negative out of $n = 40$ tubes with $v = 1$.

Graph the log relative likelihood functions and obtain 50% likelihood intervals for μ based on the two observations taken separately, and taken together.

SOLUTION. The log likelihood function based only on observation 1 is

$$l_1(\mu) = 28 \log p_1 + 12 \log(1 - p_1); \qquad p_1 = e^{-10\mu}.$$

Since $p_1 = y/n = 0.7$ at the maximum (Example 9.1.3), the maximum log likelihood is

$$l_1(\hat{\mu}_1) = 28 \log 0.7 + 12 \log 0.3 = -24.43.$$

The log relative likelihood function is then

$$r_1(\mu) = l_1(\mu) - l_1(\hat{\mu}_1) = -280\mu + 12 \log(1 - e^{-10\mu}) + 24.43.$$

Similarly, the log relative likelihood function based only on observation 2 is

$$r_2(\mu) = -37\mu + 3 \log(1 - e^{-\mu}) + 10.66.$$

For both observations together, the log LF is

$$\begin{aligned} l(\mu) &= l_1(\mu) + l_2(\mu) \\ &= -317\mu + 12 \log(1 - e^{-10\mu}) + 3 \log(1 - e^{-\mu}). \end{aligned}$$

From Example 9.2.2, the overall MLE is $\hat{\mu} = 0.04005$, and substitution of this value gives $l(\hat{\mu}) = -35.71$. The log RLF based on both observations is thus

$$r(\mu) = l(\mu) + 35.71.$$

The three log RLF's are tabulated in Table 9.3.1 and graphed in Figure 9.3.2, with $r(\mu)$ being given by the broken line. From the graphs, the following 50% likelihood intervals may be obtained:

Observation 1 only: $0.025 \le \mu \le 0.049$

Observation 2 only: $0.036 \le \mu \le 0.144$

Both observations combined: $0.029 \le \mu \le 0.053$.

Table 9.3.1. Log Relative Likelihood Functions for Example 9.3.2

μ	$r_1(\mu)$	$r_2(\mu)$	$r(\mu)$
0.005		−5.43	
0.01	−6.59	−3.55	−9.51
0.015	−3.42	−2.52	−5.32
0.018	−2.25	−2.09	−3.71
0.02	−1.66	−1.85	−2.89
0.025	−0.67	−1.37	−1.42
0.03	−0.17	−1.02	−0.57
0.04	−0.08	−0.54	−0.00
0.05	−0.76	−0.26	−0.39
0.06	−1.92	−0.09	−1.39
0.07	−3.40	−0.02	−2.80
0.08	−5.12	−0.00	−4.50
0.10		−0.10	
0.20		−1.87	
0.30		−4.50	

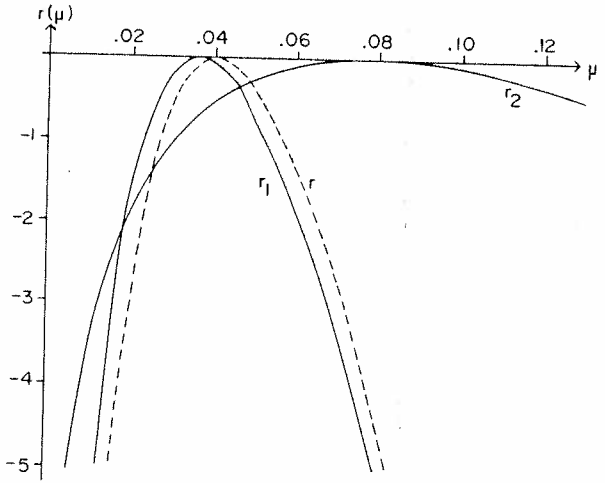

Figure 9.3.2. Combination of log RLFs from independent experiments.

The log RLF based on observation 2 only is almost flat over a large range of μ-values, indicating that this observation provides relatively little information about μ. The combined log RLF based on all the data is very nearly the same as that based on observation 1 alone.

The combined log RLF $r(\mu)$ can be obtained directly from a table or graph of $r_1(\mu)$ and $r_2(\mu)$. We form the sum $r_1(\mu)+r_2(\mu)$, and observe the value of μ at which it is greatest. This will be the overall MLE $\hat{\mu}$. The combined log RLF is then

$$r(\mu) = r_1(\mu) + r_2(\mu) - [r_1(\hat{\mu}) + r_2(\hat{\mu})].$$

If $r_1(\hat{\mu}) + r_2(\hat{\mu})$ is small (e.g. less than -2), then there exists no single value of μ which is plausible on both sets of data. The two sets of data are then in contradiction, since they point to different values for the same parameter μ. When this happens, it is generally inadvisable to combine the two data sets. Instead, the parameter should be estimated separately for each data set, and an explanation for the discrepancy should be sought (see Section 12.3).

In the present example, we find that $r_1(\hat{\mu}) + r_2(\hat{\mu}) = -0.62$. There do exist values of μ (near 0.04) which are quite plausible for both observations, and hence no contradiction is apparent. It is therefore reasonable to combine the two observations, and to base statements about μ on $r(\mu)$, the combined RLF. □

EXAMPLE 9.3.3. Relative likelihood when $\hat{\mu} = +\infty$.

Suppose that $n = 40$ test tubes are prepared, each containing $v = 10$ ml of river water, and that all of them give positive results ($y = 0$). The likelihood function of μ is then

$$L(\mu) = (1-p)^{40} = (1 - e^{-10\mu})^{40} \qquad \text{for } 0 \le \mu < \infty.$$

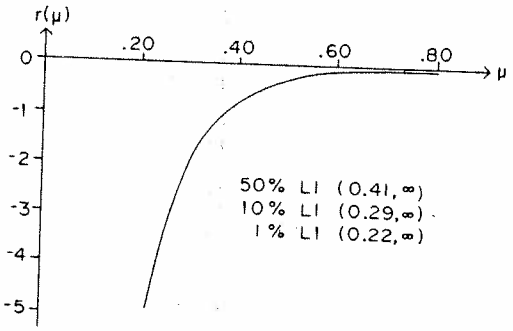

Figure 9.3.3. Log relative likelihood function when $\hat{\mu} = +\infty$.

Then, as we noted at the end of Example 9.1.3, $L(\mu)$ increases as μ increases to $+\infty$. We say that $\hat{\mu} = +\infty$, although strictly speaking $\hat{\mu}$ does not exist because this value does not belong to the parameter space.

Even when $\hat{\mu}$ does not exist, the relative likelihood function is well defined and can be used to determine the range of plausible parameter values. As μ tends to $+\infty$, $L(\mu)$ increases to 1, and hence

$$\sup_{0 \le \mu < \infty} L(\mu) = 1.$$

The relative likelihood function of μ is then

$$R(\mu) = \frac{L(\mu)}{\sup L(\mu)} = (1 - e^{-10\mu})^{40} \qquad \text{for } 0 \le \mu < \infty.$$

The log relative likelihood function,

$$r(\mu) = 40 \log(1 - e^{-10\mu}),$$

is plotted in Figure 9.3.3. We have $r(\mu) \ge -0.69$ for $\mu > 0.41$, and hence the 50% LI for μ is $(0.41, \infty)$. Any value of μ which exceeds 0.41 is very plausible in light of the data. Similarly, we have $r(\mu) \le -4.61$ for $\mu \le 0.22$, so that any value of μ less than 0.22 is extremely implausible.

PROBLEMS FOR SECTION 9.3

1.† Prepare a graph of the log RLF in Problem 9.1.1, and from it obtain 50% and 10% likelihood intervals for λ.

2. The number of westbound vehicles which pass a fixed point on a main east–west road in 10 seconds is a Poisson variate with mean μ. The numbers passing in disjoint time intervals are independent. The following table summarizes the data from 300 ten-second intervals:

No. of vehicles in 10 sec.	0	1	2	3	4	5
Frequency observed	61	107	76	45	10	1

Plot the log RLF of μ, and from the graph obtain 50% and 10% likelihood intervals for μ.

3. A company plans to purchase either machine 1 or machine 2, and has available the following performance data:

Machine 1: 0 failures in 7800 trials

Machine 2: 4 failures in 21804 trials.

Trials are independent, and the probability of failure is θ_1 for machine 1 and θ_2 for machine 2. Plot the log RLF's of θ_1 and θ_2 on the same graph. Under what conditions would you recommend the purchase of machine 2 rather than machine 1?

4.† Find the relative likelihood of $\theta = 0$ (a balanced die) in Problem 9.1.5(b).

5. (a) Plot the log RLF of the gene frequency θ in Problem 9.1.4(b).
   (b) A random sample of 200 individuals is taken from a second large population in which θ may be different. The numbers of individuals with the three blood types are found to be 48, 102, and 50. Plot the log RLF based on this sample on the graph prepared in (a).
   (c) Indicate, with reasons, whether you think that θ could be the same in both populations. If it is appropriate to do so, obtain the log RLF for θ based on both samples, and show it on the graph prepared in (a).

6. Find 50% and 10% likelihood intervals for N in Problem 9.1.7.

7.† Suppose that $r = n = 10$ and $y = 5$ in Problem 9.1.11. Which values of b have relative likelihood 50% or more? 10% or more?

8. In Problem 9.1.10(b), graph the log RLF of p and obtain a 10% LI for p.

9. The records from 200 samples in Problem 9.1.12 showed 180 with one defective, 17 with two defectives, and 3 with three defectives. Evaluate $\hat{\theta}$, plot the log RLF of θ, and obtain a 10% likelihood interval for θ.

10.† A solution in which virus particles are suspended is poured over a cell sheet. Each virus particle attacks the cells to form a plaque which is visible. The cell sheet is divided into disjoint regions of equal area. The following are the numbers of plaques observed in 20 regions:

3	2	6	0	2	2	1	4	8	1
1	3	2	0	1	5	2	3	4	1

   (a) Suppose that the virus particles are randomly and uniformly distributed over the cell sheet at the rate of λ per region. Plot the log RLF of λ and find a 10% LI.
   (b) Suppose that, for each region, the experimenter recorded only whether the result was positive (at least one plaque) or negative (no plaques). Thus for the 20 regions referred to above, we would know only that there were 18 positives and 2 negatives. Plot the new log RLF for λ on the graph prepared in (a) and find a 10% LI. Has much information about λ been lost in recording only positives and negatives?

11. The following model is proposed for the distribution of family size in a large population:

$$P(k \text{ children in family}) = \alpha^k \qquad \text{for } k = 1, 2, \ldots;$$

$$P(0 \text{ children in family}) = (1 - 2\alpha)/(1 - \alpha).$$

Here α is an unknown parameter, and $0 < \alpha < \frac{1}{2}$. Fifty families were chosen at random from the population. The observed numbers of children are summarized in the following frequency table:

No. of children	0	1	2	3	4
Frequency observed	17	22	7	3	1

   (a) Find the MLE of α and calculate estimated expected frequencies. Does the model give a reasonable fit to the data?
   (b) A large study done 20 years earlier indicated that $\alpha = 0.45$. Is this plausible for the current data?

9.4. Likelihood for Continuous Models

Continuous probability distributions are frequently used as probability models for experiments involving the measurement of time, weight, length, etc. Suppose that X has a continuous distribution with probability density function f and cumulative distribution function F, depending upon an unknown parameter θ. The experiment is performed and values of X are observed. The problem is to use the data to estimate θ, or more generally, to determine which values of θ are plausible in light of the data.

When X is a continuous variate, $f(x)$ does *not* give the probability of observing the value x. In fact, as we noted in Section 6.1, the probability of any particular real value is zero. An actual measurement of time, weight, etc. will necessarily be made to only finitely many decimal places. An observed value x will therefore correspond to some small interval of real values $a < X \leq b$, say. The probability of observing the value x is then

$$P(a < X \leq b) = \int_a^b f(x)dx = F(b) - F(a). \tag{9.4.1}$$

Suppose that, in n independent repetitions, we observe n values $x_1, x_2, \ldots, x_n$, with x_i corresponding to the real interval $[a_i, b_i]$. Because repetitions are independent, the probability of the data is obtained as a product:

$$P(E; \theta) = \prod_{i=1}^{n} P(a_i < X \leq b_i) = \prod_{i=1}^{n} [F(b_i) - F(a_i)]. \tag{9.4.2}$$

The likelihood function of θ is proportional to (9.4.2).

If the interval length $\Delta_i = b_i - a_i$ is small, then $F(b_i)$ will be close to $F(a_i)$,

and computation of the difference $F(b_i) - F(a_i)$ may introduce serious roundoff errors. In this case, we make use of (6.1.7), and approximate the area under the density function between a_i and b_i by the area of a rectangle with base Δ_i and height $f(x_i)$:

$$P(a_i < X \le b_i) = F(b_i) - F(a_i) \approx f(x_i)\Delta_i. \tag{9.4.3}$$

Some or all of the factors in (9.4.2) are approximated in this way to obtain a function which is easier to deal with computationally and mathematically.

In the most usual case, all of the measurement intervals Δ_i are small, and the approximation (9.4.3) may be applied to all of the terms in (9.4.2). This gives

$$P(E; \theta) \approx \prod_{i=1}^{n} f(x_i)\Delta_i = \left[\prod_{i=1}^{n} \Delta_i\right] \prod_{i=1}^{n} f(x_i).$$

Since the Δ_i's do not depend upon θ, *the likelihood function is proportional to the product of probability densities,*

$$L(\theta) = c \cdot \prod_{i=1}^{n} f(x_i) \tag{9.4.4}$$

where c is any convenient positive constant. This is actually an approximation, but it will be an extremely accurate one whenever the Δ_i's are all small.

It is not necessary to replace every factor in (9.4.2) by the approximation (9.4.3). For instance, it may happen that $f(x)$ changes rapidly when x is small, in which case the original terms in (9.4.2) could be retained for small values x_i, and the approximation could be used for large x_i's. Another situation where some of the terms in (9.4.2) should be retained will be discussed in the next section.

EXAMPLE 9.4.1. A certain type of electronic component is susceptible to instantaneous failure at any time. However, components do not deteriorate with age, and the chance of failure within a given time period does not depend upon the age of the component. From Section 6.2, the lifetime of such a component should have an exponential distribution, with probability density function

$$f(x) = \frac{1}{\theta} e^{-x/\theta} \qquad \text{for } x > 0,$$

where θ is the expected lifetime of such components.

Ten such components were tested independently. Their lifetimes, measured to the nearest day, were as follows:

70 11 66 5 20 4 35 40 29 8.

What values of θ are plausible in the light of the data?

SOLUTION BASED ON (9.4.4). Each observed lifetime corresponds to an interval of length $\Delta = 1$. The average lifetime is about 30, and the exponential p.d.f.

with mean $\theta = 30$ changes very little over an interval of length 1. Areas under the p.d.f. will thus be well approximated by rectangles, and (9.4.4) should give an accurate approximation. We substitute for $f(x_i)$ in (9.4.4) and take $c = 1$ to obtain

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = \theta^{-n} \exp\left(-\frac{1}{\theta}\Sigma x_i\right).$$

The log likelihood function is

$$l(\theta) = -n \log \theta - \frac{1}{\theta}\Sigma x_i.$$

The score and information functions are

$$S(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\Sigma x_i; \qquad \mathcal{I}(\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3}\Sigma x_i.$$

We may now solve the maximum likelihood equation $S(\theta) = 0$ to obtain $\hat{\theta} = \frac{1}{n}\Sigma x_i = \bar{x}$. Note that

$$\mathcal{I}(\hat{\theta}) = -\frac{n}{\hat{\theta}^2} + \frac{2n\hat{\theta}}{\hat{\theta}^3} = \frac{n}{\hat{\theta}^2} > 0$$

and hence the root obtained is a relative maximum.

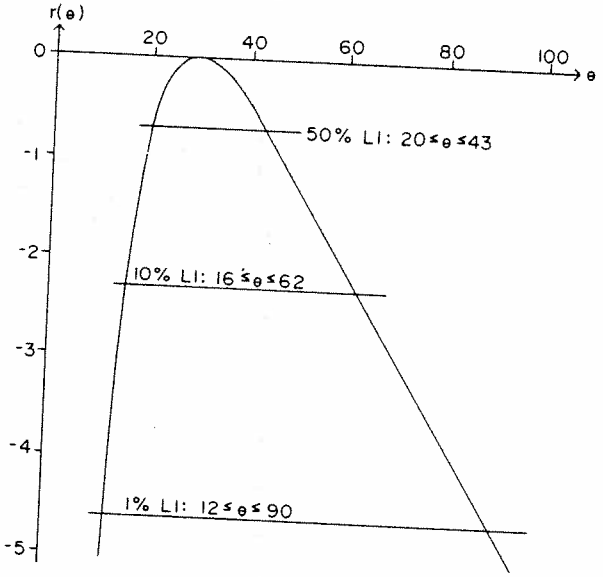

Figure 9.4.1. Log relative likelihood function for the mean based on ten observations from an exponential distribution.

The total of the $n = 10$ observed lifetimes is $\Sigma x_i = 288$, so that $\hat{\theta} = 28.8$ and

$$l(\theta) = -10 \log \theta - \frac{288}{\theta}.$$

The log relative likelihood function,

$$r(\theta) = l(\theta) - l(\hat{\theta}),$$

is plotted in Figure 9.4.1. The observations indicate a mean lifetime between 20 and 43 days (50% LI). Values of θ less than 16 days or greater than 62 days are implausible (relative likelihood less than 10%).

EXACT SOLUTION BASED ON (9.4.2). For comparison, we shall determine the exact likelihood function based on (9.4.2). The c.d.f. of the exponential distribution with mean θ is

$$F(x) = 1 - e^{-x/\theta} \qquad \text{for } x > 0.$$

An observed integer value $x > 0$ corresponds to a real interval $x \pm 0.5$, with probability

$$F(x+0.5) - F(x-0.5) = \exp\left(-\frac{x-0.5}{\theta}\right) - \exp\left(-\frac{x+0.5}{\theta}\right)$$
$$= \left[\exp\left(\frac{1}{2\theta}\right) - \exp\left(-\frac{1}{2\theta}\right)\right]\exp\left(-\frac{x}{\theta}\right).$$

Hence by (9.4.2), the probability of observed values $x_1, x_2, \ldots, x_n$ is

$$P(E;\theta) = \prod_{i=1}^{n}\left[\exp\left(\frac{1}{2\theta}\right) - \exp\left(-\frac{1}{2\theta}\right)\right]\exp(-x_i/\theta)$$
$$= \left[\exp\left(\frac{1}{2\theta}\right) - \exp\left(-\frac{1}{2\theta}\right)\right]^n \exp\left(-\frac{1}{\theta}\Sigma x_i\right).$$

The likelihood function is

$$L(\theta) = c \cdot P(E;\theta)$$

and we take $c = 1$ for convenience. The log likelihood function is

$$l(\theta) = n \log\left[\exp\left(\frac{1}{2\theta}\right) - \exp\left(-\frac{1}{2\theta}\right)\right] - \frac{1}{\theta}\Sigma x_i,$$

and the solution of the equation $S(\theta) = 0$ is

$$\hat{\theta} = \left[\log\left(\frac{\bar{x}+0.5}{\bar{x}-0.5}\right)\right]^{-1}.$$

The exact log RLF is now $r(\theta) = l(\theta) - l(\hat{\theta})$.

For the ten observations given, we find that $\hat{\theta} = 28.797$, which is very close

Table 9.4.1. Comparison of Exact and Approximate Likelihoods Based on Ten Observations from an Exponential Distribution

θ	Exact $r(\theta)$ Based on (9.4.2)	Approx. $r(\theta)$ Based on (9.4.4)	Difference (9.4.2)–(9.4.4)
5	−30.0745	−30.0906	+0.0161
10	−8.2184	−8.2221	+0.0037
12	−5.2429	−5.2453	+0.0024
15	−2.6754	−2.6767	+0.0013
20	−0.7530	−0.7536	+0.0006
25	−0.1048	−0.1050	+0.0002
40	−0.4853	−0.4850	−0.0003
60	−2.1401	−2.1397	−0.0004
80	−3.8169	−3.8165	−0.0004
100	−5.3284	−5.3279	−0.0005
200	−10.8199	−10.8194	−0.0005
300	−14.3946	−14.3941	−0.0005

to our previous result ($\hat{\theta} = 28.800$). Table 9.4.1 compares the exact log RLF with the approximate log RLF which we obtained previously from (9.4.4). The agreement is extremely close over the range $12 \le \theta \le 100$ which includes all but the most implausible parameter values. As one might expect, the agreement becomes worse as θ becomes small; for then the p.d.f. changes more rapidly over a short interval, and the approximation (9.4.3) is less accurate.

More generally, if an observation x from an exponential distribution corresponds to a real interval $x \pm h$, the ratio of the exact probability (9.4.1) to the approximate probability (9.4.3) is

$$\frac{\exp\left(-\frac{x-h}{\theta}\right) - \exp\left(-\frac{x+h}{\theta}\right)}{\frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)\cdot 2h} = \frac{e^c - e^{-c}}{2c} = 1 + \frac{c^2}{3!} + \frac{c^4}{5!} + \cdots,$$

where $c = h/\theta$ is the ratio of half the length of the measurement interval to the mean of the distribution. The approximation will be accurate whenever c is small.

PROBLEMS FOR SECTION 9.4

1.† The following are the times (in hours) between successive failures of the air-conditioning system in an aircraft:

97	51	11	4	141	18	142	68	77
80	1	16	106	206	82	54	31	216
46	111	39	63	18	191	18	163	24

(a) Assuming that these are independent observations from an exponential distribution with mean θ, find $\hat{\theta}$ and the 10% likelihood interval for θ.
(b) Prepare a frequency table for these data using classes (0, 50], (50, 100], (100, 200], and (200, ∞). Calculate estimated expected frequencies for these classes under the assumption in (a). Does the exponential distribution appear to give a reasonable model for the data?

2. Family income X is measured on a scale such that $X = 1$ corresponds to a subsistence level income. The p.d.f. of the income distribution is assumed to be

$$f(x) = \theta / x^{\theta+1} \qquad \text{for } x \geq 1$$

where $\theta > 0$. The following are the incomes of ten randomly selected families:

1.02, 1.41, 1.75, 2.31, 3.42, 4.31, 9.21, 17.4, 38.6, 392.8

Find the MLE and the 10% likelihood interval for θ.

3. It is thought that the times between particle emissions from a radioactive source are exponentially distributed with mean θ. However, the Geiger counter used to register the emissions locks for 1 unit of time after recording an emission. Thus the p.d.f. of the time X between successive recordings is

$$f(x) = \frac{1}{\theta} e^{-(x-1)/\theta} \qquad \text{for } x \geq 1.$$

The following are ten observed times between recordings:

1.47 1.46 2.20 1.36 2.90
3.71 3.89 1.29 1.86 1.81

Find the MLE and the 10% LI for θ.

4.† A manufacturing process produces fibers of varying lengths. The length of a fiber is a continuous variate with p.d.f.

$$f(x) = \theta^{-2} x e^{-x/\theta} \qquad \text{for } x > 0$$

where $\theta > 0$ is an unknown parameter. Suppose that n randomly selected fibers have lengths $x_1, x_2, \ldots, x_n$. Find expressions for the MLE and RLF of θ.

5. Let Y denote the time to failure of an electrical component. The distribution of Y is exponential with mean θ/t, where t is the temperature at which the component is operated. Suppose that n components are tested independently at temperatures $t_1, t_2, \ldots, t_n$, respectively, and their observed lifetimes are $y_1, y_2, \ldots, y_n$. Derive an expression for the MLE of θ.

6.† A laboratory method for determining the concentration of a trace metal in solution produces $N(0, \sigma^2)$ errors. If the true concentration is μ, then the measured concentration X is a random variable distributed as $N(\mu, \sigma^2)$. The value of σ is known from previous experience with the method.

(a) Let $x_1, x_2, \ldots, x_n$ be independent measurements of the same unknown concentration μ. Show that $\hat{\mu} = \bar{x}$, and that the log RLF of μ is

$$r(\mu) = -\frac{n}{2\sigma^2}(\bar{x} - \mu)^2 \qquad \text{for } -\infty < \mu < \infty.$$

Hint: Show that $\Sigma(x_i - \mu)^2 = \Sigma(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$.

(b) The original solution is diluted by half so that the concentration is now $\mu/2$, and m additional measurements $y_1, y_2, \ldots, y_m$ are taken. Find the MLE of μ based on all $n + m$ measurements.

7. A laboratory method for determining the concentration of a trace metal in solution produces independent $N(0, \sigma^2)$ errors. If the true concentration is μ, then the measured concentration X is a random variable distributed as $N(\mu, \sigma^2)$. In order to estimate σ, several measurements are taken for a solution with known concentration μ.

(a) The following 5 measurements were made on a solution with known concentration $\mu = 10$:

9.3 11.2 8.7 10.1 10.7

Plot the log RLF of σ.
(b) The following 5 measurements were made on a solution with known concentration $\mu = 20$:

21.7 19.9 20.3 20.4 19.7

Plot the log RLF of σ based on these data on the graph prepared in (a).
(c) If it is appropriate to do so, find the MLE and the 10% LI for σ based on all ten measurements.

8. A scientist makes n measurements $x_1, x_2, \ldots, x_n$ of a constant μ using a technique with known error variance σ^2, and m additional measurements $y_1, y_2, \ldots, y_m$ of μ using a technique with known error variance $k\sigma^2$. Assuming that all measurements are independent and normally distributed, find the MLE of μ. Show that, if $n = m$ and $k > 1$, then $\hat{\mu}$ is closer to $\bar{x}$ than to $\bar{y}$, and explain why this is desirable.

9. (a) Suppose that U is a continuous variate, and that U/θ has a χ^2 distribution with n degrees of freedom. Find the p.d.f. of U, and show that $\hat{\theta} = U/n$.
(b) Suppose that V is independent of U, and V/θ has a χ^2 distribution with m degrees of freedom. Find the joint p.d.f. of U and V, and show that the MLE of θ based on both U and V is $(U + V)/(n + m)$.

10.† The probability density function for a unit exponential distribution with guarantee time $c > 0$ is

$$f(x) = e^{c-x} \qquad \text{for } x \geq c.$$

Suppose that $x_1, x_2, \ldots, x_n$ are independent observations from this distribution.

(a) Show that $\hat{c} = x_{(1)}$, the smallest observation, and find the RLF of c.
(b) Find an expression for the $100p\%$ likelihood interval for c.

11. Suppose that $x_1, x_2, \ldots, x_n$ are independent observations from the continuous uniform distribution over the interval $[0, \theta]$. Show that the likelihood function of θ is proportional to θ^{-n} for $\theta \geq x_{(n)}$, and is zero otherwise. Hence determine the MLE and RLF of θ.

12.† Suppose that $x_1, x_2, \ldots, x_n$ are independent observations from the continuous uniform distribution over the interval $[\theta, 2\theta]$. Find the RLF of θ.

13. Suppose that X and Y are continuous variates with joint probability density function

$$f(x, y) = e^{-\theta x - y/\theta} \qquad \text{for } x > 0,\ y > 0.$$

Find the MLE and RLF of θ on the basis of n independent pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$.

14. Independent measurements $x_1, x_2, \ldots, x_n$ are taken at unit time intervals. For $i = 1, 2, \ldots, \theta$ the measurements come from a standardized normal distribution $N(0, 1)$. A shift in the mean occurs after time θ, and for $i = \theta + 1, \theta + 2, \ldots, n$ the measurements come from $N(1, 1)$.

(a) Show that the likelihood function of θ is proportional to

$$\exp\left\{-\sum_{i=1}^{\theta}(x_i - \tfrac{1}{2})\right\}.$$

(b) Graph the log RLF for θ on the basis of the following set of 20 consecutive measurements:

−1.26	−0.16	−0.64	0.56	−1.82	−0.76	−2.08	−0.58	0.14	
0.94	−0.58	0.78	1.80	0.58	0.02	0.86	2.30	1.80	0.84	−0.18

Which values of θ have relative likelihood 10% or more?

15.* The p.d.f. of the double exponential distribution is

$$f(x) = \tfrac{1}{2}\exp\{-|x - \theta|\} \qquad \text{for } -\infty < x < \infty,$$

where $-\infty < \theta < \infty$. Let $x_1, x_2, \ldots, x_n$ be independent observations from this distribution, and let $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ denote these n observed values arranged in nondecreasing order.

(a) Show that, if n is odd, then $\hat{\theta} = x_{(m)}$ where $n = 2m - 1$.
(b) Show that, if $n = 2m$, then $l(\theta)$ is maximized for any value of θ between $x_{(m)}$ and $x_{(m+1)}$, and so $\hat{\theta}$ is not unique.

9.5. Censoring in Lifetime Experiments

In many experiments, the quantity of interest is the lifetime (or time to failure) of a specimen; for instance, the lifetime of an electronic component, or the length of time until an aircraft component fails from metal fatigue, or the survival time of a cancer patient after a new treatment.

The probability model generally assumes the lifetime X to be a continuous variate with some particular probability density function f and cumulative distribution function F. For example, if we thought that the chance of failure did not depend upon the age of the specimen, we would assume an exponential distribution. Lifetime distributions for situations in which the risk of failure increases or decreases with age were considered in Section 6.4.

The model will usually involve one or more unknown parameters θ which require estimation from the data.

Suppose that n specimens are tested independently. If the experiment is continued sufficiently long for all of the items to have failed, the likelihood function for θ based on the n observed lifetimes $x_1, x_2, \ldots, x_n$ can be obtained as in the last section. However, one might wait a very long time indeed for all of the specimens to fail, and it is often desirable to analyze the data before this happens. One or two hardy specimens may tie up a laboratory for months or years without greatly adding to the information about θ, at the same time preventing other experiments from being undertaken. It often makes good practical sense to terminate the experiment before all n items have failed.

If the ith specimen has failed by the time the experiment terminates, we will know its lifetime x_i. This will actually correspond to a real interval $a_i < X \le b_i$, say, with probability

$$P(a_i < X \le b_i) = F(b_i) - F(a_i) \approx f(x_i)\Delta_i,$$

provided that the time interval $\Delta_i = b_i - a_i$ is small.

If the jth specimen has not failed when the experiment ends, we will not know its lifetime, and the lifetime is said to be *censored*. The *censoring time* T_j is the total time for which the specimen had been tested when the experiment ended. For this specimen, we know only that $T_j < X < \infty$, and the probability of this event is

$$P(T_j < X < \infty) = F(\infty) - F(T_j) = 1 - F(T_j).$$

The likelihood function of θ will be a product of n factors, one for each specimen tested. Suppose that m specimens fail and $n - m$ do not, so that we have m failure times $x_1, x_2, \ldots, x_m$, and $n - m$ censoring times $T_1, T_2, \ldots, T_{n-m}$. Then the likelihood function of θ will be proportional to

$$\left[\prod_{i=1}^{m} f(x_i)\Delta_i\right]\prod_{j=1}^{n-m}[1 - F(T_j)].$$

The Δ_i's do not depend upon θ and can be absorbed into the proportionality constant to give

$$L(\theta) = c\left[\prod_{i=1}^{m} f(x_i)\right]\prod_{j=1}^{n-m}[1 - F(T_j)]. \tag{9.5.1}$$

where c is any convenient positive constant. The maximum likelihood estimate and RLF can now be obtained.

Special Case: Exponential Distribution

If X is assumed to have an exponential distribution with mean θ, then

$$f(x) = \frac{1}{\theta}e^{-x/\theta}; \qquad F(x) = 1 - e^{-x/\theta} \qquad \text{for } x > 0.$$

In this case, (9.5.1) simplifies to give

$$L(\theta) = \left[\prod_{i=1}^{m} \frac{1}{\theta} e^{-x_i/\theta}\right] \prod_{j=1}^{n-m} e^{-T_j/\theta} = \theta^{-m} e^{-s/\theta},$$

where s is the total elapsed lifetime (time on test) for all n items:

$$s = \sum_{i=1}^{m} x_i + \sum_{j=1}^{n-m} T_j.$$

The log likelihood function is

$$l(\theta) = -m \log \theta - \frac{s}{\theta},$$

and solving $S(\theta) = 0$ gives $\hat{\theta} = s/m$. The log RLF is then

$$r(\theta) = l(\theta) - l(\hat{\theta}).$$

EXAMPLE 9.5.1. Consider the experiment described in Example 9.4.1. Suppose that the $n = 10$ components were placed on test simultaneously, and it was decided to terminate the experiment after 50 days. The ten actual lifetimes are shown in Figure 9.5.1. If testing stopped at 50 days, everything to the right of 50 would be hidden from view, or censored. The data would then be

$$50+ \quad 11 \quad 50+ \quad 5 \quad 20 \quad 4 \quad 35 \quad 40 \quad 29 \quad 8,$$

where 50+ indicates that the first and third lifetimes were censored at 50 days.

In the notation defined above, we have $m = 8$ lifetimes with total $11 + 5 + 20 + \cdots + 8 = 152$, and $n - m = 2$ censoring times with total $50 + 50 = 100$. The total elapsed lifetime for all 10 components is

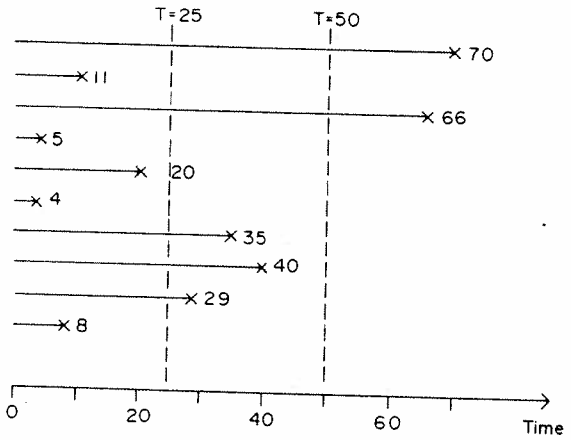

Figure 9.5.1. Diagrammatic representation of lifetime data showing two possible censoring times.

$s = 152 + 100 = 252$. Hence $\hat{\theta} = \frac{252}{8} = 31.5$, and

$$l(\theta) = -8 \log \theta - \frac{252}{\theta}.$$

If it had been decided to terminate the experiment after 25 days, the data would have been

$$25+ \quad 11 \quad 25+ \quad 5 \quad 20 \quad 4 \quad 25+ \quad 25+ \quad 25+ \quad 8.$$

There are now $m = 5$ lifetimes with total 48, and $n - m = 5$ censoring times with total 125, giving $s = 173$ and $\hat{\theta} = 34.6$. The log likelihood function is now

$$l(\theta) = -5 \log \theta - \frac{173}{\theta}.$$

Figure 9.5.2 shows the three log relative likelihood functions resulting from (i) stopping the experiment after $T = 25$ days, (ii) stopping the experiment after $T = 50$ days, and (iii) continuing the experiment until all of the components have failed (i.e. stopping at time $T > 70$). The three functions agree reasonably well for $\theta \leq 30$, indicating that plausibilities of small parameter values are affected very little even when 50% of the lifetimes are censored. However, the three curves diverge considerably for large values of

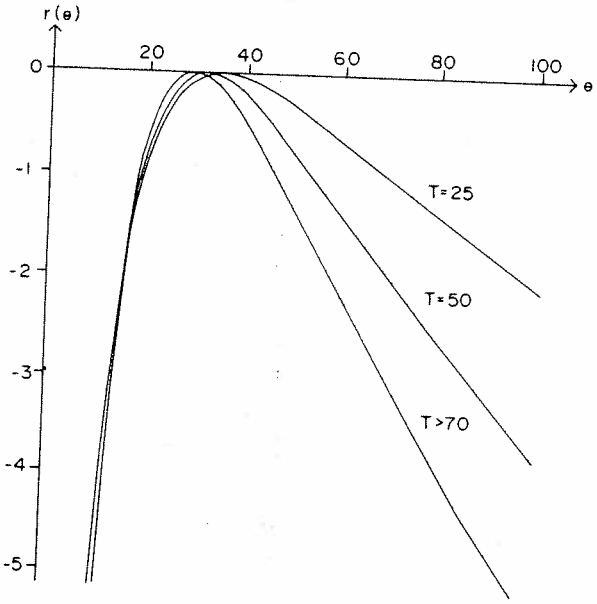

Figure 9.5.2. Log relative likelihood function for the exponential mean θ under various levels of censoring.

θ. With no censoring, values of θ greater than 62 are implausible ($R < 0.1$); with censoring at 25 days, θ can be as large as 108 before R decreases to 10%. Censoring thus makes it impossible to place as tight an upper bound on the value of θ, but has little effect on the lower bound. These results suggest that if we were primarily interested in establishing a lower bound for θ, a short experiment with heavy censoring could be quite satisfactory.

Note. In applications, the appropriate analysis will normally be that which corresponds to the pattern of censoring actually used in the experiment. However, in some cases one might also wish to examine the likelihood function that would result from more severe censoring in order to see what effect a few large lifetimes have on the analysis.

Problems for Section 9.5

1. Ten electronic components with exponentially distributed lifetimes were tested for predetermined periods of time as shown. Three of the tubes survived their test periods, and the remaining seven failed at the times shown.

Tube number	1	2	3	4	5	6	7	8	9	10
Test period	81	72	70	60	41	31	31	30	29	21
Failure time	2	–	51	–	33	27	14	24	4	–

Find the MLE and the 10% likelihood interval for the exponential mean θ.

2.† n electronic components were simultaneously placed on test. After a time T testing was stopped. It was observed that $n-k$ were still operating and that k had failed, but the times at which the failures had occurred were not known. Assuming that failure times follow an exponential distribution with mean θ, derive the maximum likelihood estimate and the relative likelihood function of θ.

3. A clinical trial was conducted to determine whether a hormone treatment benefits women who were treated previously for breast cancer. A woman entered the clinical trial when she had a recurrence. She was then treated by irradiation, and assigned to either a hormone therapy group or a control group. The observation of interest is the time until a second recurrence, which may be assumed to follow an exponential distribution with mean θ_H (hormone therapy group) or θ_C (control group). Many of the women did not have a second recurrence before the clinical trial was concluded, so that their recurrence times are censored. In the following table, a censoring time "n" means that a woman was observed for time n, and did not have a recurrence, so that her recurrence time is known to exceed n. Plot the log RLF's of θ_H and θ_C on the same graph. Is there any indication that the hormone treatment increases the mean time to recurrence?

4.†* The cumulative distribution function for the lifetime of a new type of light bulb is assumed to be

$$F(x) = 1 - \left(1 + \frac{2x}{\theta}\right)e^{-2x/\theta} \qquad \text{for } x > 0$$

	Hormone treated						Control					
Recurrence times	2	4	6	9	9	9	1	4	6	7	13	24
	13	14	18	23	31	32	25	35	35	39		
	33	34	43									
Censoring times	10	14	14	16	17	18	1	1	3	4	5	8
	18	19	20	20	21	21	10	11	13	14	14	15
	23	24	29	29	30	30	17	19	20	22	24	24
	31	31	31	33	35	37	24	25	26	26	26	28
	40	41	42	42	44	46	29	29	32	35	38	39
	48	49	51	53	54	54	40	41	44	45	47	47
	55	56					47	50	50	51		

(a) Find the probability density function, and show that the mean of this distribution is θ.

(b) Forty bulbs were tested and failures occurred at the following times (in hours):

196	327	405	537	541	660	671	710	786
940	954	1004	1004	1006	1202	1459	1474	1484
1602	1662	1666	1711	1784	1796	1799		

The remaining bulbs had not failed when testing stopped at 1800 hours. Find the MLE and the 10% likelihood interval for θ.

5.* An arrow is shot at the center of a circular target of radius 1. Let X denote the horizontal displacement and Y the vertical displacement of the point of impact from the center of the target. It is to be assumed that X and Y are independent $N(0, \sigma^2)$ random variables.

(a) Show that the probability of a shot missing the target is

$$P(X^2 + Y^2 \geq 1) = \exp\{-1/2\sigma^2\}.$$

(b) Of n independent shots, m hit the target at points (x_i, y_i) for $i = 1, 2, \ldots, m$. The other $n-m$ shots miss the target, and their points of impact are not recorded. Find the MLE of σ.

9.6. Invariance

Suppose that the probability model for an experiment depends upon an unknown parameter θ. The model then consists of a whole family of probability distributions, one for each value of θ in the parameter space Ω. For example, we might assume that the time to failure of an electronic component has an exponential distribution, with probability density function

$$f(x) = \frac{1}{\theta}e^{-x/\theta} \qquad \text{for } 0 < x < \infty, \tag{9.6.1}$$

where θ is the expected lifetime. For each value of θ belonging to $\Omega = (0, \infty)$, we have a theoretical distribution. For instance, the distribution labeled by $\theta = 1$ is

$$f(x) = e^{-x} \qquad \text{for } 0 < x < \infty, \tag{9.6.2}$$

and the distribution labeled by $\theta = 2$ is

$$f(x) = \frac{1}{2} e^{-x/2} \qquad \text{for } 0 < x < \infty. \tag{9.6.3}$$

A family of distributions can be parametrized (or labeled) in many different ways. For instance, we could equally well write (9.6.1) as

$$f(x) = \lambda e^{-\lambda x} \qquad \text{for } 0 < x < \infty$$

where $\lambda = 1/\theta$ is the failure rate. Distributions (9.6.2) and (9.6.3) are now labeled by $\lambda = 1$ and $\lambda = 0.5$, respectively. We have the choice of labeling the family of exponential distributions by values of θ, or by values of λ, or by values of any other one-to-one function of θ. We usually try to select a parametrization so that the parameter represents some interesting characteristic of the distribution, and the mathematical expressions are fairly simple.

When we say that $\theta = 1$ is ten times as likely as $\theta = 2$, we imply that the distribution labeled by $\theta = 1$ is ten times as likely as the distribution labeled by $\theta = 2$. When we say that the maximum likelihood estimate of θ is $\hat{\theta} = 1.1$, we imply that the distribution labeled by $\theta = 1.1$ is the most likely distribution. Since the method of labeling the distributions is largely arbitrary, it would seem desirable that the plausibilities assigned to the distributions should not depend upon the particular method of labeling which has been selected. In other words, the plausibilities assigned should be *invariant* under one-to-one transformations of the parameter.

An attractive property of the likelihood methods we have discussed is that they *are* invariant under one-to-one parameter transformations. For suppose that $\theta = g(\lambda)$ where g is invertible, and let $P(E; \theta)$ be the probability of the observed event E. Substituting $\theta = g(\lambda)$ in this expression gives the probability of E as a function of λ. It follows that $L(\theta)$, the likelihood function of θ, and $L_*(\lambda)$, the likelihood function of λ, are related as follows:

$$L_*(\lambda) = L(\theta) \qquad \text{where } \theta = g(\lambda).$$

Hence both functions have the same maximum value, and

$$R_*(\lambda) = R(\theta) \qquad \text{where } \theta = g(\lambda).$$

If λ_1 is any possible value of λ and $\theta_1 = g(\lambda_1)$ is the corresponding value of θ, then λ_1 and θ_1 have the same relative likelihood. Relative likelihoods do not depend upon whether we choose to work with parameter λ or parameter θ.

It follows that, if $\hat{\lambda}$ is the MLE of λ, then $\hat{\theta} = g(\hat{\lambda})$ is the MLE of θ. Similarly, θ_1 belongs to the $100p\%$ likelihood region for θ if and only if $\theta_1 = g(\lambda_1)$ where λ_1 belongs to the $100p\%$ likelihood region for λ.

EXAMPLE 9.6.1. In Example 9.4.1, we supposed that the lifetimes of electronic components were exponentially distributed, with mean lifetime θ. On the basis of ten observations, we found that $\hat{\theta} = 28.8$. The 50% LI for θ was $20 \le \theta \le 43$, and the 10% LI was $16 \le \theta \le 62$.

(a) Suppose that we are interested in the failure rate, $\lambda = 1/\theta$. Then the MLE of λ is

$$\hat{\lambda} = \frac{1}{\hat{\theta}} = \frac{1}{28.8} = 0.0347.$$

The 50% LI for λ is obtained by noting that $20 \le 1/\lambda \le 43$ if and only if $1/20 \ge \lambda \ge 1/43$. Hence the 50% LI is $0.023 \le \lambda \le 0.050$. Similarly, the 10% LI is found to be $0.016 \le \lambda \le 0.063$.

(b) Suppose that we are interested in the proportion β of such components which will last at least 25 days. Then

$$\beta = P(X \ge 25) = \int_{25}^{\infty} \frac{1}{\theta} e^{-x/\theta} dx = e^{-25/\theta},$$

which is a one-to-one function of θ. Hence the MLE of β is

$$\hat{\beta} = e^{-25/\hat{\theta}} = 0.420.$$

Since $\theta = -25/\log \beta$, the 50% LI for β is given by

$$20 \le -\frac{25}{\log \beta} \le 43$$

and solving for β gives $0.287 \le \beta \le 0.559$. Similarly, the 10% LI is $0.210 \le \beta \le 0.668$.

PROBLEMS FOR SECTION 9.6

1. Let γ denote the median lifetime of electronic components in Example 9.4.1. Show that $\gamma = \theta \log 2$, and hence obtain the MLE and the 10% likelihood interval for γ.

2. We wish to estimate p, the probability of no diseased trees in a four-acre plot, in Problem 9.1.1. One approach would be to note that 4 out of 10 plots contained no diseased trees, so that $\hat{p} = 0.4$ and $L(p) = p^4(1-p)^6$. A second approach would be to express p as a function of λ and use the invariance property of likelihood. Determine the MLE and the 10% likelihood interval for p by both methods. Under what conditions would the first method be preferable?

3.† The following table summarizes information concerning the lifetimes of one hundred V600 indicator tubes. (Ref.: D. J. Davis, *Journal of the American Statistical Association 47* (1952), 113–150).

Lifetime (hours)	0–100	100–200	200–300	300–400	400–600
Frequency observed	29	22	12	10	10
Lifetime (hours)	600–800	800+			
Frequency observed	9	8			

Suppose that the lifetimes follow an exponential distribution with mean θ.

(a) Show that the joint probability distribution of the frequencies is multinomial with probabilities

$$p_1 = P(0 < T < 100) = 1 - \beta; \qquad p_2 = P(100 < T < 200) = \beta(1-\beta); \ldots;$$

$$p_7 = P(T > 800) = \beta^8, \qquad \text{where } \beta = e^{-100/\theta}.$$

(b) Show that $\hat{\beta}$ can be obtained as a root of a quadratic equation, and deduce the value of $\hat{\theta}$.

(c) Prepare a graph of the log RLF of β. Obtain 10% and 50% likelihood intervals for β, and transform them into likelihood intervals for θ.

4.* The arrivals of westbound vehicles at a fixed point on an east–west road are random events in time. On the average there are μ arrivals per ten second interval. A traffic signal is to be installed a short distance beyond the observation point. It is desired that the signal remain at "STOP" for a time β such that the probability of holding up k or more vehicles is p.

(a) Show that $p = P(\chi^2_{(2k)} \le \beta\mu/5)$. In particular, if $k = 8$ and $p = 0.05$, then $\beta = 39.8/\mu$.

(b) Assuming that $k = 8$ and $p = 0.05$, use the data in Problem 9.3.2 to determine $\hat{\beta}$ and the 10% likelihood interval for β.

9.7. Normal Approximations

Let $l(\theta)$ denote the log likelihood function of a continuous parameter θ with possible values in Ω. Let $S(\theta) = l'(\theta)$ and $\mathcal{I}(\theta) = -l''(\theta)$ denote the score and information functions as in Section 1. We assume that $\hat{\theta}$ exists and is an interior point of Ω, and that $l(\theta)$ has a Taylor's series expansion at $\theta = \hat{\theta}$:

$$l(\theta) = l(\hat{\theta}) + \frac{\theta-\hat{\theta}}{1!}l'(\hat{\theta}) + \frac{(\theta-\hat{\theta})^2}{2!}l''(\hat{\theta}) + \frac{(\theta-\hat{\theta})^3}{3!}l'''(\hat{\theta}) + \cdots.$$

Since $l'(\hat{\theta}) = 0$ and $r(\theta) = l(\theta) - l(\hat{\theta})$, we have

$$r(\theta) = -\frac{1}{2}(\theta-\hat{\theta})^2\mathcal{I}(\hat{\theta}) + \frac{(\theta-\hat{\theta})^3}{3!}l'''(\hat{\theta}) + \cdots. \tag{9.7.1}$$

The *normal approximation to* $r(\theta)$ is defined as follows:

$$r_N(\theta) = -\frac{1}{2}(\theta-\hat{\theta})^2\mathcal{I}(\hat{\theta}). \tag{9.7.2}$$

If $|\theta - \hat{\theta}|$ is small, the cubic and higher terms in (9.7.1) are small, and hence $r(\theta) \approx r_N(\theta)$.

The effect of increasing the amount of data is to produce a sharply peaked likelihood function and shorter likelihood intervals (see Example 9.3.1). Thus, for a sufficiently large sample, $|\theta - \hat{\theta}|$ will be small and $r_N(\theta)$ will give a good approximation to $r(\theta)$ over the entire region of plausible parameter values.

The $100p\%$ likelihood region for θ is the set of θ-values such that $R(\theta) \ge p$,

or equivalently, $r(\theta) \ge \log p$. Taking $r_N(\theta) \ge \log p$ gives

$$\theta \in \hat{\theta} \pm \sqrt{(-2\log p)/\mathcal{I}(\hat{\theta})} \tag{9.7.3}$$

as an approximation to the $100p\%$ likelihood region. This is an interval centered at $\hat{\theta}$ with length

$$2\sqrt{(-2\log p)/\mathcal{I}(\hat{\theta})}.$$

The larger the value of $\mathcal{I}(\hat{\theta})$, the narrower the approximate interval will be, and hence the more information we have concerning θ. This is the reason that $\mathcal{I}(\theta)$ is called the "information function".

When the normal approximation is sufficiently accurate, all of the information concerning θ is summarized in $\hat{\theta}$ and $\mathcal{I}(\hat{\theta})$. The MLE indicates the most likely parameter value, and $\mathcal{I}(\hat{\theta})$ indicates the precision with which θ can be determined. Given these two values, likelihood intervals of any sizes desired can be obtained from (9.7.3).

How large a sample is necessary before the normal approximation $r(\theta) \approx r_N(\theta)$ can be used? This depends very much on the situation. In the first example below we find that $r(\theta) = r_N(\theta)$ exactly for all sample sizes, but in the second example the approximation is not very good even with 500 observations. Thus it is necessary to check the accuracy of the normal approximation in each new situation. We can do this by plotting both $r(\theta)$ and $r_N(\theta)$ on the same graph and verifying that they agree closely for values of θ inside, say, the 10% likelihood interval. Alternatively, we can check that a graph of the score function $S(\theta)$ is well approximated by a straight line over this interval.

EXAMPLE 9.7.1. Let $x_1, x_2, \ldots, x_n$ be independent observations from a normal distribution with unknown mean μ and known variance σ^2. If the measurement intervals are small, the likelihood function of μ is proportional to the product of probability density functions:

$$L(\mu) = c \cdot \prod_{i=1}^{n} f(x_i) = c \cdot \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2\right\}$$

by choice of c. Hence the log likelihood, score, and information functions are

$$l(\mu) = -\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2; \qquad S(\mu) = \frac{1}{\sigma^2}\Sigma(x_i-\mu); \qquad \mathcal{I}(\mu) = \frac{n}{\sigma^2}.$$

Solving $S(\mu) = 0$ gives $\hat{\mu} = \frac{1}{n}\Sigma x_i = \bar{x}$, and hence the log relative likelihood function is

$$r(\mu) = l(\mu) - l(\hat{\mu}) = -\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2 + \frac{1}{2\sigma^2}\Sigma(x_i-\bar{x})^2.$$

Upon expanding the squares and simplifying, we get

$$r(\mu) = -\frac{n}{2\sigma^2}(\mu - \bar{x})^2 = -\frac{1}{2}(\mu - \hat{\mu})^2 \mathcal{I}(\hat{\mu}).$$

In this example we have $r(\mu) = r_N(\mu)$ for all μ, and indeed it is for this reason that we call $r_N(\mu)$ the normal approximation. By (9.7.3), the exact $100p\%$ likelihood interval for μ is given by

$$\mu \in \bar{x} \pm \sigma\sqrt{(-2\log p)/n}.$$

EXAMPLE 9.7.2. Suppose that $x_1, x_2, \ldots, x_n$ are independent observations from an exponential distribution with unknown mean θ. We assume that the measurement intervals are small. Then, from Example 9.4.1, the log likelihood, score, and information functions are

$$l(\theta) = -n\log\theta - \frac{1}{\theta}\Sigma x_i; \qquad S(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\Sigma x_i;$$

$$\mathcal{I}(\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3}\Sigma x_i.$$

Also we have $\hat{\theta} = \bar{x}$ and $\mathcal{I}(\hat{\theta}) = n/\hat{\theta}^2$. Hence the relative likelihood function and normal approximation are

$$r(\theta) = l(\theta) - l(\hat{\theta}) = -n\left[\frac{\hat{\theta}}{\theta} - 1 - \log\frac{\hat{\theta}}{\theta}\right];$$

$$r_N(\theta) = -\frac{1}{2}(\theta - \hat{\theta})^2 \mathcal{I}(\hat{\theta}) = -\frac{n}{2}\left[\frac{\theta}{\hat{\theta}} - 1\right]^2.$$

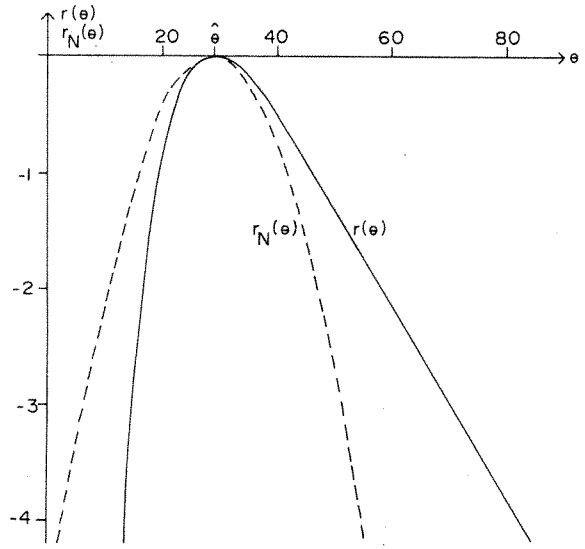

Figure 9.7.1. Log relative likelihood function and normal approximation.

These two functions are plotted in Figure 9.7.1 for the case $n = 10$, $\hat{\theta} = 28.8$ (see Example 9.4.1). The agreement is very poor because $r(\theta)$ is highly skewed, but $r_N(\theta)$ is symmetrical about the line $\theta = \hat{\theta}$. The exact 10% likelihood interval is $15.65 \le \theta \le 61.88$ (see Example 9.8.2), but (9.7.3) gives $9 \le \theta \le 48$.

With $n = 500$ and $\hat{\theta} = 28.8$, the exact 10% likelihood interval is $26.20 \le \theta \le 31.75$, and (9.7.3) gives $26.04 \le \theta \le 31.56$. The agreement, although much better than for $n = 10$, is still not very good. It is true that $r(\theta) \to r_N(\theta)$ as $n \to \infty$, but n must be very large before the normal approximation is accurate enough to use.

Parameter Transformations

We noted in Section 6 that relative likelihoods are invariant under one-to-one parameter transformations $\theta = g(\lambda)$. However, the normal approximation (9.7.2) is not invariant. Because of this, it may be possible to achieve greater accuracy by transforming from θ to a new parameter λ before approximating.

Since (9.7.2) is obtained by ignoring the cubic and higher terms in (9.7.1), it makes sense to look for a transformation which reduces the size of the cubic term relative to the quadratic term. Hopefully this will improve the accuracy of the normal approximation. We can then obtain approximate likelihood intervals for λ and transform them via g into intervals for θ.

EXAMPLE 9.7.2 (continued). Consider the family of power transformations $\theta = \lambda^a$ where $a \neq 0$. Then $\hat{\theta} = \hat{\lambda}^a$, and the log likelihood function of λ is

$$l_*(\lambda) = l(\lambda^a) = -na\log\lambda - n\hat{\lambda}^a/\lambda^a.$$

The first three derivatives with respect to λ are

$$l_*'(\lambda) = -na/\lambda + na\hat{\lambda}^a/\lambda^{a+1};$$

$$l_*''(\lambda) = na/\lambda^2 - na(a+1)\hat{\lambda}^a/\lambda^{a+2};$$

$$l_*'''(\lambda) = -2na/\lambda^3 + na(a+1)(a+2)\hat{\lambda}^a/\lambda^{a+3}.$$

Thus we have

$$\mathcal{I}_*(\hat{\lambda}) = -l_*''(\hat{\lambda}) = na^2/\hat{\lambda}^2; \qquad l_*'''(\hat{\lambda}) = na^2(a+3)/\hat{\lambda}^3.$$

The third derivative is zero for $a = -3$; that is, for $\theta = \lambda^{-3}$. Thus if $\lambda = \theta^{-1/3}$, the cubic term in the Taylor's series expansion of $l_*(\lambda)$ about $\lambda = \hat{\lambda}$ is zero, and the normal approximation to $l_*(\lambda)$ should produce accurate results.

The normal approximation to $r_*(\lambda)$ is

$$r_*(\lambda) \approx -\tfrac{1}{2}(\lambda - \hat{\lambda})^2 \mathcal{I}_*(\hat{\lambda}).$$

This is compared with $r(\theta)$ and $r_N(\theta)$ in Table 9.7.1 for the case $n = 10$, $\hat{\theta} = 28.8$. Transforming from θ to the new parameter $\lambda = \theta^{-1/3}$ has substantially improved the accuracy of the normal approximation.

Table 9.7.1. Comparison of Normal Approximations

θ	$\lambda=\theta^{-1/3}$	$r(\theta)=r_*(\lambda)$	$-\frac{1}{2}(\theta-\hat\theta)^2\mathscr{I}(\hat\theta)$	$-\frac{1}{2}(\lambda-\hat\lambda)^2\mathscr{I}_*(\hat\lambda)$
12	0.437	−5.25	−1.70	−5.17
15	0.405	−2.68	−1.15	−2.65
20	0.368	−0.75	−0.47	−0.75
25	0.342	−0.10	−0.09	−0.10
40	0.292	−0.49	−0.76	−0.48
60	0.255	−2.14	−5.87	−2.12
80	0.232	−3.82	−15.80	−3.75
100	0.215	−5.33	−30.56	−5.19

By (9.7.3), the approximate $100p\%$ LI for λ is

$$\lambda\in\hat\lambda\pm\sqrt{(-2\log p)/\mathscr{I}_*(\hat\lambda)}=\hat\lambda[1\pm\sqrt{(-2\log p)/9n}].$$

Transforming this via $\theta=\lambda^{-3}$ gives

$$\theta\in\hat\theta[1\pm\sqrt{(-2\log p)/9n}]^{-3}$$

as the approximate $100p\%$ LI for θ. For $n=10$, $\hat\theta=28.8$ this gives $15.62\le\theta\le62.16$ as the approximate 10% LI whereas the exact result is $15.65\le\theta\le61.88$. By transforming to the new parameter $\lambda=\theta^{-1/3}$, we are able to achieve greater accuracy for $n=10$ than we obtained previously with $n=500$.

Transforming the Information Function

Suppose that we change parameters from θ to λ via the one-to-one transformation $\theta=g(\lambda)$. By the invariance property, the log likelihood function of λ is

$$l_*(\lambda)=l(\theta).$$

The score function of λ is

$$S_*(\lambda)=\frac{d}{d\lambda}l_*=\frac{dl}{d\theta}\frac{d\theta}{d\lambda}=S(\theta)\cdot\frac{d\theta}{d\lambda}.$$

The information function of λ is

$$\mathscr{I}_*(\lambda)=-\frac{d}{d\lambda}S_*=-S(\theta)\frac{d^2\theta}{d\lambda^2}-\frac{dS}{d\theta}\cdot\left(\frac{d\theta}{d\lambda}\right)^2$$

$$=-S(\theta)\frac{d^2\theta}{d\lambda^2}+\mathscr{I}(\theta)\left(\frac{d\theta}{d\lambda}\right)^2.$$

At the maximum we have $S(\hat\theta)=0$, and therefore

$$\mathscr{I}_*(\hat\lambda)=\hat q^2\mathscr{I}(\hat\theta) \tag{9.7.4}$$

where $\hat q$ is the value of $d\theta/d\lambda$ at the maximum.

If $\hat\theta$ and $\mathscr{I}(\hat\theta)$ are known, $\hat\lambda$ can be found by solving the equation $\hat\theta=g(\hat\lambda)$, and $\mathscr{I}_*(\hat\lambda)$ can be found from (9.7.4). The normal approximation to $r_*(\lambda)$ can then be written down. Given $\hat\theta$ and $\mathscr{I}(\hat\theta)$, not much extra work is needed to find the normal approximation for any one-to-one function of θ.

Problems for Section 9.7

1. Obtain approximate 10% likelihood intervals for θ in Problems 9.1.4(b) and 9.1.5(b), and investigate the accuracy of the normal approximation to $r(\theta)$ in these examples.

2. Find approximate 10% likelihood intervals for θ_H and θ_C in Problem 9.5.3. Repeat using the transformed parameters $\lambda_H=\theta_H^{-1/3}$ and $\lambda_C=\theta_C^{-1/3}$ as in Example 9.7.2. Compare your results with the exact 10% LI's for θ_H and θ_C.

3.† Consider the situation described in Problem 9.4.6. A decision must be made as to the number n of measurements which will be taken to estimate the unknown concentration μ of a trace metal in solution. It is desired that the 10% LI for μ should have width at most 2 units. Determine the appropriate value of n as a function of the error variance σ^2.

4. Suppose that X has a binomial (n,θ) distribution, and consider the series expansion of $l(\theta)$ about $\theta=\hat\theta$. Show that the ratio of the cubic term to the quadratic term in this expansion is

$$\frac{2}{3}(\hat\theta-\theta)(1-2\hat\theta)/\hat\theta(1-\hat\theta).$$

Under what conditions will the normal approximation to $r(\theta)$ be satisfactory?

5. Suppose that x successes are observed in n Bernoulli trials with success probability θ. Let W denote the width of the approximate $100p\%$ LI for θ.

(a) Find an expression for W, and show that

$$W\le\sqrt{(-2\log p)/n}.$$

(b) How large must n be in order to ensure that the approximate 10% LI for θ has width at most 0.04?

6. Suppose that two independent experiments both give information about the same parameter θ, and that

$$r_1(\theta)\approx-\tfrac{1}{2}(\theta-\hat\theta_1)^2\hat{\mathscr{I}}_1;\qquad r_2(\theta)\approx-\tfrac{1}{2}(\theta-\hat\theta_2)^2\hat{\mathscr{I}}_2$$

where $\hat{\mathscr{I}}_1=\mathscr{I}_1(\hat\theta_1)$ and $\hat{\mathscr{I}}_2=\mathscr{I}_2(\hat\theta_2)$. Show that the overall MLE based on both experiments is given approximately by

$$\hat\theta=(\hat{\mathscr{I}}_1\hat\theta_1+\hat{\mathscr{I}}_2\hat\theta_2)/(\hat{\mathscr{I}}_1+\hat{\mathscr{I}}_2),$$

and that this value lies between $\hat\theta_1$ and $\hat\theta_2$.

7. (a) Let $X_1,X_2,\ldots,X_n$ be independent Poisson variates with mean μ, and consider transformations of the form $\mu=\lambda^a$ where $a\neq0$. Find the log likelihood function of λ, and show that the cubic term in the expansion of this function about $\lambda=\hat\lambda$ is zero for $a=3$.

(b) Obtain an approximate 10% LI for $\lambda = \mu^{1/3}$, and transform it to obtain an interval for μ.

(c) Suppose that $n = 10$ and $\Sigma x_i = 53$. Using a table or graph, investigate the accuracy of the normal approximations to the original log RLF of μ, and to the log RLF of the transformed parameter $\lambda = \mu^{1/3}$.

9.8. Newton's Method

In this section we describe two applications of Newton's iterative method for solving an equation.

Suppose that we wish to find a root $\hat{\theta}$ of the equation $g(\theta) = 0$. Let θ_0 be a parameter value close to $\hat{\theta}$ and consider the Taylor's series expansion of $g(\theta)$ about $\theta = \theta_0$:

$$g(\theta) = g(\theta_0) + (\theta - \theta_0)g'(\theta_0) + (\theta - \theta_0)^2 g''(\theta_0)/2! + \cdots .$$

For $|\theta - \theta_0|$ small, the quadratic and higher terms in this expansion will be small, and dropping these terms gives

$$g(\theta) \approx g(\theta_0) + (\theta - \theta_0)g'(\theta_0).$$

We are approximating $g(\theta)$ by a linear function of θ which has the same value and slope as $g(\theta)$ at $\theta = \theta_0$.

Since $g(\hat{\theta}) = 0$, we have

$$g(\theta_0) + (\hat{\theta} - \theta_0)g'(\theta_0) \approx 0,$$

and therefore

$$\hat{\theta} \approx \theta_0 - g(\theta_0)/g'(\theta_0).$$

In Newton's method we take θ_0 to be a preliminary guess at $\hat{\theta}$, and then compute a revised guess θ_1 as follows:

$$\theta_1 = \theta_0 - g(\theta_0)/g'(\theta_0). \tag{9.8.1}$$

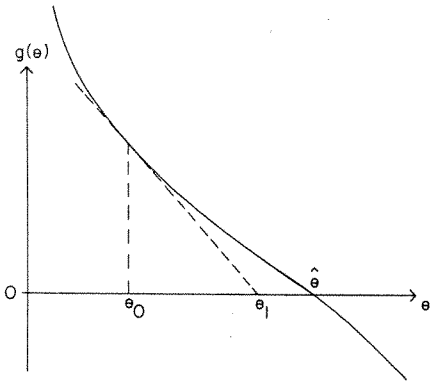

Figure 9.8.1. Solution of $g(\theta) = 0$ by Newton's method.

The revised guess is the point at which the linear approximation (tangent) to $g(\theta)$ at $\theta = \theta_0$ crosses the θ-axis (see Figure 9.8.1). We now take θ_1 as the new preliminary guess and repeat the calculation to get

$$\theta_2 = \theta_1 - g(\theta_1)/g'(\theta_1).$$

We continue this procedure until $\theta_{i+1} \approx \theta_i$, in which case $g(\theta_i) = 0$ and a root has been found.

Solving the Maximum Likelihood Equation

We noted in Section 9.1 that, under suitable conditions, $\hat{\theta}$ is a root of the maximum likelihood equation $S(\theta) = 0$, where $S(\theta) = l'(\theta)$. Taking $g(\theta) = S(\theta)$ in the above derivation, we have $g'(\theta) = S'(\theta) = -\mathcal{I}(\theta)$ (see Section 9.1). Thus the updating formula (9.8.1) becomes

$$\theta_1 = \theta_0 + S(\theta_0)/\mathcal{I}(\theta_0). \tag{9.8.2}$$

Starting with an initial guess θ_0, we repeatedly update to get $\theta_1, \theta_2, \theta_3, \ldots$. We stop as soon as $\theta_{i+1} \approx \theta_i$, so that $S(\theta_i) \approx 0$. To verify that a relative maximum has been found, we check that $\mathcal{I}(\theta_i) > 0$.

Newton's method works well in most statistical applications. If the initial guess is reasonable, the procedure usually produces an accurate approximation to $\hat{\theta}$ after only three or four iterations. The reason for this is that, for moderately large samples, $S(\theta)$ is nearly linear in θ (see Section 9.7). If $S(\theta)$ is exactly linear in θ, Newton's method produces $\hat{\theta}$ in a single iteration.

If $S(\theta) = 0$ has more than one root, Newton's method will not necessarily converge to the one desired. Difficulties can also arise if the maximum occurs on or near a boundary of the parameter space. It is a good idea to examine a graph of $l(\theta)$ before applying Newton's method.

EXAMPLE 9.8.1. Newton's method will be used to obtain the overall MLE $\hat{\mu}$ in Example 9.2.2. The score and information functions are

$$S(\mu) = \frac{120}{1-p_1} + \frac{3}{1-p_2} - 440; \qquad \mathcal{I}(\mu) = \frac{1200p_1}{(1-p_1)^2} + \frac{3p_2}{(1-p_2)^2}$$

where $p_1 = e^{-10\mu}$ and $p_2 = e^{-\mu}$. The calculations are summarized in Table 9.8.1.

A convenient choice for the initial guess μ_0 is the average of the individual estimates in Example 9.2.2:

$$\mu_0 = \frac{1}{2}(0.0357 + 0.0780) = 0.057.$$

Now we find that

$$S(\mu_0) = -109.66; \qquad \mathcal{I}(\mu_0) = 4518.16$$

and (9.8.2) gives

$$\mu_1 = 0.057 - 109.66/4518.16 = 0.03273.$$

Table 9.8.1. Solution of $S(\mu) = 0$ by Newton's Method

i	μ_i	$S(\mu_i)$	$\mathscr{I}(\mu_i)$	μ_{i+1}
0	0.057	−109.66	4518.16	0.03273
1	0.03273	83.07	13902.58	0.03871
2	0.03871	12.87	9910.74	0.04001
3	0.04001	0.41	9270.86	0.04005
4	0.04005	0.04	9252.15	0.04005

At the next step we compute

$$S(\mu_1) = 83.07; \qquad \mathscr{I}(\mu_1) = 13902.58$$

and hence obtain

$$\mu_2 = 0.03273 + 83.07/13902.58 = 0.03871.$$

Continuing in this fashion, we obtain $\hat{\mu} = 0.04005$ correct to five decimal places. Note that $\mathscr{I}(\hat{\mu}) > 0$, so a relative maximum has been found.

Likelihood Interval Calculation

In previous examples we found likelihood intervals from a graph of the log relative likelihood function $r(\theta)$. Alternatively, we can obtain the endpoints of the $100p\%$ likelihood interval by solving the equation $g(\theta) = 0$, where

$$g(\theta) = r(\theta) - \log p.$$

Usually numerical methods will be required, and Newton's iterative method can again be used.

Since $r(\theta) = l(\theta) - l(\hat{\theta})$, it follows that $g'(\theta) = l'(\theta) = S(\theta)$, and so (9.8.1) gives

$$\theta_1 = \theta_0 - [r(\theta_0) - \log p]/S(\theta_0). \tag{9.8.3}$$

Calculation of the right endpoint is illustrated in Figure 9.8.2. We begin with a preliminary estimate θ_0 for the endpoint. The revised estimate θ_1 is the θ-value at which the tangent to $g(\theta)$ at $\theta = \theta_0$ crosses the θ-axis. The calculation is repeated with the revised value as the new initial estimate. We continue in this way until convergence to the right endpoint is obtained. A second iteration is then carried out for the left endpoint.

Starting values for Newton's method can be taken from a preliminary graph of $r(\theta)$. Alternatively, they can often be obtained from the normal approximation (9.7.2), which gives

$$\theta = \hat{\theta} \pm \sqrt{(-2 \log p)/\mathscr{I}(\hat{\theta})}$$

as approximations to the interval endpoints.

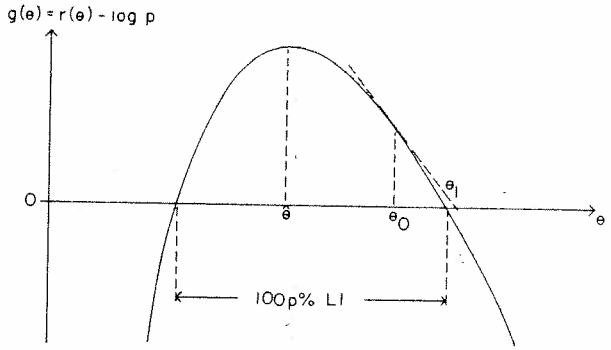

Figure 9.8.2. Solution of $r(\theta) - \log p = 0$ by Newton's method.

EXAMPLE 9.8.2. Newton's method will be used to obtain the 10% likelihood interval for θ in Example 9.4.1. For this example we have

$$l(\theta) = -10 \log \theta - \frac{288}{\theta}; \qquad S(\theta) = -\frac{10}{\theta} + \frac{288}{\theta^2}; \qquad \mathscr{I}(\theta) = -\frac{10}{\theta^2} + \frac{576}{\theta^3}.$$

The MLE is $\hat{\theta} = 28.8$, and so

$$l(\hat{\theta}) = -43.604; \qquad \mathscr{I}(\hat{\theta}) = 0.01206.$$

Thus the log relative likelihood function is

$$r(\theta) = l(\theta) - l(\hat{\theta}) = -10 \log \theta - \frac{288}{\theta} + 43.604,$$

and (9.8.4) gives

$$\theta = 28.8 \pm \sqrt{(-2 \log 0.1)/0.01206} = 28.8 \pm 19.5.$$

Table 9.8.2 shows the calculation of the left endpoint with initial estimate $28.8 - 19.5 = 9.3$. After five iterations, the left endpoint is found to be 15.65,

Table 9.8.2. Calculation of 10% LI by Newton's Method

i	θ_i	$r(\theta_i)$	$S(\theta_i)$	θ_{i+1}
0	9.30	−9.664	2.255	12.57
1	12.57	−4.621	1.027	14.83
2	14.83	−2.783	0.635	15.59
3	15.59	−2.336	0.544	15.65
4	15.65	−2.304	0.536	15.65
0	48.30	−1.133	−0.0836	62.29
1	62.29	−2.338	−0.0863	61.88
2	61.88	−2.302	−0.0864	61.88

correct to two decimal places. Similarly, the initial value for the right endpoint is $28.8 + 19.5 = 48.3$, and the final value is 61.88 after three iterations. Thus the 10% likelihood interval is $15.65 \le \theta \le 61.88$.

Problems for Section 9.8

1.† Use Newton's method to locate the maximum of the following log likelihood function:

$$l(\mu) = 100 \log \mu - 50\,\mu - 50 \log(1 - e^{-\mu}) \qquad \text{for } \mu > 0.$$

2. Suppose that the score function is linear in θ,

$$S(\theta) = a\theta + b \qquad \text{for } -\infty < \theta < \infty,$$

where a, b are constants with $a < 0$. Show that Newton's method converges to $\hat{\theta}$ in one iteration for any starting value θ_0.

3. Samples of river water are placed in test tubes and incubated. There are n_i test tubes each containing volume v_i, and y_i of these give negative reactions, indicating the absence of coliform bacteria. Altogether, data are available for m different volumes $v_1, v_2, \ldots, v_m$. It is assumed that the bacteria are distributed randomly and uniformly throughout the river water, with λ bacteria per unit volume on average.

(a) Show that the score and information functions for λ are

$$S(\lambda) = \Sigma \frac{v_i(n_i - y_i)}{1 - p_i} - \Sigma v_i n_i; \qquad \mathcal{I}(\lambda) = \Sigma \frac{v_i^2 (n_i - y_i) p_i}{(1 - p_i)^2}$$

where $p_i = \exp(-\lambda v_i)$.

(b) Using Newton's method, evaluate $\hat{\lambda}$ for the following data:

Volume	v_i	8	4	2	1
No. of test tubes	n_i	10	10	10	10
No. of negatives	y_i	0	2	3	7

4.† Use Newton's method to obtain the 10% likelihood interval for μ in Problem 9.3.2.

5. The probability that j different species of plant life are found in a randomly chosen plot of specified area is

$$p_j = \frac{(1 - e^{-\lambda})^{j+1}}{(j+1)\lambda} \qquad \text{for } j = 0, 1, 2, \ldots,$$

where $0 < \lambda < \infty$. The data obtained from an examination of 200 plots are given in the following frequency table:

No. of species	0	1	2	3	≥4
Frequency	147	36	13	4	0

(a) Obtain expressions for the log likelihood, score, and information functions of λ.
(b) Evaluate $\hat{\lambda}$ by Newton's method.
(c) Calculate estimated expected frequencies. Does the model appear to give a reasonable fit to the data?
(d) Use Newton's method to find the 10% likelihood interval for λ.

Review Problems for Chapter 9

1. (a) Red spider mites are distributed randomly and uniformly over the surface area of leaves on an apple tree. A sample of 100 leaves of unit area yielded the following results:

Number of mites	0	1	2	3	4	5	≥6
Observed frequency	16	31	22	18	10	3	0

Find the MLE and the 10% LI for λ, the expected number of mites per unit area.

(b) The following collapsed table would have been obtained if only the absence or presence of mites on a leaf had been recorded:

Number of mites	0	1 or more
Observed frequency	16	84

Find the MLE and the 10% LI for λ based on the collapsed table. Has much of the information concerning λ been lost?

2.† In a study of the spread of disease among spruce trees planted in a reforestation project, a single line of trees is selected and the number of healthy trees between successive diseased trees is counted.

Number of healthy trees	0	1	2	3	≥4	Total
Observed frequency	50	23	14	8	5	100

If the disease is non-contagious, the number X of healthy trees between successive diseased trees should have a geometric distribution, with probability function

$$f(x) = \alpha^x(1 - \alpha) \qquad \text{for } x = 0, 1, 2, \ldots$$

where $0 < \alpha < 1$.

(a) Assuming the model to be appropriate, calculate the MLE and the 10% LI for α.
(b) Calculate estimated expected frequencies under the model. Does the model give a reasonable fit to the data?

3. A shipment of 20 items contains d defectives, where d is unknown. Six items are selected at random without replacement, and only one of them is defective. Find the maximum likelihood estimate and the 50% likelihood interval for d.

4. An inoculum consists of a suspension of virulent microorganisms. To assess its strength, n animals are given a dose of 1 ml. If the dose contains one or more organism the inoculated animal will get sick, otherwise it will not.

(a) Find the probability p that an animal does not get sick as a function of λ, the density of organisms per ml of inoculum.
(b) Out of 10 animals inoculated, 6 got sick. Find the MLE and 10% likelihood interval for p.
(c) From the results in (b), obtain the MLE and 10% likelihood interval for λ.

5. An experiment was conducted to estimate γ, the 90th percentile (0.9-quantile) of the lifetime distribution of a new type of transistor. Ten transistors were tested and the observed lifetimes were

9 25 6 18 43 17 12 10 18 42

Assuming that the lifetimes follow an exponential distribution, find the maximum likelihood estimate of γ, and determine the relative likelihood of the value $\gamma = 60$.

6.†(a) The lifetimes (in hours) of certain radio tubes are independent continuous variates with cumulative distribution function

$$F(x) = 1 - e^{-x/\theta} \qquad \text{for } x > 0$$

where $\theta > 0$. Five tubes were tested simultaneously over a period of 1000 hours. One of them failed in hour 132 and another failed in hour 768. The remaining three tubes survived the test period. Obtain the log likelihood function and MLE of θ based on these results.

(b) Find the maximum likelihood estimate of ϕ, the fraction of such tubes which fail in the first 100 hours of use.

7. Suppose that events are occurring randomly in time at the constant rate of λ per minute. The numbers of events are observed in n time intervals of varying lengths, with the following results:

Length of time interval	t_1	t_2	...	t_n
Number of events	x_1	x_2	...	x_n.

Derive the likelihood function and maximum likelihood estimate of λ.

8. Let $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ be independent variates, the X_i's being $N(\mu_1, \sigma^2)$ and the Y_i's $N(\mu_2, \sigma^2)$. Both μ_1 and μ_2 are known but σ^2 is not. Find the MLE of σ^2 based on all $n + m$ measurements.

9.†One of the three children in a family comes home with the measles. Each of the other two children has probability θ of catching measles from him. If neither or both get the measles, the epidemic ends. However, if only one of them gets the disease, the remaining child has another opportunity, with probability θ, of being infected.

(a) Let X denote the total number of children in the family who are infected before the epidemic ends. Show that

$$P(X = 1) = (1 - \theta)^2; \qquad P(X = 2) = 2\theta(1 - \theta)^2; \qquad P(X = 3) = \theta^2(3 - 2\theta).$$

(b) The following data were obtained in a survey of 100 three-child families in which at least one child contracted the measles:

No. of children with measles	1	2	3
Observed frequency	48	32	20

Evaluate the MLE of θ, and calculate estimated expected frequencies under the model.

CHAPTER 10

Two-Parameter Likelihoods

In this chapter we consider likelihood methods for parameter estimation when the model involves two unknown parameters, α and β. Section 1 describes the method of maximum likelihood. The relative likelihood function and likelihood regions are considered in Section 2. Section 3 defines the maximum relative likelihood function of β, whose properties are similar to those of a one-parameter relative likelihood function. Normal approximations to the log RLF and maximum log RLF are described in Section 4.

Sections 5 and 6 deal with two applications. The estimation of the relationship between the probability of a response (e.g. death) and the dose of a drug is considered in Section 5. Section 6 describes an example from learning theory, in which the probability of a response is dependent on the results of previous trials.

Section 7 derives some results quoted in Section 1, and describes the use of Newton's method to compute points on a likelihood contour.

Most of the discussion extends readily to the case of three or more unknown parameters. However, difficulties can arise with maximum likelihood estimation and maximum relative likelihood functions when there are many unknown parameters. A brief discussion of the multi-parameter case is given in Section 8.

10.1. Maximum Likelihood Estimation

Suppose that the probability model for an experiment involves two unknown parameters, α and β. The probability of the data (observed event) E will be a function of α and β, and the joint likelihood function is proportional to this

probability:

$$L(\alpha, \beta) = c \cdot P(E; \alpha, \beta)$$

where c is positive and does not depend upon α and β. The natural logarithm of $L(\alpha, \beta)$ will be denoted by $l(\alpha, \beta)$.

The maximum likelihood estimate of (α, β) is the pair of parameter values $(\hat{\alpha}, \hat{\beta})$ which maximizes the probability of the data. Equivalently, $(\hat{\alpha}, \hat{\beta})$ is the pair of parameter values which maximizes $L(\alpha, \beta)$ and $l(\alpha, \beta)$.

In the one-parameter case we found $\hat{\theta}$ by solving the equation $S(\theta) = 0$. Now the score function is a vector with two components:

$$S(\alpha, \beta) = \begin{bmatrix} S_1(\alpha, \beta) \\ S_2(\alpha, \beta) \end{bmatrix} = \begin{bmatrix} \partial l/\partial \alpha \\ \partial l/\partial \beta \end{bmatrix}.$$

To find $(\hat{\alpha}, \hat{\beta})$, we solve a pair of simultaneous equations

$$S_1(\alpha, \beta) = 0; \qquad S_2(\alpha, \beta) = 0. \tag{10.1.1}$$

Of course, these equations need not hold if the maximum occurs on a boundary of the parameter space.

The condition for a relative maximum in the one-parameter case was $\mathscr{I}(\hat{\theta}) > 0$. Now the information function is a two-by-two symmetric matrix

$$\mathscr{I}(\alpha, \beta) = \begin{bmatrix} \mathscr{I}_{11}(\alpha, \beta) & \mathscr{I}_{12}(\alpha, \beta) \\ \mathscr{I}_{12}(\alpha, \beta) & \mathscr{I}_{22}(\alpha, \beta) \end{bmatrix} = \begin{bmatrix} -\partial^2 l/\partial \alpha^2 & -\partial^2 l/\partial \alpha \partial \beta \\ -\partial^2 l/\partial \alpha \partial \beta & -\partial^2 l/\partial \beta^2 \end{bmatrix}.$$

For a relative maximum, the matrix $\mathscr{I}(\hat{\alpha}, \hat{\beta})$ must be positive definite; that is,

$$\hat{\mathscr{I}}_{11} > 0; \qquad \hat{\mathscr{I}}_{22} > 0; \qquad \hat{\mathscr{I}}_{11}\hat{\mathscr{I}}_{22} - \hat{\mathscr{I}}_{12}^2 > 0 \tag{10.1.2}$$

where $\hat{\mathscr{I}}_{ij} = \mathscr{I}_{ij}(\hat{\alpha}, \hat{\beta})$. See Section 10.7 for a derivation of this result.

As in the one-parameter case, likelihoods are invariant under one-to-one parameter transformations. Often a parameter transformation will simplify the calculation of the maximum. The inverse transformation can then be applied to obtain the MLE's of the original parameters.

It follows from the invariance property that, if $\gamma = g(\alpha, \beta)$, then $\hat{\gamma} = g(\hat{\alpha}, \hat{\beta})$.

Calculation of $(\hat{\alpha}, \hat{\beta})$

Suppose that it is possible to solve the first equation $S_1(\alpha, \beta) = 0$ to obtain an algebraic expression for α in terms of β. Let $\hat{\alpha}(\beta)$ denote the solution of this equation. This is the MLE of α given β; that is, $\hat{\alpha}(\beta)$ is the value of α which maximizes $l(\alpha, \beta)$ when the value of β is assumed known. Substituting $\alpha = \hat{\alpha}(\beta)$ into the second equation gives

$$S_2(\hat{\alpha}(\beta), \beta) = 0.$$

This equation can then be solved for β as in the one-parameter case. We illustrate this procedure in the two examples below.

The Newton–Raphson method, which is a generalization of Newton's method, is often useful. In Newton's method, an initial guess θ_0 is improved using

$$\theta_1 = \theta_0 + S/\mathscr{I}$$

where S and $\mathscr{I}$ are evaluated at $\theta = \theta_0$. In the two-parameter case, we have

$$\begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \mathscr{I}_{11} & \mathscr{I}_{12} \\ \mathscr{I}_{12} & \mathscr{I}_{22} \end{bmatrix}^{-1} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \tag{10.1.3}$$

where the components of the score vector and information matrix are all evaluated at $\alpha = \alpha_0$, $\beta = \beta_0$. As with Newton's method, we apply (10.1.3) repeatedly until convergence is obtained, and then check that the condition (10.1.2) for a relative maximum is satisfied.

See Section 10.7 for a derivation of the Newton–Raphson method, and Section 10.5 for an example of its use.

EXAMPLE 10.1.1. Two objects with unknown weights μ_1 and μ_2 are weighed separately and together on a set of scales, giving three measurements X_1, X_2, and X_3. It is known from previous experience with the scales that measurements are independent and normally distributed about the true weights with variance 1. Thus X_1, X_2, and X_3 are independent random variables, with

$$X_1 \sim N(\mu_1, 1); \qquad X_2 \sim N(\mu_2, 1); \qquad X_3 \sim N(\mu_1 + \mu_2, 1).$$

Given observed values $x_1 = 15.6$, $x_2 = 29.3$, and $x_3 = 45.8$, what are the maximum likelihood estimates of μ_1 and μ_2?

The joint p.d.f. of X_1, X_2, and X_3 is the product of three normal p.d.f.'s:

$$f(x_1, x_2, x_3) = \left(\frac{1}{\sqrt{2\pi}}\right)^3 e^{-(x_1-\mu_1)^2/2}\, e^{-(x_2-\mu_2)^2/2}\, e^{-(x_3-\mu_1-\mu_2)^2/2}.$$

If the measurement intervals are small, $L(\mu_1, \mu_2)$ is proportional to f, and the log likelihood function is

$$l(\mu_1, \mu_2) = -\tfrac{1}{2}(x_1 - \mu_1)^2 - \tfrac{1}{2}(x_2 - \mu_2)^2 - \tfrac{1}{2}(x_3 - \mu_1 - \mu_2)^2.$$

The two components of the score function are

$$S_1(\mu_1, \mu_2) = \frac{\partial l}{\partial \mu_1} = (x_1 - \mu_1) + (x_3 - \mu_1 - \mu_2);$$

$$S_2(\mu_1, \mu_2) = \frac{\partial l}{\partial \mu_2} = (x_2 - \mu_2) + (x_3 - \mu_1 - \mu_2).$$

The second derivatives are

$$\frac{\partial^2 l}{\partial \mu_1^2} = -2; \qquad \frac{\partial^2 l}{\partial \mu_2^2} = -2; \qquad \frac{\partial^2 l}{\partial \mu_1 \partial \mu_2} = -1$$

and hence the information matrix is

$$\mathcal{I}(\mu_1, \mu_2) = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

This example is exceptional in that $\mathcal{I}(\mu_1, \mu_2)$ does not depend upon μ_1 and μ_2.

To determine $(\hat{\mu}_1, \hat{\mu}_2)$, we solve the simultaneous equations

$$S_1(\mu_1, \mu_2) = 0; \qquad S_2(\mu_1, \mu_2) = 0.$$

The first equation is

$$(x_1 - \mu_1) + (x_3 - \mu_1 - \mu_2) = 0$$

and solving this for μ_1 gives

$$\hat{\mu}_1(\mu_2) = \tfrac{1}{2}(x_1 + x_3 - \mu_2).$$

This would be the MLE of μ_1 if we knew the value of μ_2. In that case, x_1 and $x_3 - \mu_2$ are both estimates of μ_1, and the MLE is their average.

Substituting $\mu_1 = \frac{1}{2}(x_1 + x_3 - \mu_2)$ into the second equation gives

$$(x_2 - \mu_2) + x_3 - \tfrac{1}{2}(x_1 + x_3 - \mu_2) - \mu_2 = 0$$

and solving for μ_2 gives

$$\hat{\mu}_2 = \tfrac{1}{3}(2x_2 + x_3 - x_1) = 29.6.$$

Finally, we obtain

$$\hat{\mu}_1 = \tfrac{1}{2}(x_1 + x_3 - \hat{\mu}_2) = \tfrac{1}{3}(2x_1 + x_3 - x_2) = 15.9.$$

Since $\hat{\mathcal{I}}_{11} = \hat{\mathcal{I}}_{22} = 2$ and $\hat{\mathcal{I}}_{11}\hat{\mathcal{I}}_{22} - \hat{\mathcal{I}}_{12}^2 = 3$, condition (10.1.2) is satisfied, and therefore a relative maximum has been found.

When μ_2 is unknown, x_1 and $x_3 - x_2$ are both estimates of μ_1. The MLE is a weighted average of these,

$$\hat{\mu}_1 = \tfrac{2}{3}x_1 + \tfrac{1}{3}(x_3 - x_2).$$

The second estimate $x_3 - x_2$ is less precise, and therefore receives only half the weight given to x_1.

EXAMPLE 10.1.2. The following are the results, in millions of revolutions to failure, of endurance tests for 23 deep-groove ball bearings:

17.88	28.92	33.00	41.52	42.12	45.60
48.48	51.84	51.96	54.12	55.56	67.80
68.64	68.64	68.88	84.12	93.12	98.64
105.12	105.84	127.92	128.04	173.40	

The data are from page 286 of a paper by J. Lieblein and M. Zelen in *J. Res. National Bureau of Standards* (1956).

As a result of testing thousands of ball bearings, it is known that their

lifetimes have approximately a Weibull distribution. From Section 6.4, the p.d.f. of this distribution is

$$f(x) = \lambda\beta x^{\beta-1}\exp\{-\lambda x^{\beta}\} \qquad \text{for } 0 < x < \infty,$$

where λ and β are positive parameters. We wish to find $(\hat{\lambda}, \hat{\beta})$ on the basis of the above 23 measurements.

The joint p.d.f. of n independent measurements $X_1, X_2, \ldots, X_n$ from the Weibull distribution is

$$\prod_{i=1}^{n} f(x_i) = \lambda^n\beta^n\left(\prod x_i\right)^{\beta-1}\exp\{-\lambda\Sigma x_i^{\beta}\},$$

and hence the log likelihood function of λ and β is

$$l(\lambda, \beta) = n\log\lambda + n\log\beta + (\beta - 1)\Sigma\log x_i - \lambda\Sigma x_i^{\beta}.$$

The components of the score function are

$$S_1(\lambda, \beta) = \frac{\partial l}{\partial\lambda} = \frac{n}{\lambda} - \Sigma x_i^{\beta};$$

$$S_2(\lambda, \beta) = \frac{\partial l}{\partial\beta} = \frac{n}{\beta} + \Sigma\log x_i - \lambda\Sigma x_i^{\beta}\log x_i.$$

The equation $S_1(\lambda, \beta) = 0$ can be solved algebraically for λ, giving $\hat{\lambda}(\beta) = n/\Sigma x_i^{\beta}$. This is the MLE of λ when β is assumed to be known.

To obtain $\hat{\beta}$, we substitute $\lambda = n/\Sigma x_i^{\beta}$ into the equation $S_2(\lambda, \beta) = 0$ and solve for β. We have

$$g(\beta) = S_2(\hat{\lambda}(\beta), \beta)$$

$$= \frac{n}{\beta} + \Sigma\log x_i - n\Sigma x_i^{\beta}\log x_i/\Sigma x_i^{\beta}.$$

The equation $g(\beta) = 0$ may be solved numerically using Newton's method:

$$\beta_{\text{new}} = \beta_{\text{old}} - g(\beta_{\text{old}})/g'(\beta_{\text{old}}).$$

The derivative of $g(\beta)$ with respect to β is

$$g'(\beta) = -\frac{n}{\beta^2} - \frac{n\Sigma x_i^{\beta}(\log x_i)^2}{\Sigma x_i^{\beta}} + \frac{n(\Sigma x_i^{\beta}\log x_i)^2}{(\Sigma x_i^{\beta})^2}.$$

In this example we have $n = 23$ and $\Sigma\log x_i = 95.46$. Taking $\beta = 1$ as the initial guess, we obtain

$$\Sigma x_i^{\beta} = 1661; \qquad \Sigma x_i^{\beta}\log x_i = 7312; \qquad \Sigma x_i^{\beta}(\log x_i)^2 = 32572;$$

$$g(\beta) = 17.213; \qquad g'(\beta) = -28.287;$$

$$\beta_{\text{new}} = \beta - g(\beta)/g'(\beta) = 1.6085.$$

Repeating the calculations with $\beta = 1.6085$ gives $\beta_{\text{new}} = 2.0155$. Continuing in

this fashion, we find that $\hat{\beta} = 2.1021$, correct to four decimal places. We then obtain

$$\hat{\lambda} = n/\Sigma x_i^{\hat{\beta}} = 9.515 \times 10^{-5}.$$

Owing to the large amount of arithmetic, use of a computer or programmable calculator is almost essential in this example.

The parameter λ does not represent a quantity of interest, and it is usually preferable to work with parameters (θ, β) where $\lambda = \theta^{-\beta}$. By (6.4.6), the c.d.f. of the Weibull distribution is

$$F(x) = 1 - \exp\{-\lambda x^{\beta}\} = 1 - \exp\{-(x/\theta)^{\beta}\}.$$

It follows that

$$P(X \leq \theta) = F(\theta) = 1 - e^{-1} = 0.63.$$

Thus the parameter θ is directly interpretable as the 0.63-quantile of the distribution.

Since the transformation from (λ, β) to (θ, β) is one-to-one, the MLE of θ can be computed from $\hat{\lambda}$ and $\hat{\beta}$. Since $\theta = \lambda^{-1/\beta}$, the invariance property gives

$$\hat{\theta} = \hat{\lambda}^{-1/\hat{\beta}} = 81.88.$$

Problems for Section 10.1

1. Pea plants are classified according to the shape (round or angular) and color (green or yellow) of the peas they produce. According to genetic theory, the four possible plant types, RG, RY, AG, and AY have probabilities $\alpha\beta$, $\alpha(1-\beta)$, $(1-\alpha)\beta$, and $(1-\alpha)(1-\beta)$, respectively, with different plants being independent of one another. The following table shows the observed frequencies of the four types in 500 plants examined:

Plant type	RG	RY	AG	AY
Observed frequency	276	104	94	26

Find the MLE's of α and β, and calculate estimated expected frequencies under the model.

2. (a) Let $x_1, x_2, \ldots, x_n$ be independent observations from $N(\mu, \sigma^2)$, where μ and σ are both unknown. Show that

$$\hat{\mu} = \bar{x}; \qquad \hat{\sigma}^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2.$$

(b) Show that

$$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2 = \Sigma x_i^2 - (\Sigma x_i)^2/n.$$

3.† Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent normal variates with the same mean μ, but with different variances:

$$Y_i \sim N(\mu, \sigma^2/a_i) \qquad \text{for } i = 1, 2, \ldots, n$$

where $a_1, a_2, \ldots, a_n$ are known positive constants.

(a) Show that $\hat{\mu}(\sigma)$, the MLE of μ given σ, is the same for all possible values of σ.
(b) Derive expressions for $\hat{\mu}$ and $\hat{\sigma}$.
(c) Show that $\mathscr{I}(\hat{\mu}, \hat{\sigma})$, the information matrix of (μ, σ) evaluated at the maximum, is positive definite.

4. Suppose that $X_1, X_2, \ldots, X_n$ are independent normal variates with the same variance σ^2, but with different means,

$$X_i \sim N(\mu b_i, \sigma^2) \qquad \text{for } i = 1, 2, \ldots, n$$

where $b_1, b_2, \ldots, b_n$ are known constants. Find expressions for the MLE of μ and σ^2.

5.† Two treatments A and B, with success probabilities α and β, are to be tested. Subjects are treated one at a time, and the result for one subject is known before the next subject is treated. The first subject receives treatment A. Subsequently, a subject receives the same treatment as the preceding subject if a success was observed, and the other treatment if a failure was observed. Testing continues until there have been m failures with each treatment. The following data come from such an experiment with $m = 2$:

Subject	1	2	3	4	5	6	7	8	9	10
Treatment A	S	S	F			S	S	S	F	
Treatment B				S	F					F

(a) Show that, if $\alpha > \beta$, then the expected number of subjects who receive treatment A is greater than the expected number who receive treatment B.
(b) Find the log likelihood function and MLE's of α and β based on the above data with $m = 2$. Generalize your results to the case of m failures with each treatment.

6. Suppose that Y_1, Y_2, and Y_3 are independent Poisson variates with means μ_1, μ_2, and $\mu_1 + \mu_2$, respectively. Derive formulas for the maximum likelihood estimates $(\hat{\mu}_1, \hat{\mu}_2)$ based on nonzero observed values y_1, y_2, y_3.

7. The number N of eggs laid by a female robin has a Poisson distribution with mean μ. Each egg has probability θ of hatching, independently of other eggs. Given that n eggs were laid, the number Y which hatch has a binomial (n, θ) distribution.

(a) Find the joint probability function of N and Y.
(b) A biologist records n_i, the number of eggs laid, and y_i, the number which hatch, for k female robins. Find the log likelihood function and MLE's of μ and θ.

8.† The probability density function for an exponential distribution with guarantee time c is

$$f(t) = \lambda e^{-\lambda(t-c)} \qquad \text{for } t > c$$

where λ and c are positive constants. This distribution might be used as a model for the response time T in a computer system where there is a minimum response time c. Suppose that both λ and c are unknown, and that we have available n independent observations $t_1, t_2, \ldots, t_n$ from this distribution.

(a) Write down the likelihood function of λ and c, paying careful attention to the range of allowable values for λ and c.
(b) Show that, for any given λ, $L(\lambda, c)$ increases as c increases. Hence find the MLE's of c and λ.

9. Consider the situation described in Problem 9.1.9. It is suggested that, while the geometric distribution applies to most specimens, a fraction $1-\lambda$ of them have flaws and therefore always fracture on the first blow.

(a) Show that the proportions of specimens fracturing after one, two, three, and four or more blows are, respectively,

$$1-\lambda\theta,\ \lambda\theta(1-\theta),\ \lambda\theta^2(1-\theta),\ \lambda\theta^3.$$

(b) If x_i specimens are observed in the ith category ($i = 1, 2, 3, 4$; $\Sigma x_i = n$), show that

$$\hat{\theta} = \frac{x_3 + 2x_4}{x_2 + 2x_3 + 2x_4}; \qquad \hat{\lambda} = \frac{n - x_1}{n\hat{\theta}}.$$

(c) Compute estimated expected frequencies for the data given in Problem 9.1.9 and comment on the fit of the model.

10.† n individuals are randomly selected. Blood serum from each is mixed with a certain chemical compound and observed for a time T in order to record the time at which a certain color change occurs. It is observed that m individuals respond at times $t_1, t_2, \ldots, t_m$, and that the remaining $n-m$ have shown no response at the end of the observation period T. The situation is thought to be describable by a probability density function $\lambda e^{-\lambda t}$ ($t > 0$) for a fraction p of the population, and complete immunity to the reaction in the remaining fraction $1-p$. Find the maximum likelihood equations, and indicate how these can be solved for $\hat{p}$ and $\hat{\lambda}$.

11. The lengths of the gestation periods for 1000 females are summarized in the following table:

Interval (days)	Frequency	Interval (days)	Frequency
249.5–264.5	6	284.5–289.5	176
264.5–269.5	27	289.5–294.5	135
269.5–274.5	107	294.5–299.5	34
274.5–279.5	198	299.5–304.5	4
279.5–284.5	312	304.5–309.5	1

Suppose that the length of the gestation period is normally distributed with mean μ and variance σ^2.

(a) Obtain approximate values of $\hat{\mu}$ and $\hat{\sigma}^2$ by taking all times to be at the midpoints of their intervals (6 observations of 262, 27 observations of 267, etc.), and using the formulas from Problem 10.1.2(a). Compute estimated expected frequencies, and comment on the fit of the model to the data.
(b) Write down the exact likelihood function of μ and σ in terms of the $N(0, 1)$ probability integral, and indicate how $\hat{\mu}$ and $\hat{\sigma}$ could be determined exactly.

10.2. Relative Likelihood and Contour Maps

The joint relative likelihood function (RLF) of α and β is defined as follows:

$$R(\alpha, \beta) = L(\alpha, \beta)/L(\hat{\alpha}, \hat{\beta}).$$

Note that $0 \le R(\alpha, \beta) \le 1$, and $R(\hat{\alpha}, \hat{\beta}) = 1$. As in the one-parameter case, we use r to denote the natural logarithm of R:

$$r(\alpha, \beta) = \log R(\alpha, \beta) = l(\alpha, \beta) - l(\hat{\alpha}, \hat{\beta}).$$

The relative likelihood of parameter values (α_0, β_0) is

$$R(\alpha_0, \beta_0) = \frac{\text{Probability of the data when } (\alpha, \beta) = (\alpha_0, \beta_0)}{\text{Maximum probability of the data for any } \alpha, \beta}.$$

If $R(\alpha_0, \beta_0)$ is near 0, the pair (α_0, β_0) is implausible because there exist other pairs of parameter values such that the data are much more probable. The joint RLF $R(\alpha, \beta)$ ranks pairs of parameter values according to their plausibilities in light of the data.

The $100p\%$ *likelihood region* is the set of parameter values (α, β) such that $R(\alpha, \beta) \ge p$. The curve $R(\alpha, \beta) = p$ which forms the boundary of this region is called the $100p\%$ *likelihood contour*.

We may think of $R(\alpha, \beta)$ as a "mountain" of likelihood sitting on the (α, β) plane (see Figure 10.2.1). Its maximum value 1 occurs at $(\alpha, \beta) = (\hat{\alpha}, \hat{\beta})$. A convenient way to draw $R(\alpha, \beta)$ in two dimensions is by plotting contours of constant relative likelihood in the (α, β) plane. This produces a contour map similar to those used in geography and meteorology. Usually the contours will form a nested set of closed curves, roughly elliptical in shape.

EXAMPLE 10.2.1. In Example 10.1.1, the log likelihood function of the parameters μ_1 and μ_2 was found to be

$$l(\mu_1, \mu_2) = -\tfrac{1}{2}(15.6 - \mu_1)^2 - \tfrac{1}{2}(29.3 - \mu_2)^2 - \tfrac{1}{2}(45.8 - \mu_1 - \mu_2)^2.$$

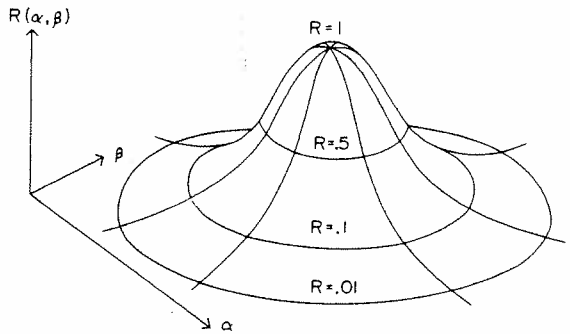

Figure 10.2.1. Two-parameter relative likelihood function.

Since $\hat{\mu}_1 = 15.9$ and $\hat{\mu}_2 = 29.6$, its maximum is

$$l(\hat{\mu}_1, \hat{\mu}_2) = -\tfrac{1}{2}[(0.3)^2 + (0.3)^2 + (0.3)^2] = -0.135.$$

Hence the log relative likelihood function is

$$r(\mu_1, \mu_2) = l(\mu_1, \mu_2) + 0.135.$$

The 50%, 10%, and 1% likelihood contours are shown in Figure 10.2.2. For instance, the 10% contour is given by $r(\mu_1, \mu_2) = \log 0.1$; that is,

$$-\tfrac{1}{2}(15.6 - \mu_1)^2 - \tfrac{1}{2}(29.3 - \mu_2)^2 - \tfrac{1}{2}(45.8 - \mu_1 - \mu_2)^2 + 0.135 = \log 0.1.$$

This is the equation of an ellipse centered at $(\hat{\mu}_1, \hat{\mu}_2)$. The 10% likelihood region is the set of all parameter values lying on or inside this ellipse.

The broken lines in Figure 10.2.2 show the outer limits of the 10% likelihood region. For all points (μ_1, μ_2) in the 10% likelihood region we have $14.15 \le \mu_1 \le 17.65$ and $27.85 \le \mu_2 \le 31.35$. These are called 10% *maximum likelihood intervals* for μ_1 and for μ_2 (see Section 10.3), and parameter values outside these intervals are implausible.

Note that, although 14.15 and 27.85 are within the 10% intervals for μ_1 and μ_2, the pair of values (14.15, 27.85) is extremely implausible. It is possible that μ_1 might be as small as 14.15, but if it is, then μ_2 is likely to be larger than 27.85. The axes of the elliptical contours are not parallel to the coordinate axes, and for this reason we cannot estimate μ_1 and μ_2 independently of one another. See Section 10.3 for further discussion.

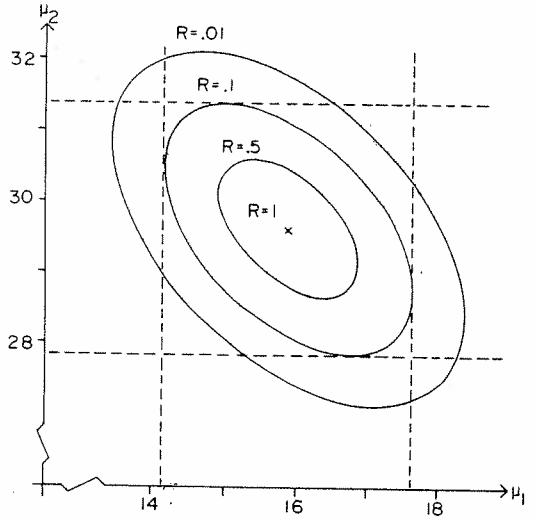

Figure 10.2.2. Contour map for $R(\mu_1, \mu_2)$ in Example 10.2.1. The broken lines show 10% maximum likelihood intervals.

EXAMPLE 10.2.2. Consider the lifetime data from a Weibull distribution in Example 10.1.2. We shall work with the parameters (θ, β), where θ is the 0.63-quantile and β is the shape parameter. We obtain $l(\theta, \beta)$ by substituting $\lambda = \theta^{-\beta}$ into the expression for $l(\lambda, \beta)$ in Example 10.1.2:

$$l(\theta, \beta) = -n\beta \log \theta + n \log \beta + (\beta - 1)\Sigma \log x_i - \theta^{-\beta}\Sigma x_i^{\beta}.$$

We showed that $\hat{\beta} = 2.1021$ and $\hat{\theta} = 81.88$, so the maximum of the log likelihood function is

$$l(\hat{\theta}, \hat{\beta}) = -113.691.$$

The log relative likelihood function of θ and β is then

$$r(\theta, \beta) = l(\theta, \beta) + 113.691.$$

Perhaps the simplest way to construct a contour map is from a tabulation of $R(\theta, \beta) = e^{r(\theta, \beta)}$ over a lattice of (θ, β) values. Table 10.2.1 gives values of $R(\theta, \beta)$ near the maximum, and the curve $R(\theta, \beta) = 0.5$ is sketched in. This is the innermost curve on the contour map of Figure 10.2.3. The 10% and 1% contours can be found in a similar way from a tabulation of $R(\theta, \beta)$ over a larger region.

The value $\beta = 1$ is of special interest, since for $\beta = 1$ the Weibull distribution simplifies to an exponential distribution. Note that the line $\beta = 1$ lies entirely outside the 1% contour in Figure 10.2.3. If $\beta = 1$, there does not exist a value of θ for which $R(\theta, \beta) \ge 0.01$; in fact, the maximum of $R(\theta, 1)$ is about 0.0004. It is therefore highly unlikely that $\beta = 1$, and the simpler exponential distribution model is not suitable for these data. Since $\beta > 1$, the ball bearings are deteriorating with age (see Section 6.4).

The broken lines in Figure 10.2.3 show the outer limits of the 10% likelihood region. The 10% maximum likelihood intervals are $64.2 \le \theta \le 103.1$, and $1.45 \le \beta \le 2.86$. Parameter values outside these intervals are implausible.

Table 10.2.1. Relative Likelihood Function $R(\beta, \theta)$

	$\theta = 72$	75	78	81	84	87	90	93
$\beta = 2.6$	0.019	0.066	0.155	0.261	0.338	0.351	0.306	0.230
2.5	0.047	0.136	0.275	0.418	0.501	0.495	0.416	0.307
2.4	0.100	0.245	0.437	0.605	0.679	0.641	0.525	0.383
2.3	0.184	0.387	0.619	0.791	0.839	0.764	0.613	0.443
2.2	0.291	0.539	0.783	0.934	0.945	0.835	0.660	0.474
2.1	0.400	0.661	0.885	0.994	0.967	0.835	0.653	0.469
2.0	0.477	0.715	0.890	0.952	0.897	0.761	0.591	0.427
1.9	0.493	0.679	0.796	0.817	0.750	0.628	0.487	0.354
1.8	0.441	0.565	0.630	0.625	0.563	0.468	0.364	0.267
1.7	0.341	0.411	0.439	0.424	0.377	0.312	0.244	0.181

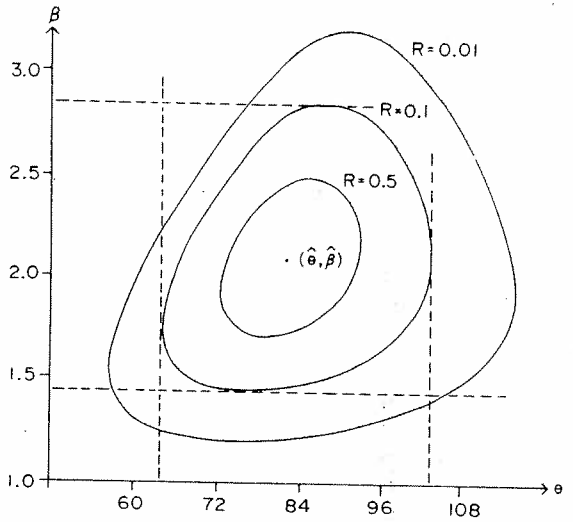

Figure 10.2.3. Contours of constant relative likelihood for the Weibull distribution parameters in Example 10.2.2.

Note on Calculation

In the preceding example we constructed the contour map by tabulating $R(\theta, \beta)$ over a lattice of (θ, β)-values. A large table of values, and therefore a great deal of arithmetic, may be needed to produce an accurate contour map. An alternative procedure is to use Newton's method to solve for points on a particular contour. See Section 10.7 for details.

Before computing contours, the possibility of transforming parameters should be considered. The calculation and interpretation of the contour map will be easier if the axes of the contours are roughly parallel to the parameter axes. This is an additional reason for working with (θ, β) rather than (λ, β) in the Weibull distribution example.

Problems for Section 10.2

1. Find the log relative likelihood function of α and β in Problem 10.1.1. Plot the 10% likelihood contour, and obtain 10% maximum likelihood intervals for α and β.

2.† A zoologist wishes to investigate the survival of fish in an isolated pond during the winter. The population may change by death during the period, but not by birth, immigration, or emigration. He catches n fish, marks them, and returns them. On two subsequent occasions he takes a sample of fish from the pond, observes which of his marked fish are in the sample, and returns them. He finds that x_1 of the marked fish are caught in the first sample only, x_2 in the second sample only, x_3 in both samples, and x_4 in neither sample. He assumes that each individual independently has a probability ϕ of survival between sampling periods, and a probability p of being caught in any sample if it is alive at the time of the sample.

   (a) Show that the probabilities of the four classes of recapture are $\alpha(1-\alpha)$, $\alpha\beta$, α^2, and $1-\alpha-\alpha\beta$, respectively, where $\alpha = \phi p$ and $\beta = \phi(1-p)$.
   (b) Show that
   $$\hat{\beta} = \frac{x_2}{x_2 + x_4}\left(\frac{1-\hat{\alpha}}{\hat{\alpha}}\right); \qquad \hat{\alpha} = \frac{x_1 + 2x_3}{n + x_1 + x_3}.$$
   (c) Suppose that the observed frequencies are 15, 11, 9, and 29, respectively. Find the MLE's of ϕ and p, and compute estimated expected frequencies.
   (d) Find 10% maximum likelihood intervals for ϕ and p based on the data in (c).

3. Suppose that, in Example 10.1.2, testing had stopped at 75 million revolutions. The last 8 lifetimes would then have been censored. Thus we would have $m = 15$ failure times $x_1 = 17.88, \ldots, x_{15} = 68.88$ and 8 equal censoring times $T_1 = \cdots = T_8 = 75$. Find the MLE's of θ and β and prepare a contour map similar to Figure 10.2.3. What effect does the censoring have on the estimation of θ and β?

4. Eighteen identical ball bearings were placed in test machines and subjected to a fixed radial load. The following are the numbers of hours the individual bearings endured at 2000 r.p.m.

183	355	538	618	697	834	862	887	1056
1147	1351	1506	1578	1607	1683	1710	2020	2410

   The lifetimes are modeled as independent observations from a Weibull distribution with c.d.f.
   $$F(x) = 1 - \exp\{-(x/\theta)^\beta\} \qquad \text{for } x > 0$$
   where θ and β are positive.

   (a) Find $\hat{\theta}$ and $\hat{\beta}$.
   (b) Plot the 10% likelihood contour, and obtain the 10% maximum likelihood interval for each parameter.
   (c) Would it be reasonable to assume an exponential distribution model for these data?

10.3. Maximum Relative Likelihood

It may be that, although the probability model involves two unknown parameters α and β, only one of them, say β, is of real interest. The joint RLF ranks *pairs* of values (α, β) according to their plausibilities in the light of the data. However, what we would like is a summary of the information concerning parameter β only.

The *maximum relative likelihood function* of β is obtained by maximizing $R(\alpha, \beta)$ over α with β fixed:

$$R_{\max}(\beta) = \max_{\alpha} R(\alpha, \beta) = R(\hat{\alpha}(\beta), \beta).$$

Here $\hat{\alpha}(\beta)$ is the MLE of α given β, which may be found by solving the equation $S_1(\alpha, \beta) = 0$ (see Section 10.1). The natural logarithm of $R_{\max}$ is

$$r_{\max}(\beta) = r(\hat{\alpha}(\beta), \beta) = l(\hat{\alpha}(\beta), \beta) - l(\hat{\alpha}, \hat{\beta}), \tag{10.3.1}$$

which is the difference between the restricted maximum of $l(\alpha, \beta)$ with β fixed and the unrestricted maximum.

The joint RLF can be pictured as a mountain of likelihood sitting in the (α, β) plane (see Figure 10.2.1). The maximum RLF of β is the profile or silhouette of $R(\alpha, \beta)$ when it is viewed from a distant point on the α-axis. Similarly, $R_{\max}(\alpha)$ is the silhouette of the likelihood mountain when it is viewed from a distant point on the β-axis.

The properties of $R_{\max}(\beta)$ are similar to those of a one-parameter RLF. For instance, we have

$$0 \le R_{\max}(\beta) \le 1; \qquad R_{\max}(\hat{\beta}) = 1.$$

If $R_{\max}(\beta_0)$ is near 0, there does not exist a parameter value α_0 such that the pair (α_0, β_0) is plausible, and hence β_0 is an implausible value of β. On the other hand, if $R_{\max}(\beta_0)$ is near 1, then there exists at least one plausible pair of values (α_0, β_0), and thus β_0 is not an implausible value of β.

The $100p\%$ *maximum likelihood interval* (or region) for β is the set of all β values for which $R_{\max}(\beta) \ge p$. This interval contains those β values such that, for some α, the pair (α, β) belongs to the $100p\%$ likelihood region. Ten percent maximum likelihood intervals are shown with broken lines in Figures 10.2.2 and 10.2.3.

EXAMPLE 10.3.1. Consider the situation described in Examples 10.1.1 and 10.2.1. The joint log likelihood function is

$$l(\mu_1, \mu_2) = -\tfrac{1}{2}(x_1 - \mu_1)^2 - \tfrac{1}{2}(x_2 - \mu_2)^2 - \tfrac{1}{2}(x_3 - \mu_1 - \mu_2)^2$$

and the MLE's are

$$\hat{\mu}_1 = \tfrac{1}{3}(2x_1 + x_3 - x_2) = 15.9; \qquad \hat{\mu}_2 = \tfrac{1}{3}(2x_2 + x_3 - x_1) = 29.6.$$

From Example 10.1.1, the MLE of μ_1 given μ_2 is

$$\hat{\mu}_1(\mu_2) = \tfrac{1}{2}(x_1 + x_3 - \mu_2).$$

The maximum log RLF of μ_2 is

$$r_{\max}(\mu_2) = l(\hat{\mu}_1(\mu_2), \mu_2) - l(\hat{\mu}_1, \hat{\mu}_2).$$

After substitution and simplification, we obtain

$$r_{\max}(\mu_2) = -\tfrac{3}{4}(\mu_2 - \hat{\mu}_2)^2.$$

Taking $r_{\max}(\mu_2) \ge \log 0.1$ gives the 10% maximum likelihood interval $27.85 \le \mu_2 \le 31.35$. Similarly, we find that

$$r_{\max}(\mu_1) = -\tfrac{3}{4}(\mu_1 - \hat{\mu}_1)^2,$$

and the 10% interval is $14.15 \le \mu_1 \le 17.65$. These intervals are shown in Figure 10.2.2.

EXAMPLE 10.3.2. Consider the analysis of failure times from a Weibull distribution, as previously discussed in Examples 10.1.2 and 10.2.2. The log RLF of λ and β is

$$r(\lambda, \beta) = n \log \lambda + n \log \beta + (\beta - 1)\Sigma \log x_i - \lambda \Sigma x_i^\beta + 113.691.$$

From Example 10.1.2, the MLE of λ given β is $\hat{\lambda}(\beta) = n/\Sigma x_i^\beta$. Hence the maximum log RLF of β is

$$\begin{aligned} r_{\max}(\beta) &= r(\hat{\lambda}(\beta), \beta) \\ &= n \log(n/\Sigma x_i^\beta) + n \log \beta + (\beta - 1)\Sigma \log x_i - n + 113.691. \end{aligned}$$

This function is plotted with a solid line in Figure 10.3.1. The broken line shows the normal approximation to $r_{\max}(\beta)$ (see Section 10.4).

The 10% maximum likelihood interval for β is $1.45 \le \beta \le 2.86$. This can be obtained from Figure 10.3.1, or from Figure 10.2.3, or by the numerical methods described in Section 9.8.

Next we find the maximum log RLF of θ, the 0.63-quantile of the distribution. The joint log RLF of θ and β is

$$r(\theta, \beta) = -n\beta \log \theta + n \log \beta + (\beta - 1)\Sigma x_i - \theta^{-\beta}\Sigma x_i^\beta + 113.691.$$

We find $\hat{\beta}(\theta)$, the MLE of β given θ, by solving the equation $S_2(\theta, \beta) = 0$. Then we obtain

$$r_{\max}(\theta) = r(\theta, \hat{\beta}(\theta)).$$

Numerical methods are required to solve for $\hat{\beta}(\theta)$. For instance, when $\theta = 80$ we find by Newton's method that $S_2(80, \beta) = 0$ for $\beta = 2.0764$. Thus $\hat{\beta}(80) = 2.0764$, and

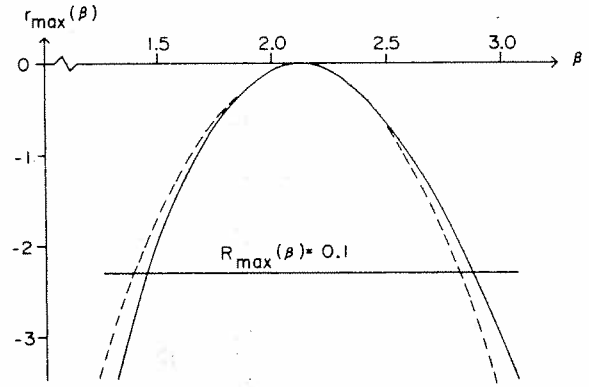

Figure 10.3.1. Maximum log RLF for β in the Weibull distribution example. The normal approximation is shown with a broken line.

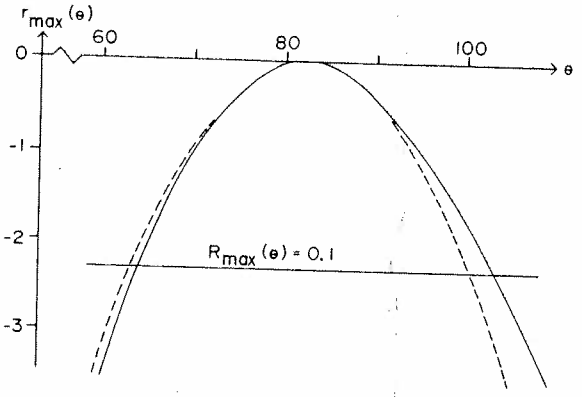

Figure 10.3.2. Maximum log RLF for θ in the Weibull distribution example. The normal approximation is shown with a broken line.

$$r_{\max}(80) = r(80, 2.0764) = -0.024.$$

Similarly, we find that $S_2(70, \beta) = 0$ for $\beta = 1.8810$, and

$$r_{\max}(70) = r(70, 1.8810) = -1.023.$$

Figure 10.3.2 shows $r_{\max}(\theta)$ and the normal approximation from Section 10.4. The 10% maximum likelihood interval is $64.2 \le \theta \le 103.1$.

Factorization

Suppose that the joint likelihood function of α and β factors into a function of α times a function of β:

$$L(\alpha, \beta) = g(\alpha) \cdot h(\beta) \qquad \text{for all } \alpha, \beta. \tag{10.3.2}$$

Then $\hat{\alpha}(\beta)$ is not a function of β, and

$$R_{\max}(\beta) = L(\hat{\alpha}(\beta), \beta)/L(\hat{\alpha}, \hat{\beta}) = h(\beta)/h(\hat{\beta}).$$

It follows that

$$R(\alpha, \beta) = R_{\max}(\alpha) \cdot R_{\max}(\beta);$$

$$r(\alpha, \beta) = r_{\max}(\alpha) + r_{\max}(\beta).$$

When (10.3.2) holds, α and β can be estimated independently. Graphs of $r_{\max}(\alpha)$ and $r_{\max}(\beta)$ will then provide a complete summary of the information concerning (α, β), and it is not necessary to consider a contour map.

If (10.3.2) does not hold, the range of plausible values for one parameter will depend upon the value of the other parameter. This information cannot be recovered from just the maximum RLF's and a contour map will be required.

By careful design of the experiment and choice of the parameters, it may be possible to arrange that (10.3.2) is true, at least approximately. We can then treat the two-parameter problem as a pair of one-parameter problems, thus simplifying both the analysis and the interpretation. Advance planning to achieve factorization of the likelihood function becomes progressively more important as the number of unknown parameters increases.

EXAMPLE 10.3.1 (continued). The joint log likelihood function of μ_1 and μ_2 contains a product term $\mu_1\mu_2$, and hence the likelihood function does not factor into a function of μ_1 times a function of μ_2. As Figure 10.2.2 shows, the range of plausible values for μ_1 depends upon the value of μ_2. In particular, the most likely value of μ_1 is

$$\hat{\mu}_1(\mu_2) = \tfrac{1}{2}(x_1 + x_3 - \mu_2)$$

which decreases as μ_2 increases.

Suppose that we work instead with parameters (θ_1, θ_2) where $\theta_1 = \mu_1 + \mu_2$ and $\theta_2 = \mu_1 - \mu_2$. Their MLE's are

$$\hat{\theta}_1 = \hat{\mu}_1 + \hat{\mu}_2 = \tfrac{1}{3}(x_1 + x_2 + 2x_3);$$

$$\hat{\theta}_2 = \hat{\mu}_1 - \hat{\mu}_2 = x_1 - x_2.$$

Upon substituting $\mu_1 = (\theta_1 + \theta_2)/2$, $\mu_2 = (\theta_1 - \theta_2)/2$ and simplifying we find that the log RLF of θ_1 and θ_2 is

$$r(\theta_1, \theta_2) = -\tfrac{3}{4}(\theta_1 - \hat{\theta}_1)^2 - \tfrac{1}{4}(\theta_2 - \hat{\theta}_2)^2$$

$$= r_{\max}(\theta_1) + r_{\max}(\theta_2).$$

The log relative likelihood of any pair of values (μ_1, μ_2) can be found by computing the corresponding values of θ_1 and θ_2 and then summing $r_{\max}(\theta_1)$ and $r_{\max}(\theta_2)$.

PROBLEMS FOR SECTION 10.3

1. Derive the maximum RLF's of α and β in Problem 10.1.1, and find the 10% maximum likelihood interval for α.

2.†(a) Let X and Y be independent Poisson variates with means μ and $\lambda\mu$, respectively. Derive the maximum RLF of λ.

(b) In a certain city there were 47 murders during the year prior to abolition of the death penalty. There were 57 murders the year after abolition. Assuming these to be observed values of independent Poisson variates with means μ and $\lambda\mu$, find the 10% maximum likelihood interval for λ. Is it plausible that the murder rate has not changed?

3. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be independent exponential variates. The X_i's have mean θ and the Y_i's have mean $\lambda\theta$ where λ and θ are positive unknown parameters.

(a) Derive expressions for $\hat{\lambda}$ and $\hat{\theta}$.
(b) Show that the maximum RLF of λ is

$$R_{\max}(\lambda) = 2^{2n}\left(\frac{\bar{y}}{\lambda\bar{x}}\right)^n\left(1+\frac{\bar{y}}{\lambda\bar{x}}\right)^{-2n}.$$

(c) The following are the observed survival times for 12 subjects:

Treatment A:	9	186	25	6	44	115
Treatment B:	1	18	6	25	14	45

Survival times are modeled as independent exponential variates with mean θ for treatment A and mean $\lambda\theta$ for treatment B. Obtain the 10% maximum likelihood interval for λ. Do these data clearly demonstrate the superiority of the first treatment?

4. Let $x_1, x_2, \ldots, x_n$ be independent observations from $N(\mu, \sigma^2)$ where both μ and σ are unknown.

(a) Show that the maximum relative likelihood function of μ is

$$R_{\max}(\mu) = \left[\frac{\Sigma(x_i-\mu)^2}{n\hat{\sigma}^2}\right]^{-n/2} = \left[1+\left(\frac{\hat{\mu}-\mu}{\hat{\sigma}}\right)^2\right]^{-n/2}$$

for $-\infty < \mu < \infty$. Hence show that the $100p\%$ maximum likelihood interval for μ has the form

$$\mu \in \hat{\mu} \pm c\hat{\sigma}$$

where c is a function of p and n.
(b) Show that the maximum relative likelihood function of σ is

$$R_{\max}(\sigma) = \left(\frac{\hat{\sigma}}{\sigma}\right)^n \exp\left\{\frac{n}{2}\left[1-\left(\frac{\hat{\sigma}}{\sigma}\right)^2\right]\right\} \qquad \text{for } \sigma > 0.$$

5. Find the maximum RLF of μ_2 in Problem 10.1.6.

6.† Find the maximum RLF's of λ and c in Problem 10.1.8.

7. Find the maximum relative likelihood function for θ in Problem 10.1.9. Show that this is the same as the relative likelihood function for θ based on the conditional distribution of X_2, X_3, and X_4 given X_1.

8.† Find the maximum relative likelihood function for λ in Problem 10.1.10.

9. Show that a one-to-one parameter transformation from (α, β) to (γ, β) does not affect the maximum RLF of β. The maximum RLF of β can be found by maximizing the joint RLF of α and β over α, or by maximizing the joint RLF of γ and β over γ.

10.4. Normal Approximations

In Section 9.7 we derived the normal approximation

$$r(\theta) \approx -\tfrac{1}{2}(\theta-\hat{\theta})^2 \mathscr{I}(\hat{\theta})$$

by ignoring cubic and higher terms in the Taylor's series expansion of $l(\theta)$ about $\theta = \hat{\theta}$. A similar derivation in the two-parameter case gives

$$r(\alpha, \beta) \approx -\tfrac{1}{2}(\alpha-\hat{\alpha})^2\hat{\mathscr{I}}_{11} - \tfrac{1}{2}(\beta-\hat{\beta})^2\hat{\mathscr{I}}_{22} - (\alpha-\hat{\alpha})(\beta-\hat{\beta})\hat{\mathscr{I}}_{12}, \tag{10.4.1}$$

where $\hat{\mathscr{I}}_{ij} = \mathscr{I}_{ij}(\hat{\alpha}, \hat{\beta})$ as in Section 10.1. If we take

$$\theta = \begin{bmatrix}\alpha\\ \beta\end{bmatrix}; \qquad \hat{\theta} = \begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix}; \qquad \mathscr{I}(\hat{\theta}) = \mathscr{I}(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix}\hat{\mathscr{I}}_{11} & \hat{\mathscr{I}}_{12}\\ \hat{\mathscr{I}}_{12} & \hat{\mathscr{I}}_{22}\end{bmatrix},$$

the approximation may be written

$$r(\theta) \approx -\tfrac{1}{2}(\theta-\hat{\theta})'\mathscr{I}(\hat{\theta})(\theta-\hat{\theta}),$$

which shows its similarity to the one-parameter result.

When (10.4.1) applies, the likelihood contours are close to ellipses centered at $(\hat{\alpha}, \hat{\beta})$. As in the one-parameter case, the normal approximation is not invariant, and a one-to-one transformation from (α, β) to new parameters (μ, ν) may substantially improve its accuracy. The information matrix for the new parameters is

$$\mathscr{I}_*(\hat{\mu}, \hat{\nu}) = \hat{Q}'\mathscr{I}(\hat{\alpha}, \hat{\beta})\hat{Q}. \tag{10.4.2}$$

Here Q is the two-by-two matrix of derivatives of the old parameters with respect to the new:

$$Q = \begin{bmatrix}\partial\alpha/\partial\mu & \partial\alpha/\partial\nu\\ \partial\beta/\partial\mu & \partial\beta/\partial\nu\end{bmatrix}.$$

We evaluate Q at the MLE to obtain $\hat{Q}$. The proof of (10.4.2) is similar to the proof of the one-parameter result (9.7.4).

Differentiating (10.4.1) with respect to α gives an approximation to $S_1(\alpha, \beta)$,

$$S_1(\alpha, \beta) \approx -(\alpha-\hat{\alpha})\hat{\mathscr{I}}_{11} - (\beta-\hat{\beta})\hat{\mathscr{I}}_{12},$$

which is linear in α and β. Setting this equal to zero and solving for α gives

$$\hat{\alpha}(\beta) \approx \hat{\alpha} + (\hat{\beta}-\beta)\hat{\mathscr{I}}_{12}/\hat{\mathscr{I}}_{11}. \tag{10.4.3}$$

If we now substitute for α in (10.4.1) and simplify, we obtain the following normal approximation to $r_{\max}(\beta)$:

$$r_{\max}(\beta) \approx -\tfrac{1}{2}(\beta-\hat{\beta})^2[\hat{\mathscr{I}}_{22} - \hat{\mathscr{I}}_{12}^2/\hat{\mathscr{I}}_{11}]. \tag{10.4.4}$$

This has the same form as the normal approximation which we derived in Section 9.7 for the one-parameter case. The quantity in square brackets is positive by (10.1.2).

The inverse of the information matrix is

$$\begin{bmatrix}\hat{\mathscr{I}}_{11} & \hat{\mathscr{I}}_{12}\\ \hat{\mathscr{I}}_{12} & \hat{\mathscr{I}}_{22}\end{bmatrix}^{-1} = \frac{1}{\hat{\mathscr{I}}_{11}\hat{\mathscr{I}}_{22} - \hat{\mathscr{I}}_{12}^2}\begin{bmatrix}\hat{\mathscr{I}}_{22} & -\hat{\mathscr{I}}_{12}\\ -\hat{\mathscr{I}}_{12} & \hat{\mathscr{I}}_{11}\end{bmatrix}$$

and the (2, 2)-element of the inverse is

$$\hat{\mathscr{I}}^{22} = \hat{\mathscr{I}}_{11}/(\hat{\mathscr{I}}_{11}\hat{\mathscr{I}}_{22} - \hat{\mathscr{I}}_{12}^2) = (\hat{\mathscr{I}}_{22} - \hat{\mathscr{I}}_{12}^2/\hat{\mathscr{I}}_{11})^{-1}.$$

Thus the normal approximation (10.4.4) can also be written as

$$r_{\max}(\beta) \approx -\tfrac{1}{2}(\beta - \hat{\beta})^2/\hat{\mathscr{I}}^{22}. \tag{10.4.5}$$

EXAMPLE 10.4.1. Consider the normal distribution example of the preceding three sections. The log likelihood function of μ_1 and μ_2 is

$$l(\mu_1, \mu_2) = -\tfrac{1}{2}(x_1 - \mu_1)^2 - \tfrac{1}{2}(x_2 - \mu_2)^2 - \tfrac{1}{2}(x_3 - \mu_1 - \mu_2)^2,$$

which is a second-degree polynomial in μ_1 and μ_2. As a result, the approximations (10.4.1), (10.4.3), and (10.4.4) hold exactly. Since $\hat{\mathscr{I}}_{11} = \hat{\mathscr{I}}_{22} = 2$ and $\hat{\mathscr{I}}_{12} = 1$, we have

$$r(\mu_1, \mu_2) = -(\mu_1 - \hat{\mu}_1)^2 - (\mu_2 - \hat{\mu}_2)^2 - (\mu_1 - \hat{\mu}_1)(\mu_2 - \hat{\mu}_2).$$

The contours of constant relative likelihood are ellipses as shown in Figure 10.2.2. Also, since

$$\hat{\mathscr{I}}_{11} - \hat{\mathscr{I}}_{12}^2/\hat{\mathscr{I}}_{22} = \hat{\mathscr{I}}_{22} - \hat{\mathscr{I}}_{12}^2/\hat{\mathscr{I}}_{11} = \tfrac{3}{2},$$

the maximum log RLF's are

$$r_{\max}(\mu_1) = -\tfrac{3}{4}(\mu_1 - \hat{\mu}_1)^2; \qquad r_{\max}(\mu_2) = -\tfrac{3}{4}(\mu_2 - \hat{\mu}_2)^2.$$

In Example 10.3.1 we transformed parameters from (μ_1, μ_2) to (θ_1, θ_2) where $\mu_1 = (\theta_1 + \theta_2)/2$ and $\mu_2 = (\theta_1 - \theta_2)/2$. Differentiating the μ_i's with respect to the θ_i's gives

$$Q = \begin{bmatrix} \partial\mu_1/\partial\theta_1 & \partial\mu_1/\partial\theta_2 \\ \partial\mu_2/\partial\theta_1 & \partial\mu_2/\partial\theta_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}.$$

The derivatives are not functions of the parameters because the transformation is linear. Thus $\hat{Q} = Q$, and (10.4.5) gives

$$\mathscr{I}_*(\hat{\theta}_1, \hat{\theta}_2) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}^t \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

The log likelihood function is a second-degree polynomial in θ_1 and θ_2, and the approximations hold exactly. From (10.4.1) and (10.4.4) we obtain

$$r(\theta_1, \theta_2) = -\tfrac{3}{4}(\theta_1 - \hat{\theta}_1)^2 - \tfrac{1}{4}(\theta_2 - \hat{\theta}_2)^2;$$

$$r_{\max}(\theta_1) = -\tfrac{3}{4}(\theta_1 - \hat{\theta}_1)^2; \qquad r_{\max}(\theta_2) = -\tfrac{1}{4}(\theta_2 - \hat{\theta}_2)^2.$$

EXAMPLE 10.4.2. Figure 10.2.3 shows contours of constant relative likelihood for parameters (θ, β) in the Weibull distribution example. The 10% and 1% contours are not elliptical in shape, and a sample size larger than $n = 23$ is needed before the normal approximation (10.4.1) will give accurate results.

The maximum log RLF's of β and θ are shown as solid curves in Figures 10.3.1 and 10.3.2. To obtain the normal approximations to these curves, we

must first evaluate $\mathscr{I}(\hat{\theta}, \hat{\beta})$. An expression for $l(\theta, \beta)$ is given in Example 10.2.2. We find the second derivatives of $l(\theta, \beta)$ and change their signs to get $\mathscr{I}(\theta, \beta)$. Substituting $\theta = \hat{\theta}$ and $\beta = \hat{\beta}$ then gives

$$\mathscr{I}(\hat{\theta}, \hat{\beta}) = \begin{bmatrix} 0.01516 & -0.13046 \\ -0.13046 & 10.379 \end{bmatrix}.$$

From this we compute

$$\hat{\mathscr{I}}_{11} - \hat{\mathscr{I}}_{12}^2/\hat{\mathscr{I}}_{22} = 0.01352; \qquad \hat{\mathscr{I}}_{22} - \hat{\mathscr{I}}_{12}^2/\hat{\mathscr{I}}_{11} = 9.256$$

and then (10.4.4) gives

$$r_{\max}(\theta) \approx -\tfrac{1}{2}(\theta - 81.88)^2(0.01352);$$

$$r_{\max}(\beta) \approx -\tfrac{1}{2}(\beta - 2.1021)^2(9.256).$$

These functions are plotted as broken curves in Figures 10.3.1 and 10.3.2. The agreement is not too bad. The normal approximations give 10% maximum likelihood intervals $63.4 \le \theta \le 100.3$ and $1.40 \le \beta \le 2.81$, while the exact results are $64.2 \le \theta \le 103.1$ and $1.45 \le \beta \le 2.86$.

PROBLEMS FOR SECTION 10.4

1. In Problem 10.2.4, evaluate the information matrix $\mathscr{I}(\hat{\theta}, \hat{\beta})$. Find approximate 10% maximum likelihood intervals for θ and β, and compare them with the exact results.

2.† (a) Evaluate $\mathscr{I}(\hat{\mu}, \hat{\lambda})$ and find an approximate 10% maximum likelihood interval for λ in Problem 10.3.2(b).
   (b) Transform parameters from (μ, λ) to (α, β) where $\alpha = \log\mu$ and $\beta = \log\lambda$. Calculate the information matrix $\mathscr{I}_*(\hat{\alpha}, \hat{\beta})$. Obtain an approximate 10% maximum LI for β and transform it to give an interval for λ.
   (c) Compare the results of (a) and (b) with the exact 10% interval. Does the logarithmic transformation seem to improve the normal approximation?

3. Prove result (10.4.2) for transforming information matrices.

4. Consider a one-to-one parameter transformation from (α, β) with information matrix $\hat{\mathscr{I}} = \mathscr{I}(\hat{\alpha}, \hat{\beta})$ to (γ, β) with information matrix $\hat{\mathscr{I}}_* = \mathscr{I}_*(\hat{\gamma}, \hat{\beta})$.
   (a) Show that

$$\hat{\mathscr{I}}_*^{-1} = \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \hat{\mathscr{I}}^{-1} \begin{bmatrix} a & 0 \\ b & 1 \end{bmatrix}$$

   where $a = \dfrac{\partial\gamma}{\partial\alpha}$ and $b = \dfrac{\partial\gamma}{\partial\beta}$ are evaluated at the maximum.
   (b) Show that $\hat{\mathscr{I}}_*^{22} = \hat{\mathscr{I}}^{22}$. Hence the normal approximation (10.4.5) does not depend upon whether we work with parameters (α, β) or with parameters (γ, β).

(c) Show that

$$\hat{\mathscr{I}}^{11} = a^2\hat{\mathscr{I}}^{11} + 2ab\hat{\mathscr{I}}^{12} + b^2\hat{\mathscr{I}}^{22}$$
$$= \frac{a^2\hat{\mathscr{I}}_{22} - 2ab\hat{\mathscr{I}}_{12} + b^2\hat{\mathscr{I}}_{11}}{\hat{\mathscr{I}}_{11}\hat{\mathscr{I}}_{22} - \hat{\mathscr{I}}_{12}^2}.$$

Hence it is possible to approximate $r_{\max}(\gamma)$, where $\gamma = g(\alpha, \beta)$, by using just the results computed for α and β.

5.* An alternative method for deriving a normal approximation to $r_{\max}(\beta)$ is to expand $r_{\max}(\beta)$ about $\beta = \hat{\beta}$ and then ignore cubic and higher terms. Show that this procedure also leads to the approximation (10.4.2).

10.5. A Dose-Response Example

Suppose that a drug is administered in k different doses $d_1, d_2, \ldots, d_k$. Dosage is usually taken to be the log concentration of the active ingredient, so that $d \to -\infty$ as the concentration approaches zero. Suppose that each subject either responds to the drug or does not respond, so that the response is quantal (all or nothing). For instance, when an insecticide is applied, insects either respond (die) or do not respond (survive). When a beneficial drug is administered, an improvement in the patient's condition might be taken as a response, and a lack of improvement as no response.

Let $p(d)$ denote the probability of response for a subject who receives dose d of the drug. We expect $p(d)$ to be a smooth nondecreasing function of d. To simplify matters, we assume that $p(d) \to 0$ as $d \to -\infty$, and $p(d) \to 1$ as $d \to +\infty$; that is, we assume that no subjects will respond if the dose is very small, and all subjects will respond to a very large dose. These assumptions are not always reasonable. There may be some subjects who would respond naturally without the drug, and others may be immune to the drug. For discussion of these situations, see D.J. Finney, *Probit Analysis*, 3rd edition (1971), published by the Cambridge University Press.

When these assumptions hold, the dose-response curve will be as shown in Figure 10.5.1, and $p(d)$ has the same mathematical properties as the c.d.f. of a continuous distribution.

This result can also be obtained by imagining that different subjects have different tolerances to the drug. Let D represent the minimum dose required to produce a response in a randomly chosen subject. A dose d will produce a response if and only if the tolerance of the individual is at most d. Thus the probability of a response at dose d is

$$p(d) = P(D \le d) = F(d)$$

where F is the cumulative distribution function of the random variable D.

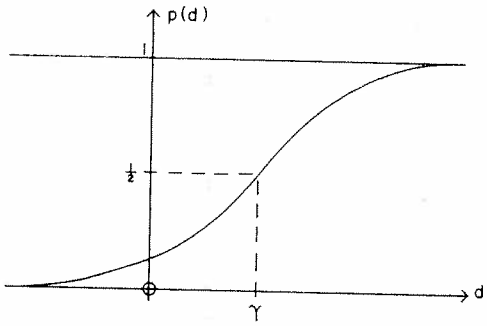

Figure 10.5.1. A typical dose-response curve.

Probit Model

Suppose that the tolerance D is normally distributed with mean μ and variance σ^2. Then

$$p(d) = P(D \le d) = P\left(Z \le \frac{d-\mu}{\sigma}\right) = F(\alpha + \beta d)$$

where $\alpha = -\mu/\sigma$, $\beta = 1/\sigma$, and F is the standardized normal c.d.f. (6.6.3). This can also be written

$$F^{-1}(p) = \alpha + \beta d$$

where F^{-1} is the inverse of the $N(0, 1)$ c.d.f., and is called a probit dose-response model.

Logistic Model

The logistic distribution is similar in shape to $N(0, 1)$ and has c.d.f.

$$G(z) = 1 - \frac{1}{1+e^z} \qquad \text{for } -\infty < z < \infty.$$

An advantage of the logistic distribution is that its c.d.f. can be evaluated without numerical integration. Replacing F by G in the above derivation gives

$$p(d) = G(\alpha + \beta d) = 1 - \frac{1}{1+e^{\alpha+\beta d}}. \tag{10.5.1}$$

Solving $G(z) = p$ gives $z = \log\frac{p}{1-p}$, and hence the model may be rewritten

$$\log\frac{p}{1-p} = \alpha + \beta d. \tag{10.5.2}$$

This is called the logistic dose-response model, and $\log \frac{p}{1-p}$ is called the *log-odds* or *logistic transform* of p.

Both the logistic and the probit models are commonly used in analyzing data from dose-response tests. The two models lead to quite similar results, and a very large amount of data would be needed to show that one was better than the other. The calculations are a bit simpler for the logistic model, and for this reason we shall use it in what follows.

Maximum Likelihood Estimates

Suppose that n_i subjects receive dose d_i, and that Y_i of these respond ($i = 1, 2, \ldots, k$). Then Y_i has a binomial distribution with parameters n_i and p_i, where

$$p_i = 1 - (1 + e^{\alpha + \beta d_i})^{-1}.$$

If different subjects are used for different doses, the Y_i's will be independent, and their joint probability function is

$$f(y_1, y_2, \ldots, y_k) = \prod_{i=1}^{k} \binom{n_i}{y_i} p_i^{y_i}(1 - p_i)^{n_i - y_i}.$$

The likelihood and log likelihood functions are

$$L(\alpha, \beta) = \prod_{i=1}^{k} p_i^{y_i}(1 - p_i)^{n_i - y_i} = \prod_{i=1}^{k} \left[\frac{p_i}{1 - p_i}\right]^{y_i} (1 - p_i)^{n_i};$$

$$l(\alpha, \beta) = \sum_{i=1}^{k} \left[y_i \log \frac{p_i}{1 - p_i} + n_i \log(1 - p_i) \right]$$

$$= \sum_{i=1}^{k} [y_i(\alpha + \beta d_i) + n_i \log(1 - p_i)].$$

Note that

$$\frac{\partial p_i}{\partial \alpha} = (1 + e^{\alpha + \beta d_i})^{-2} e^{\alpha + \beta d_i} = p_i(1 - p_i);$$

$$\frac{\partial p_i}{\partial \beta} = (1 + e^{\alpha + \beta d_i})^{-2} e^{\alpha + \beta d_i} d_i = p_i(1 - p_i) d_i.$$

Using these results, one can easily show that

$$S_1(\alpha, \beta) = \frac{\partial l}{\partial \alpha} = \Sigma(y_i - \mu_i);$$

$$S_2(\alpha, \beta) = \frac{\partial l}{\partial \beta} = \Sigma(y_i - \mu_i) d_i$$

where $\mu_i = n_i p_i$. Differentiating again gives

$$\mathcal{I}_{11}(\alpha, \beta) = -\frac{\partial}{\partial \alpha} S_1(\alpha, \beta) = \Sigma n_i \frac{\partial p_i}{\partial \alpha} = \Sigma v_i$$

where $v_i = n_i p_i(1 - p_i)$. Similarly, we obtain

$$\mathcal{I}_{12}(\alpha, \beta) = \Sigma v_i d_i; \qquad \mathcal{I}_{22}(\alpha, \beta) = \Sigma v_i d_i^2.$$

The MLE's are found by solving the simultaneous equations

$$S_1(\alpha, \beta) = 0; \qquad S_2(\alpha, \beta) = 0.$$

In general, these equations must be solved numerically, and the Newton–Raphson method (10.1.3) can be used.

EXAMPLE 10.5.1. $k = 5$ different doses of an insecticide were applied under standardized conditions to samples of an insect species. The results are shown in Table 10.5.1. We assume that p, the probability that an insect dies, is related to the dose via the logistic model (10.5.1). We wish to find the maximum likelihood estimates $(\hat{\alpha}, \hat{\beta})$.

Based only on the data for dose d_i, we would estimate p_i by y_i/n_i and the log-odds by

$$\log \frac{y_i/n_i}{1 - y_i/n_i} = \log \frac{y_i}{n_i - y_i}.$$

These values are given in the last row of the table, and are plotted versus the dose in Figure 10.5.2. A straight line has been drawn in by eye. If the logistic model holds, then (10.5.2) implies that the five points should be scattered about a straight line. The agreement with a straight line is very good in this example.

From Figure 10.5.2, we see that $\alpha \approx -5$ and $\beta \approx 3$, and we use these as starting values for the Newton–Raphson method. Taking $\alpha = -5$ and $\beta = 3$, we compute p_i, $\mu_i = n_i p_i$, and $v_i = n_i p_i(1 - p_i)$ for $i = 1, 2, \ldots, 5$. Using these values and the d_i's from Table 10.5.1, we then get

$$S_1 = 11.195; \qquad S_2 = 19.031;$$

$$\mathcal{I}_{11} = 40.11; \qquad \mathcal{I}_{12} = 66.85; \qquad \mathcal{I}_{22} = 118.44.$$

Table 10.5.1. Data from a Dose-Response Experiment

Concentration (mg/l)	2.6	3.8	5.1	7.7	10.2
Log concentration d_i	0.96	1.34	1.63	2.04	2.32
Number of insects n_i	50	48	46	49	50
Number killed y_i	6	16	24	42	44
Fraction killed	0.12	0.33	0.52	0.86	0.88
$\log(y_i/(n_i - y_i))$	−1.99	−0.69	0.09	1.79	1.99

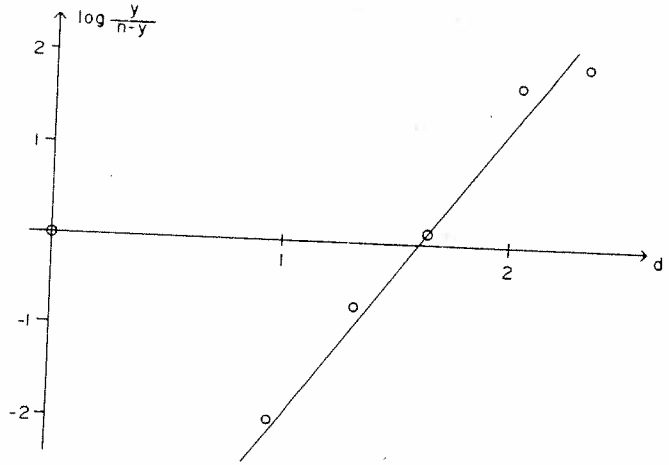

Figure 10.5.2. Plot of estimated log-odds versus dose.

The inverse of the information matrix is

$$\mathscr{I}^{-1} = \begin{bmatrix} 40.11 & 66.85 \\ 66.85 & 118.44 \end{bmatrix}^{-1} = \begin{bmatrix} 0.4196 & -0.2368 \\ -0.2368 & 0.1421 \end{bmatrix}$$

and by (10.1.3), the improved estimates are given by

$$\begin{bmatrix} -5 \\ 3 \end{bmatrix} + \begin{bmatrix} 0.4196 & -0.2368 \\ -0.2368 & 0.1421 \end{bmatrix} \begin{bmatrix} 11.195 \\ 19.031 \end{bmatrix} = \begin{bmatrix} -4.8094 \\ 3.0531 \end{bmatrix}.$$

We now repeat the calculations with $\alpha = -4.8094$ and $\beta = 3.0531$. After two more iterations, we obtain

$$\hat{\alpha} = -4.8869; \qquad \hat{\beta} = 3.1035$$

correct to four decimals. The maximum of the log likelihood is

$$l(\hat{\alpha}, \hat{\beta}) = -119.894,$$

and the information matrix is

$$\mathscr{I}(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix} 39.091 & 62.785 \\ 62.785 & 107.491 \end{bmatrix}.$$

The estimated dose response model is

$$\hat{p} = 1 - (1 + e^{-4.8869 + 3.1035d})^{-1},$$

and from this we can find the estimated kill probability $\hat{p}$ for any given dose d. For instance, at concentration 6 mg/l, the dose is $d = \log 6$, and the estimated kill probability is $\hat{p} = 0.662$.

Using this result, we find that the estimated kill probability at concentration 2.6 is $\hat{p}_1 = 0.1277$. The estimated expected number of insects killed is $\hat{\mu}_1 = n_1 \hat{p}_1 = 6.39$, and the expected number surviving is $n_1(1 - \hat{p}_1) = 43.61$. Table 10.5.2 shows the observed and expected frequencies for the five doses used in the experiment. The agreement is very close, indicating that the logistic model gives a good description of the data.

Table 10.5.2. Observed Frequencies of Insects Killed and Surviving, and Expected Frequencies Under a Logistic Model

Concentration	Number killed Observed (expected)	Number surviving Observed (expected)	Total
2.6	6 (6.39)	44 (43.61)	50
3.8	16 (15.47)	32 (32.53)	48
5.1	24 (24.94)	22 (21.06)	46
7.7	42 (39.68)	7 (9.32)	49
10.2	44 (45.53)	6 (4.47)	50

Estimation of the ED50

The ED50 is the dose γ, say, which would be required to produce a 50% response rate (see Figure 10.5.1). Since $p(\gamma) = \frac{1}{2}$, we have

$$0 = \log \frac{p(\gamma)}{1 - p(\gamma)} = \alpha + \beta\gamma,$$

and it follows that $\gamma = -\alpha/\beta$. By the invariance property, we have $\hat{\gamma} = -\hat{\alpha}/\hat{\beta}$.

Usually γ is of more interest than the intercept parameter α, and so we consider a parameter transformation from (α, β) to (γ, β). The logistic model (10.5.2) then becomes

$$\log \frac{p_i}{1 - p_i} = \beta(d_i - \gamma).$$

The log RLF of (α, β) is

$$r(\alpha, \beta) = l(\alpha, \beta) - l(\hat{\alpha}, \hat{\beta}).$$

Substituting $\alpha = -\gamma\beta$ gives the log RLF of (γ, β):

$$r_*(\gamma, \beta) = r(-\gamma\beta, \beta) = l(-\gamma\beta, \beta) - l(\hat{\alpha}, \hat{\beta}).$$

EXAMPLE 10.5.1 (continued). The MLE of the ED50 is

$$\hat{\gamma} = -\hat{\alpha}/\hat{\beta} = 1.5746$$

and the log RLF of (γ, β) is

$$r_*(\gamma, \beta) = l(-\gamma\beta, \beta) + 119.894.$$

Figure 10.5.3 shows contours of constant relative likelihood in the (γ, β) plane. The contours are close to elliptical in shape, and thus the normal approximations of Section 10.4 should give accurate results. Since the axes of the ellipses are nearly parallel to the parameter axes, the range of plausible values for γ is nearly independent of the value of β.

If contours are plotted in the (α, β) plane, their axes are not parallel to the coordinate axes, and the range of plausible values for α is strongly dependent upon the value of β. This is another reason for changing parameters from (α, β) to (γ, β).

To find $r_{\max}(\gamma)$, it is necessary to maximize $l(-\gamma\beta, \beta)$ over β for fixed γ. Define

$$g(\beta) = \frac{\partial}{\partial \beta} l(-\gamma\beta, \beta) = \Sigma(d_i - \gamma)(y_i - \mu_i);$$

$$g'(\beta) = \frac{\partial}{\partial \beta} g(\beta) = -\Sigma(d_i - \gamma)^2 v_i$$

where $\mu_i = n_i p_i$ and $v_i = n_i p_i(1 - p_i)$ as before. For any given value of γ, we can solve the equation $g(\beta) = 0$ by Newton's method to obtain $\hat{\beta}(\gamma)$, and then calculate

$$r_{\max}(\gamma) = l(-\gamma\hat{\beta}(\gamma), \hat{\beta}(\gamma)) + 119.894.$$

After repeating this procedure for several γ values, a graph of this function can be prepared (see Figure 10.5.4).

To find the normal approximation to $r_{\max}(\gamma)$, we need the information matrix $\mathscr{I}_*(\hat{\gamma}, \hat{\beta})$ for the new parameters. The matrix of derivatives of (α, β) with

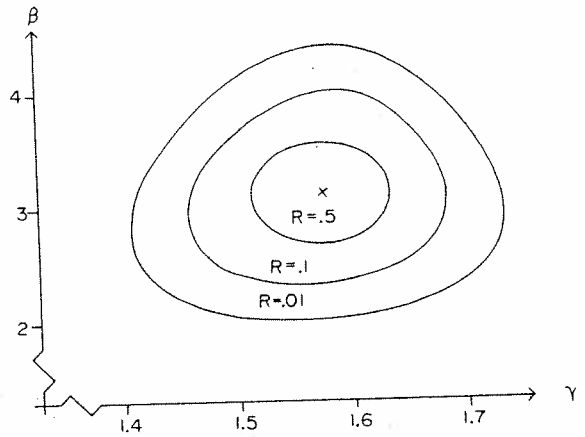

Figure 10.5.3. Contours of constant relative likelihood in the (γ, β) plane.

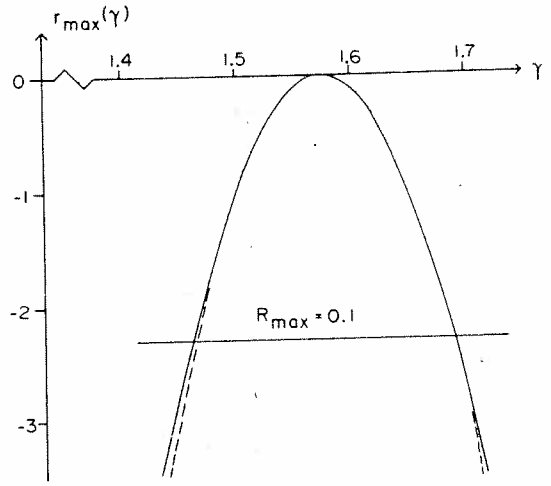

Figure 10.5.4. Maximum log RLF (solid curve) and normal approximation (broken curve) for the ED50.

respect to (γ, β) is

$$Q = \begin{bmatrix} \partial\alpha/\partial\gamma & \partial\alpha/\partial\beta \\ \partial\beta/\partial\gamma & \partial\beta/\partial\beta \end{bmatrix} = \begin{bmatrix} -\beta & -\gamma \\ 0 & 1 \end{bmatrix}$$

and now (10.4.2) gives

$$\mathscr{I}_*(\hat{\gamma}, \hat{\beta}) = \begin{bmatrix} -\hat{\beta} & -\hat{\gamma} \\ 0 & 1 \end{bmatrix}^t \mathscr{I}(\hat{\alpha}, \hat{\beta}) \begin{bmatrix} -\hat{\beta} & -\hat{\gamma} \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 376.527 & -3.819 \\ -3.819 & 6.691 \end{bmatrix}.$$

From this we calculate

$$\hat{\mathscr{I}}^*_{11} - (\hat{\mathscr{I}}^*_{12})^2/\hat{\mathscr{I}}^*_{22} = 374.35$$

and hence the normal approximation is

$$r_{\max}(\gamma) \approx -\tfrac{1}{2}(\gamma - \hat{\gamma})^2(374.35).$$

The approximation is shown with a broken curve in Figure 10.5.4, and the agreement with the exact curve is very close. The approximate 10% maximum likelihood interval is $1.464 \le \gamma \le 1.686$, and the exact result is $1.460 \le \gamma \le 1.686$.

Problems for Section 10.5

1. The following table gives the number of beetles which died within 6 days and the number which survived at each of six concentrations of an insecticide.

Concentration	0.711	0.852	0.959	1.066	1.202	1.309
Number dead	15	24	26	24	29	29
Number alive	35	25	24	26	21	20

Assume that the log-odds in favor of death is a linear function of the dose d,

$$\log \frac{p}{1-p} = \alpha + \beta d$$

where d is the log concentration.

(a) Prepare a graph to check whether the model seems reasonable, and from it obtain initial estimates of α and β.
(b) Obtain the maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}$ by the Newton–Raphson method.
(c) Estimate the concentration of the insecticide which is required to obtain a 50% kill probability.
(d) Find the 10% maximum likelihood intervals for α and β.

2.† The probability of a normal specimen after radiation dose d is assumed to be $p = e^{\alpha+\beta d}$ where α and β are constants. The following table gives the number of normal specimens and the total number tested at each of five doses:

d = Radiation dose	0	1	2	3	4
y = Number of normals	4357	3741	3373	2554	1914
n = Number tested	4358	3852	3605	2813	2206

(a) Plot $\log(y/n)$ against d to check whether the model seems reasonable, and obtain rough estimates of α and β from the graph.
(b) Find the maximum likelihood equations and solve numerically for $\hat{\alpha}$ and $\hat{\beta}$ using the Newton–Raphson method or otherwise. Plot the 10% likelihood contour, and obtain 10% maximum likelihood intervals for β and e^{α}.

3.† The number of particles emitted in unit time from a radioactive source has a Poisson distribution. The strength of the source is decaying exponentially with time, and the mean of the Poisson distribution on the jth day is $\mu_j = \alpha\beta^j$ $(j = 0, 1, \ldots, n)$. Independent counts $x_0, x_1, \ldots, x_n$ of the number of emissions in unit time are obtained on these $n+1$ days. Find the maximum likelihood equations and indicate how these may be solved for $\hat{\alpha}$ and $\hat{\beta}$.

4. Observations $y_1, y_2, \ldots, y_n$ are taken on the number of plankton in unit-volume samples of seawater at temperatures $x_1, x_2, \ldots, x_n$. The y_i's are modeled as observed values of independent Poisson variates $Y_1, Y_2, \ldots, Y_n$, where

$$\mu_i = E(Y_i) = \exp(\alpha + \beta x_i).$$

(a) Show that the log likelihood function is

$$l(\alpha, \beta) = \Sigma(y_i \log \mu_i - \mu_i).$$

(b) Find the score vector and information matrix for α and β, and describe how to obtain $\hat{\alpha}$ and $\hat{\beta}$ by the Newton–Raphson method.

(c) Show that $\hat{\beta}$ is a root of the equation

$$(\Sigma x_i y_i)(\Sigma e^{\beta x_i}) - (\Sigma y_i)(\Sigma x_i e^{\beta x_i}) = 0,$$

and describe how $\hat{\beta}$ can be found by Newton's method.
(d) Derive the maximum RLF of β.

5. The survival time Y_i of an individual with tumor size x_i has an exponential distribution with mean

$$\theta_i = E(Y_i) = \exp(\alpha + \beta x_i)$$

where α and β are unknown parameters. Suppose that n survival times $y_1, y_2, \ldots, y_n$ with corresponding tumor sizes $x_1, x_2, \ldots, x_n$ are observed.

(a) Show that the score vector and information matrix of α and β are as follows:

$$S = \begin{bmatrix} \Sigma(r_i - 1) \\ \Sigma x_i(r_i - 1) \end{bmatrix}; \qquad \mathcal{I} = \begin{bmatrix} \Sigma r_i & \Sigma x_i r_i \\ \Sigma x_i r_i & \Sigma x_i^2 r_i \end{bmatrix}$$

where $r_i = y_i/\theta_i$.
(b) Show that the determinant of $\mathcal{I}$ is Σr_i times

$$\Sigma r_i (x_i - \bar{x})^2$$

where $\bar{x} = (\Sigma r_i x_i)/\Sigma r_i$. Hence verify that condition (10.1.2) is satisfied.
(c) Derive an expression for the MLE of α when β is given, and describe a numerical procedure for evaluating $\hat{\beta}$.

10.6. An Example from Learning Theory

In their book *Stochastic Models for Learning* (Wiley, 1955), R.R. Bush and F. Mosteller develop general probabilistic learning models and apply them to a variety of learning experiments. One of the most interesting applications is to the Solomon–Wynne experiment (R.L. Solomon and L.C. Wynne, Traumatic Avoidance Learning: Acquisition in Normal Dogs, *Psych. Monog.* **67** (1953), No. 4). We shall first describe this experiment, then develop the model, and finally use likelihood methods to estimate the two parameters of the model.

In the Solomon–Wynne experiment, 30 dogs learned to avoid an intense electric shock by jumping a barrier. The lights were turned out in the dog's compartment and the barrier was raised. Ten seconds later, an intense shock was applied through the floor of the compartment to the dog's feet, and was left on until the dog escaped over the barrier. The dog could avoid the shock only by jumping the barrier during the ten-second interval after the lights were turned out and before the shock was administered. Each trial could thus be classified as a shock trial, or as an avoidance trial. The experimental record of 30 dogs, each of which had 25 trials, is shown in Table 10.6.1, with 0

Table 10.6.1. Data from 25 Trials with 30 Dogs in the Solomon–Wynne Experiment

	0 = Shock trial; 1 = Avoidance trial				
	Trial numbers				
	0–4	5–9	10–14	15–19	20–24
Dog 13	0 0 1 0 1	0 1 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
16	0 0 0 0 0	0 0 1 0 0	0 0 0 0 1	1 1 1 1 1	1 1 1 1 1
17	0 0 0 0 0	1 1 0 1 1	0 0 1 1 0	1 0 1 1 1	1 1 1 1 1
18	0 1 1 0 0	1 1 1 1 0	1 0 1 0 1	1 1 1 1 1	1 1 1 1 1
21	0 0 0 0 0	0 0 0 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
27	0 0 0 0 0	0 1 1 1 1	0 0 1 0 1	1 1 1 1 1	1 1 1 1 1
29	0 0 0 0 0	1 0 0 0 0	0 0 1 1 1	1 1 1 1 1	1 1 1 1 1
30	0 0 0 0 0	0 0 1 1 0	0 1 1 1 1	1 1 1 1 1	1 1 1 1 1
32	0 0 0 0 0	1 0 1 0 1	1 0 1 0 0	0 1 1 1 1	1 0 1 1 0
33	0 0 0 0 1	0 0 1 1 0	1 0 1 1 1	1 1 1 1 1	1 1 1 1 1
34	0 0 0 0 0	0 0 0 0 0	1 1 1 1 1	1 0 1 1 1	1 1 1 1 1
36	0 0 0 0 0	1 1 1 1 1	0 0 1 1 1	1 1 1 1 1	1 1 1 1 1
37	0 0 0 1 1	0 1 0 0 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
41	0 0 0 0 1	0 1 1 0 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
42	0 0 0 1 0	1 1 0 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
43	0 0 0 0 0	0 0 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
45	0 1 0 1 0	0 0 1 0 1	1 1 1 0 1	1 1 1 1 1	1 1 1 1 1
47	0 0 0 0 1	0 1 0 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
48	0 1 0 0 0	0 1 0 0 0	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
46	0 0 0 0 1	1 0 1 0 1	1 0 1 0 1	1 1 1 1 1	1 1 1 1 1
49	0 0 0 1 1	1 1 1 0 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
50	0 0 1 0 1	0 1 1 1 1	1 1 1 1 1	1 0 0 1 1	1 1 1 1 1
52	0 0 0 0 0	0 0 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
54	0 0 0 0 0	0 0 0 1 1	1 0 1 0 0	0 1 1 0 1	1 1 1 1 1
57	0 0 0 0 0	0 1 0 1 1	1 1 0 1 0	1 1 1 1 1	1 1 1 1 1
59	0 0 1 0 1	1 1 0 1 1	0 1 1 1 1	1 1 1 1 1	1 1 1 1 1
67	0 0 0 0 1	0 1 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1
66	0 0 0 1 0	1 0 1 1 1	0 1 0 1 1	1 1 1 1 1	1 1 1 1 1
69	0 0 0 0 1	1 0 0 1 1	1 0 1 0 1	0 1 0 1 1	1 1 1 1 1
71	0 0 0 0 1	1 1 1 1 1	0 1 0 1 1	1 1 1 1 1	1 1 1 1 1

denoting a shock trial and 1 an avoidance trial. (The dogs are numbered 13, 16, etc. for identification purposes, and no use is made of these numbers in the analysis.) Initially, all of the dogs received shocks on almost every trial, but by trial 20, all except dog number 32 had learned to avoid the shock by jumping the barrier.

The Model

Consider the sequence of trials for one dog. As in Table 10.6.1 we take $y_j = 1$ if the dog avoids shock at trial j, and $y_j = 0$ otherwise ($j = 0, 1, \ldots, 24$). Because of learning at trial $j-1$, the probability that the dog receives shock should be smaller at trial j than at trial $j-1$. The amount by which the probability decreases may well depend upon whether there was shock or avoidance at trial $j-1$. We wish to compare the effectiveness of shock trials and avoidance trials in teaching the dog to avoid future shocks.

Let ϕ_j be the probability that the dog receives a shock at trial j, given its past history in trials 0 through $j-1$. Let x_j be the number of times that the dog has avoided shock in trials 0 through $j-1$, so that

$$x_j = y_0 + y_1 + \cdots + y_{j-1}.$$

The number of previous shock trials is then $j - x_j$. Since all dogs were given a shock at trial 0, we assume that $\phi_0 = 1$. For $j > 0$ we assume that

$$\phi_j = A^{x_j} B^{j - x_j} \tag{10.6.1}$$

where $0 \le A \le 1$ and $0 \le B \le 1$. We call A the *avoidance parameter* and B the *shock parameter*. The model can also be written

$$\log \phi_j = x_j \alpha + (j - x_j)\beta \tag{10.6.2}$$

where $\alpha = \log A$ and $\beta = \log B$.

It is easy to show that

$$\frac{\phi_j}{\phi_{j-1}} = \begin{cases} A & \text{if } y_{j-1} = 1; \\ B & \text{if } y_{j-1} = 0. \end{cases}$$

The probability of a shock decreases by the factor A if there was an avoidance at trial $j-1$, or by the factor B if there was a shock at trial $j-1$. If A is small, then the effect of an avoidance trial is to greatly reduce the chance of future shock. If $A = 1$, nothing is learned from an avoidance trial. If $A < B$, then more is learned from an avoidance trial than from a shock trial.

The Log Likelihood Function

The joint probability function of $Y_0, Y_1, \ldots, Y_{24}$ can be written as a product of 25 factors:

$$f(y_0, y_1, \ldots, y_n) = f(y_0) \cdot f(y_1 | y_0) \cdot f(y_2 | y_0, y_1) \cdot \cdots.$$

Given the results of trials 0 through $j-1$, the probability function of Y_j is

$$f(y_j | y_0, \ldots, y_{j-1}) = \phi_j^{1-y_j}(1 - \phi_j)^{y_j} = \begin{cases} \phi_j & \text{for } y_j = 0; \\ 1 - \phi_j & \text{for } y_j = 1. \end{cases}$$

Since $f(y_0) = 1$ for $y_0 = 1$, this term makes no contribution, and therefore

$$f(y_0, y_1, \ldots, y_{24}) = \prod_{j=1}^{24} \phi_j^{1-y_j}(1-\phi_j)^{y_j}.$$

The log likelihood function based on the data from a single dog is thus

$$\sum_{j=1}^{24} [(1-y_j)\log\phi_j + y_j \log(1-\phi_j)].$$

Now we assume that results for different dogs are independent, and that α and β have the same values for all 30 dogs. Then the log likelihood function based on all of the data is

$$l(\alpha, \beta) = \sum_{i=1}^{30}\sum_{j=1}^{24} [(1-y_{ij})\log\phi_{ij} + y_{ij}\log(1-\phi_{ij})]$$
$$= \alpha T_1 + \beta T_2 + \Sigma\Sigma y_{ij}\log(1-\phi_{ij})$$

where

$$\log\phi_{ij} = x_{ij}\alpha + (j - x_{ij})\beta;$$
$$x_{ij} = y_{i0} + y_{i1} + \cdots + y_{ij-1};$$
$$T_1 = \Sigma\Sigma(1-y_{ij})x_{ij};$$
$$T_2 = \Sigma\Sigma(1-y_{ij})(j-x_{ij}).$$

It is easily shown that

$$\frac{\partial\phi_{ij}}{\partial\alpha} = x_{ij}\phi_{ij}; \qquad \frac{\partial\phi_{ij}}{\partial\beta} = (j-x_{ij})\phi_{ij}.$$

The components of the score function are then

$$S_1(\alpha, \beta) = T_1 - \Sigma\Sigma y_{ij}x_{ij}\phi_{ij}/(1-\phi_{ij});$$
$$S_2(\alpha, \beta) = T_2 - \Sigma\Sigma y_{ij}(j-x_{ij})\phi_{ij}/(1-\phi_{ij}).$$

The components of the information matrix are

$$\mathcal{I}_{11}(\alpha, \beta) = \Sigma\Sigma y_{ij}x_{ij}^2\phi_{ij}/(1-\phi_{ij})^2;$$
$$\mathcal{I}_{12}(\alpha, \beta) = \Sigma\Sigma y_{ij}x_{ij}(j-x_{ij})\phi_{ij}/(1-\phi_{ij})^2;$$
$$\mathcal{I}_{22}(\alpha, \beta) = \Sigma\Sigma y_{ij}(j-x_{ij})^2\phi_{ij}/(1-\phi_{ij})^2.$$

Numerical Results

The above expressions involve sums of 720 terms, and one would certainly not wish to attempt the calculations by hand! Before high speed computers were available, the analysis of data such as these was very difficult, and often depended upon approximations whose accuracy could not be checked. However, now we require only a minute or two on a fast computer to find the

MLE's and plot contours, thus obtaining an exact summary of the information concerning the parameters.

A preliminary tabulation of $l(\alpha, \beta)$ indicates that the maximum occurs near $(\alpha, \beta) = (-0.1, -0.2)$. Taking these as initial values, the MLE's can be found by the Newton–Raphson method (10.1.3). After three iterations we obtain

$$\hat{\alpha} = -0.24091; \qquad \hat{\beta} = -0.07872$$
$$l(\hat{\alpha}, \hat{\beta}) = -273.987$$
$$\mathcal{I}(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix} 2451 & 2784 \\ 2784 & 10277 \end{bmatrix}.$$

The MLE's of the original parameters A, B are

$$\hat{A} = e^{\hat{\alpha}} = 0.786; \qquad \hat{B} = e^{\hat{\beta}} = 0.924.$$

The log RLF of A and B is

$$r_*(A, B) = l(\log A, \log B) + 273.987,$$

and by (10.4.2), the information matrix is

$$\mathcal{I}_*(\hat{A}, \hat{B}) = \begin{bmatrix} 1/\hat{A} & 0 \\ 0 & 1/\hat{B} \end{bmatrix}^t \mathcal{I}(\hat{\alpha}, \hat{\beta}) \begin{bmatrix} 1/\hat{A} & 0 \\ 0 & 1/\hat{B} \end{bmatrix} = \begin{bmatrix} 3969 & 3833 \\ 3833 & 12029 \end{bmatrix}.$$

Figure 10.6.1 shows contours of constant relative likelihood in the (A, B) plane. Since $A < B$ over the entire region of plausible values, an avoidance trial is clearly more effective than a shock trial in reducing the probability of future shock. In fact, since $\hat{A} \approx \hat{B}^3$, about as much is learned in one avoidance

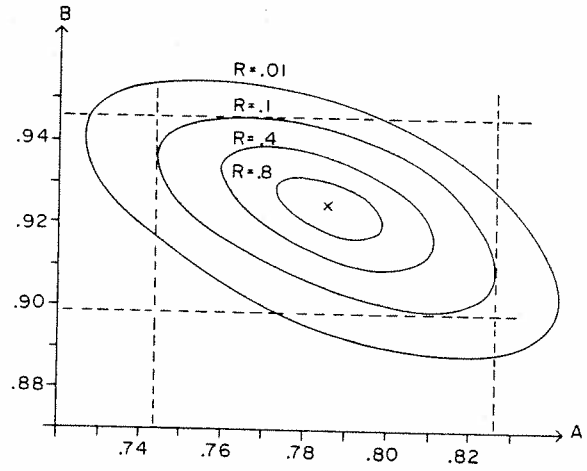

Figure 10.6.1. Contours of constant relative likelihood for the avoidance and shock parameters.

trial as in three shock trials. Note also that the experiment determines the value of B more precisely than the value of A, and that the parameters A, B cannot be estimated independently of one another.

The contours in Figure 10.6.1 are nearly elliptical, and therefore we would expect the normal approximations of Section 10.4 to be fairly accurate. By (10.4.4), we have

$$r_{\max}(A) \approx -\tfrac{1}{2}(A-\hat{A})^2(3969-3833^2/12029);$$
$$r_{\max}(B) \approx -\tfrac{1}{2}(B-\hat{B})^2(12029-3833^2/3969).$$

From these, we obtain approximate 10% maximum likelihood intervals $0.745 \le A \le 0.827$ and $0.901 \le B \le 0.948$. The exact results from Figure 10.6.1 are $0.744 \le A \le 0.826$ and $0.899 \le B \le 0.946$.

*10.7. Some Derivations

In this section, derivations will be given for some results quoted earlier in the chapter. First we shall derive the Newton–Raphson procedure for evaluating the MLE $(\hat{\alpha}, \hat{\beta})$ in the two-parameter case. Then conditions (10.1.1) and (10.1.2) which must be satisfied at a relative maximum will be established. Finally, the use of Newton's method to solve for points on a likelihood contour will be described.

The Newton–Raphson Method

Newton's method for solving the equation $g(\theta)=0$ was derived at the beginning of Section 9.8. We considered a linear approximation to $g(\theta)$ at $\theta=\theta_0$:

$$g(\theta) \approx g(\theta_0) + (\theta-\theta_0)g'(\theta_0).$$

The approximation has the same value and slope as $g(\theta)$ at the point $\theta=\theta_0$. Since $g(\hat{\theta})=0$, we have

$$\hat{\theta} \approx \theta_0 - g(\theta_0)/g'(\theta_0).$$

Beginning with a preliminary guess θ_0, we apply this result repeatedly to obtain $\hat{\theta}$ such that $g(\hat{\theta})=0$.

In the two-parameter case we need to solve a pair of simultaneous equations

$$g(\alpha,\beta)=0; \qquad h(\alpha,\beta)=0.$$

*This section may be omitted on first reading.

Let (α_0, β_0) be a preliminary estimate of the root $(\hat{\alpha}, \hat{\beta})$, and consider the linear approximations

$$g(\alpha,\beta) \approx g(\alpha_0,\beta_0) + (\alpha-\alpha_0)\frac{\partial g}{\partial \alpha} + (\beta-\beta_0)\frac{\partial g}{\partial \beta}$$

$$h(\alpha,\beta) \approx h(\alpha_0,\beta_0) + (\alpha-\alpha_0)\frac{\partial h}{\partial \alpha} + (\beta-\beta_0)\frac{\partial h}{\partial \beta}.$$

Here the derivatives are to be evaluated at $\alpha=\alpha_0$, $\beta=\beta_0$. These linear approximations can be derived by truncating the bivariate Taylor's series expansions of g and h at (α_0, β_0). They have the same values and first derivatives as g and h at the point (α_0, β_0).

Since $g(\hat{\alpha}, \hat{\beta})=0$ and $h(\hat{\alpha}, \hat{\beta})=0$, we have

$$(\hat{\alpha}-\alpha_0)\frac{\partial g}{\partial \alpha} + (\hat{\beta}-\beta_0)\frac{\partial g}{\partial \beta} \approx -g(\alpha_0,\beta_0);$$

$$(\hat{\alpha}-\alpha_0)\frac{\partial h}{\partial \alpha} + (\hat{\beta}-\beta_0)\frac{\partial h}{\partial \beta} \approx -h(\alpha_0,\beta_0).$$

Solving these linear equations for $\hat{\alpha}-\alpha_0$ and $\hat{\beta}-\beta_0$ gives

$$\begin{bmatrix} \hat{\alpha}-\alpha_0 \\ \hat{\beta}-\beta_0 \end{bmatrix} \approx -\begin{bmatrix} \frac{\partial g}{\partial \alpha} & \frac{\partial g}{\partial \beta} \\ \frac{\partial h}{\partial \alpha} & \frac{\partial h}{\partial \beta} \end{bmatrix}^{-1} \begin{bmatrix} g \\ h \end{bmatrix}$$

and therefore

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \approx \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} - \begin{bmatrix} \frac{\partial g}{\partial \alpha} & \frac{\partial g}{\partial \beta} \\ \frac{\partial h}{\partial \alpha} & \frac{\partial h}{\partial \beta} \end{bmatrix}^{-1} \begin{bmatrix} g \\ h \end{bmatrix}. \tag{10.7.1}$$

Beginning with a preliminary guess (α_0, β_0), we apply (10.7.1) repeatedly until convergence is obtained. This generalization of Newton's method is called the Newton–Raphson method.

Solving the Maximum Likelihood Equations

To obtain the MLE $(\hat{\alpha}, \hat{\beta})$ in the two-parameter case, we need to solve the simultaneous equations

$$S_1(\alpha,\beta)=0; \qquad S_2(\alpha,\beta)=0$$

where S_1 and S_2 are the components of the score vector (see Section 10.1). In

this case we have

$$\frac{\partial g}{\partial \alpha} = \frac{\partial S_1}{\partial \alpha} = \frac{\partial^2 l}{\partial \alpha^2} = -\mathscr{I}_{11}(\alpha, \beta),$$

and similarly for the other derivatives. Thus (10.7.1) becomes

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \approx \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \mathscr{I}_{11} & \mathscr{I}_{12} \\ \mathscr{I}_{12} & \mathscr{I}_{22} \end{bmatrix}^{-1} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \tag{10.7.2}$$

where S_1, S_2, and the $\mathscr{I}_{ij}$'s are all evaluated at $\alpha = \alpha_0$, $\beta = \beta_0$. This is the result stated in Section 10.1.

In many applications, $l(\alpha, \beta)$ is approximately a quadratic function of α and β near $(\hat{\alpha}, \hat{\beta})$ (see Section 10.4). Then S_1 and S_2 will be nearly linear in α and β near $(\hat{\alpha}, \hat{\beta})$. The Newton–Raphson procedure will then converge quickly provided that the initial guess (α_0, β_0) is not too far from $(\hat{\alpha}, \hat{\beta})$.

In Example 10.1.1 the log-likelihood function $l(\mu_1, \mu_2)$ is a second-degree polynomial in μ_1 and μ_2, and the components of the score function are linear in μ_1 and μ_2. In this case (10.7.2) will yield $(\hat{\mu}_1, \hat{\mu}_2)$ in one iteration, no matter what initial values are taken for μ_1 and μ_2.

Derivation of (10.1.1) and (10.1.2)

If $l(\alpha, \beta)$ has a relative maximum at $(\hat{\alpha}, \hat{\beta})$, then it must be "down hill" from $(\hat{\alpha}, \hat{\beta})$ in all directions.

Suppose that we move away from $(\hat{\alpha}, \hat{\beta})$ along a line at angle ϕ to the α-axis (see Figure 10.7.1). For points (α, β) on this line we have

$$\alpha = \hat{\alpha} + d\cos\phi; \qquad \beta = \hat{\beta} + d\sin\phi$$

where d is the distance from (α, β) to $(\hat{\alpha}, \hat{\beta})$. Along this line the log likelihood

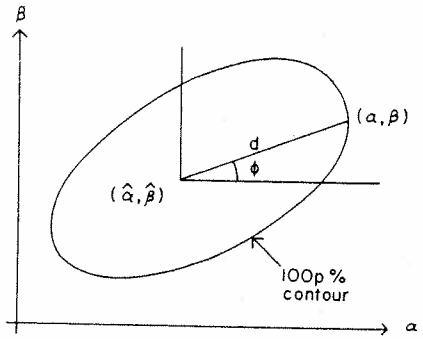

Figure 10.7.1. Calculation of a point on the $100p\%$ likelihood contour by Newton's method.

function is

$$h(d) = l(\alpha, \beta) = l(\hat{\alpha} + d\cos\phi, \hat{\beta} + d\sin\phi).$$

For each possible angle ϕ, $h(d)$ must have a relative maximum at $d = 0$. It follows that, for all ϕ,

$$\frac{\partial h(d)}{\partial d} = 0 \qquad \text{and} \qquad \frac{\partial^2 h(d)}{\partial d^2} < 0 \qquad \text{at } d = 0.$$

Using the chain rule, we find that

$$\frac{\partial h(d)}{\partial d} = \frac{\partial l}{\partial \alpha}\frac{\partial \alpha}{\partial d} + \frac{\partial l}{\partial \beta}\frac{\partial \beta}{\partial d} = S_1\cos\phi + S_2\sin\phi.$$

Since this must be zero for all ϕ when $d = 0$, it follows that

$$S_1(\hat{\alpha}, \hat{\beta}) = 0; \qquad S_2(\hat{\alpha}, \hat{\beta}) = 0$$

which is (10.1.1).

Upon differentiating again, we find that

$$-\frac{\partial^2 h(d)}{\partial d^2} = \mathscr{I}_{11}\cos^2\phi + 2\mathscr{I}_{12}\cos\phi\sin\phi + \mathscr{I}_{22}\sin^2\phi.$$

Since this must be positive for all ϕ when $d = 0$, it follows that

$$\hat{\mathscr{I}}_{11}\cos^2\phi + 2\hat{\mathscr{I}}_{12}\cos\phi\sin\phi + \hat{\mathscr{I}}_{22}\sin^2\phi > 0 \tag{10.7.3}$$

for all ϕ. It is not difficult to show that this condition is satisfied if and only if (10.1.2) holds.

Calculation of Contours

In Example 10.2.2 we constructed a contour map by tabulating the joint relative likelihood function over a lattice of parameter values. A great deal of calculation may be needed to produce an accurate contour map in this way.

As an alternative, it is possible to use Newton's method to obtain points on the $100p\%$ contour. We select an angle ϕ as in Figure 10.7.1 and iterate to find d such that (α, β) lies on the contour. By repeating this procedure for various angles ϕ, we can obtain enough points on the contour to permit accurate plotting.

The equation of the $100p\%$ contour is

$$r(\alpha, \beta) - \log p = 0$$

and so we consider the function

$$g(d) = r(\hat{\alpha} + d\cos\phi, \hat{\beta} + d\sin\phi) - \log p.$$

Since $r(\alpha, \beta) = l(\alpha, \beta) - l(\hat{\alpha}, \hat{\beta})$, we have

$$g(d) = h(d) - l(\hat{\alpha}, \hat{\beta}) - \log p$$

where $h(d)$ was defined earlier in this section. Thus the derivative of $g(d)$ with respect to d is

$$g'(d) = h'(d) = S_1 \cos \phi + S_2 \sin \phi.$$

To solve the equation $g(d) = 0$ by Newton's method, we begin with an initial guess d_{old}, and compute

$$d_{\text{new}} = d_{\text{old}} - g(d_{\text{old}})/g'(d_{\text{old}}).$$

This calculation is repeated until convergence is obtained. An initial guess can be obtained from a preliminary tabulation of $r(\alpha, \beta)$, or from the large-sample results of Section 10.4.

*10.8. Multi-Parameter Likelihoods

In this section we shall briefly consider likelihood methods in the multi-parameter case. These methods will generally produce satisfactory results provided that the number of unknown parameters is small in comparison with the number of independent observations. However, as we shall see, serious difficulties may arise when many unknown parameters are to be estimated from a relatively small amount of data.

Suppose that the probability model involves a vector of k unknown parameters, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Then the log likelihood function $l(\theta_1, \theta_2, \ldots, \theta_k)$ is a function of k variables, and the MLE $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ is the vector of parameter values which maximizes this function.

In the k-parameter case, the score function $S(\theta)$ is a k-dimensional vector whose ith component is $S_i(\theta) = \partial l/\partial \theta_i$. The information function $\mathcal{I}(\theta)$ is a symmetric $k \times k$ matrix whose (i, j) component is $-\partial^2 l/\partial \theta_i \partial \theta_j$.

Usually $\hat{\theta}$ can be found by solving the k simultaneous equations $S_i(\theta) = 0$ for $i = 1, 2, \ldots, k$. The condition for a relative maximum is that $\mathcal{I}(\hat{\theta})$ must be non-negative definite. Numerical methods may be needed to find $\hat{\theta}$, and the Newton–Raphson method is often useful.

The relative likelihood function is

$$R(\theta_1, \ldots, \theta_k) = L(\theta_1, \ldots, \theta_k)/L(\hat{\theta}_1, \ldots, \hat{\theta}_k).$$

In principle, one can plot this function to determine ranges of plausible values for the parameters. In practice, it is difficult to plot and interpret the RLF when $k > 2$, and so parameters are usually considered one or two at a time.

The maximum relative likelihood function of θ_1 is obtained by holding θ_1 fixed and maximizing the joint RLF over the other $k - 1$ parameters:

$$R_{\max}(\theta_1) = \underset{\theta_1 \text{ fixed}}{\text{maximum}} R(\theta_1, \theta_2, \ldots, \theta_k).$$

If the number of unknown parameters is small in comparison with the number of independent observations, then the maximum RLF has properties similar to those of a one-parameter RLF. Maximum likelihood intervals or regions can be used to summarize the information available concerning θ_1. The $100p\%$ maximum likelihood region in the set of parameter values such that $R_{\max}(\theta_1) \geq p$ (see Section 10.3).

The amount of information available for estimating θ_1 may well depend upon which other parameters $\theta_2, \ldots, \theta_k$ are to be estimated from the same data. For instance, an observation y whose distribution depends on the product $\theta_1 \theta_2$ will give information about θ_1 if θ_2 is known, but it may yield little or no information about θ_1 if θ_2 is unknown. In general, one would expect to be able to estimate θ_1 more precisely if the values of $\theta_2, \ldots, \theta_k$ were all known.

The maximum relative likelihood function does not adequately adjust for the possible loss of information about θ_1 due to the estimation of $\theta_2, \ldots, \theta_k$. As a result, the maximum RLF may suggest that θ_1 can be estimated more precisely than is appropriate in the situation.

This is not a serious problem when there are many observations and only a few parameters, since then only a small fraction of the information will be lost due to estimation of $\theta_2, \ldots, \theta_k$. However, as the following examples show, serious difficulties may arise when many parameters require estimation from only a small amount of data.

EXAMPLE 10.8.1. Suppose that a single measurement y is taken from a normal distribution $N(\mu, \sigma^2)$.

If μ is known, then y provides information concerning the magnitude of σ^2. The closer y is to μ, the smaller the variance is likely to be. The MLE of σ^2 is $\hat{\sigma}^2 = (y - \mu)^2$, and likelihood intervals for σ^2 can be found in the usual way.

If μ is unknown, then a single observation y will tell us nothing about the variance. The proper conclusion is that it is impossible to estimate σ^2 from a single observation if μ is unknown.

If we apply the method of maximum likelihood here, we find that $\hat{\mu} = y$ and $\hat{\sigma}^2 = (y - \hat{\mu})^2 = 0$. No matter what is observed, σ^2 is estimated to be zero. This is a silly estimate. It arises because the method of maximum likelihood treats $\hat{\mu}$ as though it were the known value of μ. The method does not allow for the fact that all of the information provided by y is lost in the process of estimating μ.

EXAMPLE 10.8.2. Suppose that we observe values of $n + 1$ independent random variables $Y_0, Y_1, \ldots, Y_n$ where $Y_i \sim N(\mu_i, \sigma^2)$. Suppose that μ_0 is known, but that $\mu_1, \mu_2, \ldots, \mu_n$, and σ are all unknown. In this case there are $n + 1$ unknown parameters to be estimated from $n + 1$ observations.

*This section may be omitted on first reading.

As in the preceding example, it can be argued that the information provided by $y_1, y_2, \ldots, y_n$ is lost in estimating $\mu_1, \mu_2, \ldots, \mu_n$. Only observation u_0 gives information about σ^2. Based on just this observation, the log likelihood function is

$$l(\sigma) = -\log \sigma - \frac{1}{2\sigma^2}(y_0 - \mu_0)^2,$$

and the MLE of σ^2 is $(y_0 - \mu_0)^2$.

The likelihood function based on all $n+1$ observations is

$$l(\mu_1, \ldots, \mu_n, \sigma) = -(n+1)\log \sigma - \frac{1}{2\sigma^2}\sum_{i=0}^{n}(y_i - \mu_i)^2.$$

By a straightforward calculation we find that $\hat{\mu}_i = y_i$ for $i = 1, 2, \ldots, n$, and

$$\hat{\sigma}^2 = \frac{1}{n+1}\left[(y_0 - \mu_0)^2 + \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2\right] = \frac{1}{n+1}(y_0 - \mu_0)^2.$$

If n is large, this estimate will be much too small. Furthermore, because of the serious underestimation of σ^2, maximum likelihood intervals for $\mu_1, \mu_2, \ldots, \mu_n$ will be too narrow. The analysis does not take account of the fact that the information provided by n of the $n+1$ observations is lost in estimating $\mu_1, \mu_2, \ldots, \mu_n$.

EXAMPLE 10.8.3. Consider $2n$ independent measurements X_i, Y_i for $i = 1, 2, \ldots, n$, where $X_i \sim N(\mu_i, \sigma^2)$ and $Y_i \sim N(\mu_i, \sigma^2)$. Altogether there are $n+1$ unknown parameters $\mu_1, \mu_2, \ldots, \mu_n$, and σ to be estimated from $2n$ observations.

It can be shown that the log likelihood function based on all $2n$ observations is

$$l(\mu_1, \mu_2, \ldots, \mu_n, \sigma) = -2n \log \sigma - \frac{1}{2\sigma^2}\Sigma[(x_i - \mu_i)^2 + (y_i - \mu_i)^2].$$

Upon setting the derivatives of l equal to zero and solving, we find that

$$\hat{\mu}_i = (x_i + y_i)/2 \qquad \text{for } i = 1, 2, \ldots, n;$$

$$\hat{\sigma}^2 = \frac{1}{2n}\Sigma[(x_i - \hat{\mu}_i)^2 + (y_i - \hat{\mu}_i)^2]$$

$$= \frac{1}{4n}\Sigma(x_i - y_i)^2.$$

It can be shown that, if n is large, then $\hat{\sigma}^2$ will be close to $\sigma^2/2$ with probability near 1. Both $\hat{\sigma}^2$ and maximum likelihood intervals will seriously underestimate σ^2.

To see the similarity with the situation in the preceding example, we consider the one-to-one transformation from (X_i, Y_i) to (S_i, T_i), where

$$S_i \equiv (X_i - Y_i)/2 \sim N(0, \sigma^2/2)$$

$$T_i \equiv (X_i + Y_i)/2 \sim N(\mu_i, \sigma^2/2).$$

All of the information provided by $T_1, T_2, \ldots, T_n$ is lost in estimating $\mu_1, \mu_2, \ldots, \mu_n$. There are effectively only n observations $S_1, S_2, \ldots, S_n$ which provide information about σ^2 when the μ_i's are unknown. The log likelihood function of σ based on the marginal distribution of the S_i's is

$$l(\sigma) = -n \log \sigma - \frac{1}{\sigma^2}\Sigma s_i^2$$

which is maximized for

$$\hat{\sigma}^2 = \frac{2}{n}\Sigma s_i^2 = \frac{1}{2n}\Sigma(x_i - y_i)^2.$$

The original analysis underestimated σ^2 by 50% because it did not allow for the fact that half of the information about σ^2 is lost in estimating $\mu_1, \mu_2, \ldots, \mu_n$.

Marginal and Conditional Likelihoods

The preceding three examples show some of the difficulties which can arise when many parameters are to be estimated from a relatively small amount of data.

Sometimes these difficulties can be overcome by carefully analyzing the situation and discarding all but the relevant part of the data. It may be possible to identify a marginal (or conditional) distribution which carries all of the relevant information concerning the parameter of interest. The likelihood function derived from this distribution is then called a marginal (or conditional) likelihood function.

In Example 10.8.2 we argued that $Y_1, Y_2, \ldots, Y_n$ do not provide any information about σ^2 when $\mu_1, \mu_2, \ldots, \mu_n$ are unknown. The marginal log likelihood function of σ based on the distribution of Y_0 is

$$l(\sigma) = -\log \sigma - \frac{1}{2\sigma^2}(y_0 - \mu_0)^2 \qquad \text{for } \sigma > 0.$$

Note that, since Y_0 is independent of $Y_1, Y_2, \ldots, Y_n$, the conditional distribution of Y_0 given $Y_1, Y_2, \ldots, Y_n$ is the same as the marginal distribution. Hence $l(\sigma)$ can also be regarded as a conditional log likelihood function.

Similarly, in Example 10.8.3 the marginal log likelihood function of σ based on $S_1, S_2, \ldots, S_n$ is

$$l(\sigma) = -n \log \sigma - \frac{1}{\sigma^2}\Sigma s_i^2 \qquad \text{for } \sigma > 0.$$

Since the S_i's are distributed independently of the T_i's, this can also be interpreted as the conditional log likelihood function of σ given $T_1, T_2, \ldots, T_n$.

In both examples, satisfactory results are obtained if we apply likelihood methods to the appropriate marginal or conditional likelihood function of σ.

CHAPTER 11

Frequency Properties

The purpose of this chapter is to investigate some theoretical properties of the estimation procedures described in Chapters 9 and 10.

It is customary to evaluate and compare statistical procedures by examining how they would behave in a series of hypothetical repetitions of the experiment. One imagines that the experiment which gave rise to the data is to be repeated over and over again under identical conditions. One then determines how frequently the statistical procedure would give correct results in this series of repetitions.

Quantities such as $\hat{\theta}$ and the endpoints of the 10% likelihood interval would vary from one repetition of the experiment to another. Their probability distributions in repetitions of the experiment are called their sampling distributions. Some exact sampling distributions are derived in Section 1, and these are used to calculate coverage probabilities in Section 2. These sections also introduce the likelihood ratio statistic, which will be used extensively in Chapter 12. A chi-square approximation to the distribution of the likelihood ratio statistic is investigated in Section 3.

Most Statistics textbooks advocate the use of confidence intervals, and these are defined in Section 4. It is shown that, under some mild conditions, likelihood intervals are confidence intervals or approximate confidence intervals. It is suggested that the best way to construct confidence intervals is by calculating likelihood intervals of the appropriate size. Intervals constructed in this way will give valid information summaries in particular applications, as well as having the desired frequency properties in a series of imaginary repetitions.

Section 5 gives results on coverage probabilities and confidence regions for the two-parameter case. Section 6 defines the expected information function and illustrates its use in planning experiments. Finally, Section 7 gives a brief discussion of bias in parameter estimation.

11.1. Sampling Distributions

Likelihood methods for parameter estimation were described in Chapters 9 and 10. Usually one starts with a set of data from an experiment. One then attempts to formulate a probability model which describes how the data were generated. If this model involves an unknown parameter θ, the probability of the data is found as a function of θ. From this one gets the likelihood function, MLE, log relative likelihood function, and likelihood intervals for θ.

The purpose of this chapter is to investigate how the log RLF and related quantities, such as $\hat{\theta}$ and likelihood intervals, would behave in a hypothetical series of repetitions of the experiment. Thus we imagine that the experiment is to be repeated over and over again under identical conditions. We assume that θ has the same value θ_0 (called the true value of θ) in all repetitions.

If the experiment were repeated, we would most likely get a different set of data. The likelihood function depends on the data, and so repeating the experiment would likely produce a different $r(\theta)$. The MLE and endpoints of the 10% likelihood interval would vary from one repetition to the next, and can be modeled as random variables. Their probability distributions in repetitions of the experiment with θ fixed are called their *sampling distributions*. In principle, it is possible to derive exact sampling distributions from the probability model. However, this is often very difficult to do, and approximations will be required (see Section 11.3).

Various other quantities associated with the log RLF can be studied. The most important of these for statistical applications is the *likelihood ratio statistic*,

$$D \equiv -2r(\theta_0). \tag{11.1.1}$$

It is more convenient to work with D rather than $r(\theta_0)$ because D is non-negative whereas $r(\theta_0)$ is non-positive. The factor 2 is included in (11.1.1) because it slightly simplifies matters in normal distribution examples.

In repetitions of the experiment, both $r(\theta_0)$ and D would vary according to the data obtained. D would be small whenever the data were such that the true value θ_0 was a likely value of θ, and D would be large whenever the data were such that θ_0 was unlikely. We can consider D as a random variable, and its sampling distribution can be derived from the model.

Likelihood ratio statistics will be used extensively in the next chapter, and a more general definition is given there. In the language of Section 12.2, D is the likelihood ratio statistic for testing the hypothesis $\theta = \theta_0$.

Note that the repetitions to which we keep referring are purely imaginary. Real experiments do not get repeated over and over under identical conditions. The series of repetitions is invented by the statistician to give a theoretical framework within which statistical procedures may be investigated and compared. It is often possible to imagine many different ways in which the experiment could be repeated, and the answers obtained will depend to some extent on the series of repetitions considered. We shall return to this troublesome point in Chapter 15.

EXAMPLE 11.1.1. Suppose that $n = 10$ people are chosen at random from a large homogeneous population and are tested for tuberculosis. The aim is to estimate θ, the proportion of people with TB in the population. We showed in Example 9.1.1 that if x of the 10 tested have TB, the log likelihood function of θ is

$$l(\theta) = x \log \theta + (10 - x) \log (1 - \theta) \qquad \text{for } 0 < \theta < 1.$$

The MLE is $\hat{\theta} = x/10$, and the log RLF is $r(\theta) = l(\theta) - l(\hat{\theta})$.

For instance, if 3 out of 10 are diseased, we have $\hat{\theta} = 0.3$, $l(\hat{\theta}) = -6.109$, and

$$r(\theta) = 3 \log \theta + 7 \log (1 - \theta) + 6.109 \qquad \text{for } 0 < \theta < 1.$$

This function is plotted with a solid line in Figure 11.1.1. We find that $r(\theta) \geq -2.30$ for $0.072 \leq \theta \leq 0.635$, and this is the 10% LI for θ based on the observation $x = 3$.

Now imagine that the experiment is to be repeated over and over, with θ fixed at a particular value θ_0. Then X, the number with TB in the sample, is a random variable with probability function

$$f(x) = \binom{10}{x} \theta_0^x (1 - \theta_0)^{10-x} \qquad \text{for } x = 0, 1, \ldots, 10.$$

Since the observed X would vary from one repetition to another, so would $l(\theta)$, $\hat{\theta}$ and $r(\theta)$. For instance, if we observed $X = 2$, the log RLF would be

$$r(\theta) = 2 \log \theta + 8 \log \theta + 5.004,$$

and if we observed $x = 0$, the log RLF would be

$$r(\theta) = 10 \log (1 - \theta).$$

Altogether, there are 11 possible log RLF's corresponding to the 11 possible values of X. Figure 11.1.1 shows six of these, corresponding to $X = 0, 1, \ldots, 5$.

Various features of the log RLF can be studied. First consider $\hat{\theta}$, the

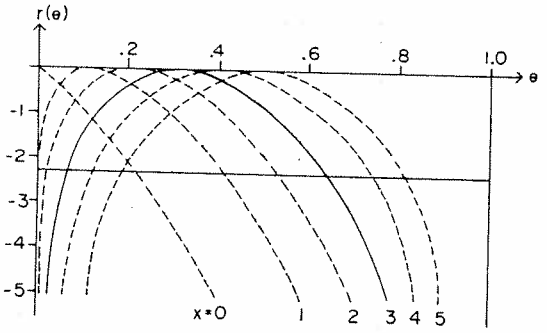

Figure 11.1.1. Six of the eleven possible log relative likelihood functions in a binomial example.

location of the maximum of $r(\theta)$. In repetitions of the experiment, $\hat{\theta}$ is a random variable with 11 possible values (see the second column of Table 11.1.1). The probabilities of these values can be found from the binomial distribution of X, and they depend on the true value θ_0. The last three columns of Table 11.1.1 give these probabilities for $\theta_0 = 0.1$, for $\theta_0 = 0.2$, and for $\theta_0 = 0.3$.

Similarly, the left and right endpoints of the 10% LI are now regarded as random variables. There are 11 possible intervals $[A, B]$ as shown in Table 11.1.1, and their probabilities are given in the last three columns of the table for $\theta_0 = 0.1$, for $\theta_0 = 0.2$, and for $\theta_0 = 0.3$. Using the tabulated values, one can investigate how the 10% LI would behave in repetitions of the experiment with θ fixed (see the next section).

Finally, consider the likelihood ratio statistic

$$D \equiv -2r(\theta_0) \equiv -2[l(\theta_0) - l(\hat{\theta})].$$

For any particular θ_0, there are 11 possible values for D corresponding to the 11 possible values of X. The probabilities of these values can be found from the binomial distribution of X. Table 11.1.2 gives the sampling distribution of D (i.e. its possible values and their probabilities) for $\theta_0 = 0.1$, for $\theta_0 = 0.2$, and for $\theta_0 = 0.3$.

The value of D which corresponds to a relative likelihood of 10% is $-2 \log 0.1 = 4.61$. From Table 11.1.2 we find that

$$P(D \leq 4.61) = P(X \leq 3) = 0.987 \qquad \text{for } \theta_0 = 0.1;$$

$$P(D \leq 4.61) = P(X \leq 5) = 0.992 \qquad \text{for } \theta_0 = 0.2.$$

Similarly, for $\theta_0 = 0.3$ we have

$$P(D \leq 4.61) = P(1 \leq X \leq 6) = 0.961.$$

Table 11.1.1. Sampling Distributions of the MLE and the 10% LI $[A, B]$ in a Binomial Example

				Probability		
X	$\hat{\theta}$	A	B	$\theta_0 = 0.1$	$\theta_0 = 0.2$	$\theta_0 = 0.3$
0	0.0	0.000	0.206	0.349	0.107	0.028
1	0.1	0.004	0.403	0.387	0.268	0.121
2	0.2	0.029	0.530	0.194	0.302	0.233
3	0.3	0.072	0.635	0.057	0.201	0.267
4	0.4	0.128	0.725	0.011	0.088	0.200
5	0.5	0.196	0.804	0.001	0.026	0.103
6	0.6	0.275	0.872	0.000	0.006	0.037
7	0.7	0.365	0.928	0.000	0.001	0.009
8	0.8	0.470	0.971	0.000	0.000	0.001
9	0.9	0.597	0.996	0.000	0.000	0.000
10	1.0	0.794	1.000	0.000	0.000	0.000

Table 11.1.2. Sampling Distribution of the Likelihood Ratio Statistic $D \equiv -2r(\theta_0)$ in a Binomial Example

	$\theta_0 = 0.1$		$\theta_0 = 0.2$		$\theta_0 = 0.3$	
x	$D(x)$	$f(x)$	$D(x)$	$f(x)$	$D(x)$	$f(x)$
0	2.11	0.349	4.46	0.107	7.13	0.028
1	0.00	0.387	0.73	0.268	2.33	0.121
2	0.89	0.194	0.00	0.302	0.51	0.233
3	3.07	0.057	0.56	0.201	0.00	0.267
4	6.22	0.011	2.09	0.088	0.45	0.200
5	10.22	0.001	4.46	0.026	1.74	0.103
6	15.01	0.000	7.64	0.006	3.84	0.037
7	20.65	0.000	11.65	0.001	6.78	0.009
8	27.25	0.000	16.64	0.000	10.68	0.001
9	35.16	0.000	22.91	0.000	15.88	0.000
10	46.05	0.000	32.19	0.000	24.08	0.000

In all three cases, values of D greater than 4.61 would occur very rarely in repetitions of the experiment. The true value θ_0 would almost always have a relative likelihood of 10% or more, and therefore a D-value of 4.61 or less.

EXAMPLE 11.1.2. Suppose that an experiment involves taking a single measurement x which is modeled as the observed value of a random variable $X \sim N(\theta, 1)$. If the measurement interval is small, the log likelihood function of θ is

$$l(\theta) = -\tfrac{1}{2}(x-\theta)^2 \qquad \text{for } -\infty < \theta < \infty,$$

from which we obtain $\hat{\theta} = x$ and $l(\hat{\theta}) = 0$. The log RLF is

$$r(\theta) = -\tfrac{1}{2}(x-\theta)^2 \qquad \text{for } -\infty < \theta < \infty.$$

Upon solving $r(\theta) \geq \log p$, we find that the $100p\%$ likelihood interval for θ is given by

$$x - c \leq \theta \leq x + c$$

where $c = \sqrt{-2 \log p}$.

Now imagine that the experiment is to be repeated over and over again with θ fixed at a particular value θ_0. The probability distribution of X in repetitions with $\theta = \theta_0$ is $N(\theta_0, 1)$. Since the observed value of X would vary from one repetition to the next, so would $r(\theta)$, $\hat{\theta}$, and the endpoints of likelihood intervals. We can now think of $\hat{\theta}$ as a random variable, $\hat{\theta} \equiv X$, with sampling distribution

$$\hat{\theta} \sim N(\theta_0, 1).$$

Similarly, the endpoints of the $100p\%$ LI are random variables $A \equiv X - c$ and $B \equiv X + c$, with sampling distributions

$$A \sim N(\theta_0 - c, 1); \qquad B \sim N(\theta_0 + c, 1).$$

The likelihood ratio statistic is

$$D \equiv -2r(\theta_0) \equiv (X - \theta_0)^2,$$

which would also vary from one repetition of the experiment to another. To find the sampling distribution of D, we note that, by (6.6.5),

$$Z \equiv X - \theta_0 \sim N(0, 1).$$

Thus, by (6.9.8), we have

$$D \equiv Z^2 \sim \chi^2_{(1)}.$$

In this example the sampling distribution of the likelihood ratio statistic is $\chi^2_{(1)}$ for all possible parameter values θ_0. The situation is much simpler than in the preceding example, where the sampling distribution of D depended on the true value θ_0.

To calculate probabilities for D, we can use either Table B4 for the chi-square distribution, or Table B2 for $N(0, 1)$ (see Appendix B). For instance,

$$P(D \leq 4.61) = P(\chi^2_{(1)} \leq 4.61) \approx 0.97$$

by interpolating in Table B4. For greater accuracy, we note that

$$P(D \leq 4.61) = P(Z^2 \leq 4.61) = P(|Z| \leq 2.146) = 0.968$$

from Table B2. This result is true for all θ_0, whereas in the preceding example we found that $P(D \leq 4.61)$ depended on the true value θ_0.

EXAMPLE 11.1.3. Suppose that we observe n measurements $x_1, x_2, \ldots, x_n$ which we model as observed values of independent $N(\mu, \sigma^2)$ variates. As in Example 9.7.1 we assume that σ is known and that the measurement intervals are small. Then the MLE of μ is $\hat{\mu} = \bar{x} = \frac{1}{n}\Sigma x_i$, the log RLF is

$$r(\mu) = -\frac{n}{2\sigma^2}(\bar{x} - \mu)^2 \qquad \text{for } -\infty < \mu < \infty,$$

and the $100p\%$ likelihood interval is

$$\bar{x} - \frac{c\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{c\sigma}{\sqrt{n}}$$

where $c = \sqrt{-2 \log p}$.

Now imagine that the experiment is to be repeated over and over with $u = \mu_0$. The observed x_i's would vary from one repetition to the next, and thus so would $\bar{x}$. By (6.6.8) we have $\bar{X} \sim N(\mu_0, \sigma^2/n)$, and therefore

$$\hat{\mu} \equiv \bar{X} \sim N(\mu_0, \sigma^2/n).$$

This is the sampling distribution of the MLE in repetitions of the experiment with $\mu = \mu_0$. Similarly, the endpoints of the 100p% LI are now regarded as random variables $\bar{X} - c\sigma/\sqrt{n}$ and $\bar{X} + c\sigma/\sqrt{n}$.

By (11.1.1), the likelihood ratio statistic is

$$D \equiv -2r(\mu_0) \equiv \frac{n}{\sigma^2}(\bar{X} - \mu_0)^2 \equiv Z^2$$

where Z is the standard form of $\bar{X}$:

$$Z \equiv (\bar{X} - \mu_0)/\sqrt{\sigma^2/n} \sim N(0, 1).$$

It follows by (6.9.8) that for all μ_0 the sampling distribution of D is χ^2 with one degree of freedom. The same result was obtained in the preceding example, which is the special case $n = \sigma = 1$.

Note on Notation

In previous chapters we have used capital letters for random variables and the corresponding small letters for their possible values. We shall no longer follow this convention in all cases. In particular, we shall use $\hat{\theta}$ to represent the MLE of θ whether we are thinking of it as a random variable in repetitions of the experiment, or as the particular value computed from the data. Also, we shall use $r(\theta_0)$ for the log relative likelihood of θ_0 whether we are considering it to be fixed or random, because we are already using R to mean e^r.

11.2. Coverage Probability

Suppose that the probability model for an experiment involves a single unknown parameter θ. As in Section 11.1 we imagine a series of repetitions of the experiment with θ fixed at θ_0. Further, imagine that an interval $[A, B]$ is to be computed from the data in the same way at each repetition. For instance, $[A, B]$ could be the 10% likelihood interval for θ.

Owing to random variability in the data, the interval $[A, B]$ would vary from one repetition to the next. Its endpoints A, B can be modeled as random variables. The sampling distributions of A and B can be derived from the probability model, and they will generally depend upon θ_0.

Since the interval $[A, B]$ would vary from one repetition to the next, it might sometimes fail to include the true value θ_0. Hopefully this would happen only rarely, and the interval $[A, B]$ would usually contain, or cover, the true value θ_0.

The *coverage probability* of the random interval $[A, B]$ is the probability that the interval $[A, B]$ includes, or covers, the true parameter value θ_0:

$$\mathrm{CP}(\theta_0) = P(A \le \theta_0 \le B | \theta = \theta_0). \qquad (11.2.1)$$

This is not a conditional probability. Rather, the notation is meant to emphasize that the true parameter value θ_0 is to be used in computing the coverage probability.

The coverage probability $\mathrm{CP}(\theta_0)$ is the fraction of the time that the interval $[A, B]$ would include the true value θ_0 in a large number of repetitions of the experiment with $\theta = \theta_0$. Note that A and B are the random variables in (11.2.1) and θ_0 is fixed.

EXAMPLE 11.2.1. Imagine that the experiment described in Example 11.1.1 is to be repeated over and over again with θ fixed at θ_0. Each time the 10% likelihood interval for θ is to be calculated. We want to know what fraction of the time this interval would contain the true value θ_0.

First consider repetitions with $\theta = 0.1$. The 11 possible 10% likelihood intervals are listed in Table 11.1.1, and from the table we see that $A \le 0.1 \le B$ whenever $0 \le X \le 3$. Thus the coverage probability of the 10% LI in repetitions with $\theta = 0.1$ is

$$\mathrm{CP}(0.1) = P(A \le 0.1 \le B | \theta = 0.1) = P(0 \le X \le 3 | \theta = 0.1).$$

Now from the 5th column of Table 11.1.1 we get

$$\mathrm{CP}(0.1) = 0.349 + 0.387 + 0.194 + 0.057 = 0.987.$$

The coverage probability in repetitions with $\theta = 0.2$ is

$$\mathrm{CP}(0.2) = P(A \le 0.2 \le B | \theta = 0.2) = P(0 \le X \le 5 | \theta = 0.2).$$

Using the 6th column of Table 11.1.1 we get

$$\mathrm{CP}(0.2) = 0.107 + 0.268 + \cdots + 0.026 = 0.992.$$

Similarly, we find that

$$\mathrm{CP}(0.3) = P(A \le 0.3 \le B | \theta = 0.3) = 0.961.$$

In repetitions of the experiment with $\theta_0 = 0.1$, the 10% LI for θ would include, or cover, the true value 98.7% of the time. Similarly, the 10% LI would cover the true value θ_0 in 99.2% of repetitions with $\theta_0 = 0.2$, and in 96.1% of repetitions with $\theta_0 = 0.3$.

In this example, the coverage probability depends on the true value θ_0. It can be shown that $\mathrm{CP}(\theta_0)$ varies from a low of 89.3% when θ_0 is 0.206 or 0.794 to a high approaching 100% as θ_0 tends to 0 or 1. Owing to the discreteness of the binomial distribution, $\mathrm{CP}(\theta_0)$ is not a continuous function. For instance, we see from Table 11.1.1 that the upper endpoint of the 10% LI corresponding to $X = 0$ is 0.206. In computing $\mathrm{CP}(\theta_0)$, we would include $P(X = 0)$ for $\theta_0 \le 0.206$ but not for $\theta_0 > 0.206$. As θ_0 increases through 0.206, the coverage probability suddenly decreases by

$$P(X = 0 | \theta_0 = 0.206) = (0.794)^{10} = 0.100.$$

Similarly, $\mathrm{CP}(\theta_0)$ will have a discontinuity at each of the other endpoints of the 11 possible likelihood intervals.

EXAMPLE 11.2.2. In Example 11.1.2, the $100p\%$ likelihood interval for θ has the form

$$X - c \le \theta \le X + c$$

where $c = \sqrt{-2 \log p}$. By (11.2.1), the coverage probability of this interval in repetitions of the experiment with $\theta = \theta_0$ is

$$\begin{aligned} \mathrm{CP}(\theta_0) &= P(X - c \le \theta_0 \le X + c | \theta = \theta_0) \\ &= P(-c \le X - \theta_0 \le c | \theta = \theta_0). \end{aligned}$$

The distribution of X is $N(\theta_0, 1)$, so (6.6.5) gives

$$Z \equiv X - \theta_0 \sim N(0, 1).$$

It follows that

$$\mathrm{CP}(\theta_0) = P(-c \le Z \le c)$$

where $Z \sim N(0, 1)$ and $c = \sqrt{-2 \log p}$. For any given p, the coverage probability can be found from Table B1 or B2, and it does not depend on θ_0.

If $p = 0.1$, then $c = \sqrt{-2 \log 0.1} = 2.146$, and so the coverage probability of the 10% LI is

$$\mathrm{CP} = P(-2.146 \le Z \le 2.146) = 0.968.$$

The 10% LI is $X \pm 2.146$, and it would include the true value of θ in 96.8% of repetitions with θ fixed. Similarly, it can be shown that the coverage probabilities of the 50% and 1% likelihood intervals are 0.761 and 0.9976, respectively (see Table 11.2.1).

From Table B1 we note that

$$P(-1.96 \le Z \le 1.96) = 0.95.$$

Thus to obtain 95% coverage probability we require $c = 1.96$. Now solving $c = \sqrt{-2 \log p}$ for p gives

$$p = e^{-c^2/2} = 0.147.$$

The 14.7% likelihood interval for θ is $X \pm 1.96$. It would cover the true value of θ in 95% of repetitions of the experiment with θ fixed. Similarly, we can show that the 25.8% and 3.6% likelihood intervals have coverage probabilities 0.9 and 0.99, respectively.

Table 11.2.1. Coverage Probabilities of $100p\%$ Likelihood Intervals in Examples 11.2.2 and 11.2.3

p	0.5	0.1	0.01	0.258	0.147	0.036
CP	0.761	0.968	0.9976	0.9	0.95	0.99

EXAMPLE 11.2.3. In Example 11.1.3 the $100p\%$ likelihood interval for μ has the form

$$\bar{X} - \frac{c\sigma}{\sqrt{n}} \le \mu \le \bar{X} + \frac{c\sigma}{\sqrt{n}}$$

where $c = \sqrt{-2 \log p}$. The coverage probability of this interval in repetitions of the experiment with $\mu = \mu_0$ is

$$\begin{aligned} \mathrm{CP}(\mu_0) &= P\left(\bar{X} - \frac{c\sigma}{\sqrt{n}} \le \mu_0 \le \bar{X} + \frac{c\sigma}{\sqrt{n}} | \mu = \mu_0 \right) \\ &= P\left(-c \le \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \le c | \mu = \mu_0 \right). \end{aligned}$$

The probability distribution of $\bar{X}$ when $\mu = \mu_0$ is $N(\mu_0, \sigma^2/n)$, and therefore

$$Z \equiv \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0, 1).$$

Thus we have

$$\mathrm{CP}(\mu_0) = P(-c \le Z \le c)$$

where $Z \sim N(0, 1)$ and $c = \sqrt{-2 \log p}$. For every p, the coverage probability of the $100p\%$ likelihood interval is the same as in the preceding example.

Use of the Likelihood Ratio Statistic

In Example 11.2.1 we used a list of all possible 10% likelihood intervals in calculating the coverage probability. Essentially the same method of calculation was used in Examples 11.2.2 and 11.2.3, except that there we were able to use formulas for the interval endpoints.

It is not really necessary to determine all possible likelihood intervals, either numerically or by formula, in order to determine their coverage probability. Instead, one can find coverage probabilities of likelihood intervals from the sampling distribution of the likelihood ratio statistic.

The $100p\%$ likelihood interval (or region) for θ is the set of parameter values such that $R(\theta) \ge p$, or equivalently, $r(\theta) \ge \log p$. A particular parameter value θ_0 belongs to the $100p\%$ LI for θ if and only if $r(\theta_0) \ge \log p$. Since $D \equiv -2r(\theta_0)$, it follows that θ_0 belongs to the $100p\%$ LI if and only if $D \le -2 \log p$. Therefore, the coverage probability of the $100p\%$ likelihood interval is

$$\begin{aligned} \mathrm{CP}(\theta_0) &= P(\theta_0 \text{ belongs to } 100p\% \text{ LI} | \theta = \theta_0) \\ &= P(D \le -2 \log p | \theta = \theta_0). \end{aligned} \tag{11.2.2}$$

EXAMPLE 11.2.1 (continued). Since $-2 \log 0.1 = 4.61$, the coverage probability of the 10% likelihood interval is

$$\mathrm{CP}(\theta_0) = P(D \le 4.61 | \theta = \theta_0).$$

Columns 2 and 3 of Table 11.1.2 give the sampling distribution of D for $\theta = 0.1$, and we see that $D \le 4.61$ whenever $0 \le X \le 3$. It follows that

$$\mathrm{CP}(0.1) = P(0 \le X \le 3 | \theta = 0.1) = 0.987$$

as before. Similarly, when $\theta = 0.2$ we have $D \le 4.61$ for $0 \le X \le 5$, and so

$$\mathrm{CP}(0.2) = P(0 \le X \le 5 | \theta = 0.2) = 0.992.$$

Note that (11.2.2) permits us to find coverage probabilities without having to calculate all possible intervals. For instance, since $-2 \log 0.5 = 1.386$, the coverage probability of the 50% likelihood interval is

$$\mathrm{CP}(\theta_0) = P(D \le 1.386 | \theta = \theta_0).$$

This can be found from Table 11.1.2 for $\theta_0 = 0.1$, 0.2, and 0.3. It is not necessary to calculate the 11 possible 50% likelihood intervals in order to evaluate the coverage probability.

EXAMPLES 11.2.2 and 11.2.3 (continued). We showed in Section 11.1 that for both of these examples, the sampling distribution of D is $\chi^2_{(1)}$ for all parameter values. Thus, in both examples, the coverage probability of the $100p\%$ likelihood interval is

$$\mathrm{CP} = P(\chi^2_{(1)} \le -2 \log p).$$

It follows from (6.9.8) that, for all $d > 0$,

$$P(\chi^2_{(1)} \le d) = P(Z^2 \le d) = P(-\sqrt{d} \le Z \le \sqrt{d})$$

where $Z \sim N(0, 1)$. Thus we have

$$\mathrm{CP} = P(-c \le Z \le c)$$

where $Z \sim N(0, 1)$ and $c = \sqrt{-2 \log p}$. This is the same result that we obtained previously by working with formulas for the likelihood intervals.

PROBLEMS FOR SECTION 11.2

1. In Example 11.2.1, find the coverage probability of the 20% likelihood interval for θ in repetitions of the experiment with $\theta = 0.1$. Repeat for $\theta = 0.2$ and for $\theta = 0.3$.
2. In Example 11.2.1, show that the $100p\%$ likelihood interval for θ has the same coverage probability in repetitions with $\theta = 1 - \theta_0$ as in repetitions with $\theta = \theta_0$.
3. † Consider the situation described in Example 11.2.2.
   (a) Find the coverage probability of the 20% likelihood interval for μ in repetitions of the experiment with μ fixed.

   (b) Determine p so that the $100p\%$ likelihood interval for θ will have coverage probability 0.975.
4. A deck of 1000 cards contains a card from each of 901 denominations 0, 1, ..., 900 and 99 extra cards from some unknown denomination θ. The possible values of θ are 0, 1, ..., 900. One card is selected at random from the deck and its denomination X is observed.
   (a) Show that $\hat{\theta} = x$ and

$$R(\theta) = \begin{cases} 1 & \text{for } \theta = x; \\ 0.01 & \text{for } \theta \ne x. \end{cases}$$

   (b) Show that the likelihood ratio statistic $D = -2r(\theta_0)$ has just two possible values, 0 and 4 log 10, with probabilities 0.1 and 0.9, respectively.
   (c) Show that, for $p > 0.01$, the $100p\%$ likelihood interval for θ consists of a single value, and its coverage probability is 0.10.

11.3. Chi-Square Approximations

We noted in Section 11.2 that coverage probabilities of likelihood intervals can be found from the sampling distribution of the likelihood ratio statistic $D \equiv -2r(\theta_0)$. The exact distribution of D was derived under binomial and normal models in the examples of Section 11.1. We found that the distribution of D depends on the true value θ_0 in the binomial example. However, in the normal examples, the distribution of D is $\chi^2_{(1)}$ for all θ_0.

It is usually difficult to derive the exact sampling distribution of D. Fortunately a good approximation is available. It turns out that, in many situations, the distribution of D is remarkably close to $\chi^2_{(1)}$:

$$D \equiv -2r(\theta_0) \approx \chi^2_{(1)}. \tag{11.3.1}$$

The results in Table 11.2.1, which are exact for the normal distribution examples, will provide good approximations in many situations with non-normal distributions. Coverage probabilities of 25.8%, 14.7%, and 3.6% likelihood intervals will usually be close to 0.9, 0.95, and 0.99, respectively.

The Central Limit Theorem can be used to establish (11.3.1) when the data are observed values of n IID random variables $X_1, X_2, \ldots, X_n$. It can be shown that, under some mild regularity conditions,

$$\lim_{n \to \infty} P(D \le d | \theta = \theta_0) = P(\chi^2_{(1)} \le d) \tag{11.3.2}$$

for all $d > 0$. A brief justification of this result is given at the end of the section.

The most important condition required in proving (11.3.2) is that the range of the X_i's must not depend upon θ. Also, it is assumed that θ_0 is an interior point of the parameter space. The χ^2 approximation need not hold if θ_0 is on the boundary of the parameter space. If θ_0 is close to the boundary, a very large value of n may be needed before the χ^2 distribution provides a good

approximation. The limiting distribution of D is $\chi^2_{(1)}$ for all interior parameter values θ_0, but the sample size needed to achieve reasonable accuracy may depend upon θ_0.

In the following three examples, the accuracy of (11.3.1) is investigated in situations where the exact sampling distribution of D can be derived fairly easily. Using the exact sampling distribution, we shall calculate

$$P(D \leq 2.706), \qquad P(D \leq 3.841), \qquad P(D \leq 6.635).$$

The values 2.706, 3.841, and 6.635 were chosen because they are the 90%, 95%, and 99% points of $\chi^2_{(1)}$ (see Table B4). Thus, if (11.3.1) holds, the three probabilities should be close to 0.9, 0.95, and 0.99.

By (11.2.2), $P(D \leq d)$ is the coverage probability of the $100p\%$ likelihood interval where $d = -2 \log p$; that is, $p = e^{-d/2}$. Since $e^{-2.706/2} = 0.258$, $P(D \leq 2.706)$ is the coverage probability of the 25.8% likelihood interval. Similarly, $P(D \leq 3.841)$ and $P(D \leq 6.635)$ are coverage probabilities of 14.7% and 3.6% likelihood intervals. The three examples provide a comparison of the exact coverage probabilities of these three likelihood intervals with their approximate coverage probabilities from (11.3.1).

EXAMPLE 11.3.1. Suppose that n people are tested for tuberculosis as in Example 11.1.1. In that example we took $n = 10$ and derived the exact sampling distribution of D for $\theta_0 = 0.1$, for $\theta_0 = 0.2$, and for $\theta_0 = 0.3$.

The exact distribution of D when $\theta_0 = 0.1$ is given in columns 2 and 3 of Table 11.1.2. We see that $D \leq 2.706$ for $X \leq 2$, and therefore

$$P(D \leq 2.706) = P(X \leq 2) = 0.349 + 0.387 + 0.194 = 0.930.$$

Similarly, we have

$$P(D \leq 3.841) = P(X \leq 3) = 0.987;$$

$$P(D \leq 6.635) = P(X \leq 4) = 0.998.$$

These results are shown in the first row of Table 11.3.1. The values in the next two rows of this table are obtained in a similar fashion from the last four columns of Table 11.1.2. The last row of Table 11.3.1 gives the approximate probabilities according to (11.3.1). The other rows are obtained by redoing the calculations of Example 11.1.1 with $n = 20$, and then repeating them again with $n = 50$.

In this example, D is a discrete random variable having $n + 1$ possible values, one for each value of X. The approximating χ^2 distribution is continuous. For this reason, one would not expect the approximation (11.3.1) to be very accurate when n is as small as 10. As n increases, so does the number of possible values for D. When n is very large, the discreteness no longer matters, and the distribution of D will be well approximated by $\chi^2_{(1)}$.

The limiting distribution of D is $\chi^2_{(1)}$ for all θ_0 such that $0 < \theta_0 < 1$. If θ_0 is near $\frac{1}{2}$, (11.3.1) gives fairly accurate results for $n = 20$. However, a much larger n is needed if θ_0 is close to 0 or 1.

Table 11.3.1. Exact and Approximate Probabilities for the Likelihood Ratio Statistic in a Binomial Example

n	θ_0	$P(D \leq 2.706)$	$P(D \leq 3.841)$	$P(D \leq 6.635)$
10	0.1	0.930	0.987	0.998
	0.2	0.859	0.859	0.992
	0.3	0.924	0.961	0.961
20	0.1	0.835	0.867	0.998
	0.2	0.899	0.956	0.986
	0.3	0.917	0.947	0.987
50	0.1	0.908	0.942	0.992
	0.2	0.891	0.951	0.988
	0.3	0.912	0.957	0.987
χ^2 approx.		0.9	0.95	0.99

EXAMPLE 11.3.2. Suppose that an experiment yields n counts which are modeled as observed values of independent Poisson-distributed variates $X_1, X_2, \ldots, X_n$ with expected value μ. From Example 9.1.2, the log likelihood function of μ is

$$l(\mu) = t \log \mu - n\mu \qquad \text{for } \mu > 0$$

where $t = \Sigma x_i$. The MLE is $\hat{\mu} = t/n$, and the log RLF is

$$r(\mu) = l(\mu) - l(\hat{\mu}) = t \log \frac{\mu}{\hat{\mu}} - n\mu + n\hat{\mu}$$

$$= -t \log \frac{t}{n\mu} - n\mu + t.$$

Now imagine a series of repetitions of the experiment with μ fixed at μ_0. The total count $T \equiv \Sigma X_i$ would vary from one repetition to the next, and is modeled as a random variable. By the corollary to Example 4.6.1, the probability distribution of T is Poisson with mean $m = n\mu_0$.

The likelihood ratio statistic is

$$D \equiv -2r(\mu_0) \equiv 2\left[T \log \frac{T}{m} + m - T\right].$$

D is a discrete variate with one possible value for each value of T. For any given m we can substitute $T = 0, 1, 2, \ldots$ to obtain the possible values of D. Their probabilities are obtained from the Poisson distribution

$$P(T = t) = m^t e^{-m}/t! \qquad \text{for } t = 0, 1, 2, \ldots .$$

For instance, suppose that $m = 10$. Then

$$D \equiv 2\left[T \log \frac{T}{10} + 10 - T\right],$$

which can be calculated for $T = 0, 1, 2, \ldots$. (The term $T \log \frac{T}{10}$ is taken to be 0 when $T = 0$.) From these calculated values, we find that $D \leq 2.706$ for $6 \leq T \leq 15$. Since T has a Poisson distribution with $m = 10$, it follows that

$$P(D \leq 2.706) = \sum_{t=6}^{15} 10^t e^{-10}/t! = 0.884.$$

Similarly, we obtain

$$P(D \leq 3.841) = P(5 \leq T \leq 16) = 0.944;$$

$$P(D \leq 6.635) = P(4 \leq T \leq 19) = 0.986.$$

These probabilities are recorded in the second row of Table 11.3.2, and the last row gives the approximate probabilities from (11.3.1). The other rows of the table are found by repeating the calculations with $m = 5$, 20, and 40. The table suggests that the χ^2 approximation is reasonably good for $n\mu_0 \geq 10$.

Note that the exact distribution of D, and hence the accuracy of the χ^2 approximation, depends only on the product $n\mu_0$. If μ_0 is small, a very large value of n will be needed before (11.3.1) is applicable. If μ_0 is large, (11.3.1) will give accurate results even for $n = 1$.

As in the preceding example, the likelihood ratio statistic is a discrete variate whereas the χ^2 approximation is continuous. When m is small, there are only a few values with appreciable probabilities and so we cannot expect (11.3.1) to be accurate. When m is large, there are many more D-values with non-negligible probabilities and so the effects of discreteness will be less serious.

EXAMPLE 11.3.3. Suppose that an experiment yields n lifetimes which are modeled as observed values of IID exponential variates with mean θ. From Examples 9.4.1 and 9.7.2, the MLE of θ is $\hat{\theta} = t/n$ where $t = \Sigma x_i$, and the log

Table 11.3.2. Exact and Approximate Probabilities for the Likelihood Ratio Statistic in a Poisson Example

$n\mu_0$	$P(D \leq 2.706)$	$P(D \leq 3.841)$	$P(D \leq 6.635)$
5	0.928	0.928	0.988
10	0.884	0.944	0.986
20	0.881	0.958	0.990
40	0.886	0.951	0.991
χ^2 approx.	0.9	0.95	0.99

RLF is

$$r(\theta) = -n\left[\frac{\hat{\theta}}{\theta} - 1 - \log \frac{\hat{\theta}}{\theta}\right] = -n\left[\frac{t}{n\theta} - 1 - \log \frac{t}{n\theta}\right].$$

Now imagine that the experiment is repeated over and over with θ fixed at θ_0. The total lifetime $T \equiv \Sigma X_i$ is now modeled as a continuous random variable. Thus the likelihood ratio statistic

$$D \equiv -2r(\theta_0) \equiv 2n\left[\frac{T}{n\theta_0} - 1 - \log \frac{T}{n\theta_0}\right]$$

is also a continuous variate. Note that

$$D \equiv 2n[Y - 1 - \log Y]$$

where $Y \equiv T/n\theta_0$, and so

$$P(D \leq d) = P(Y - 1 - \log Y \leq d/2n).$$

Consider the function

$$g(y) = y - 1 - \log y \qquad \text{for } y > 0.$$

Since $g'(y) = 1 - \frac{1}{y}$ and $g''(y) = 1/y^2$, we see that $g(y)$ has a unique minimum value $g(1) = 0$. Also, $g(y) \to \infty$ as $y \to 0$ and as $y \to \infty$. Thus, for every $d > 0$, there will exist two values y_1, y_2 with $y_1 < 1 < y_2$ such that $g(y) \leq d/2n$ if and only if $y_1 \leq y \leq y_2$. To find these values, we can solve the equation $g(y) - d/2n = 0$ by Newton's method. We then have

$$P(D \leq d) = P(y_1 \leq Y \leq y_2) = P\left(y_1 \leq \frac{T}{n\theta_0} \leq y_2\right).$$

To evaluate this probability, we note that, from Problem 6.9.7, the variate $U \equiv 2T/\theta_0$ has a χ^2 distribution with $2n$ degrees of freedom. Therefore,

$$P(D \leq d) = P\left(2ny_1 \leq \frac{2T}{\theta_0} \leq 2ny_2\right) = P(2ny_1 \leq \chi^2_{(2n)} \leq 2ny_2).$$

For instance, suppose that $n = 5$ and we wish to evaluate $P(D \leq 2.706)$. Then $d/2n = 0.2706$, and y_1, y_2 are the roots of the equation

$$y - 1 - \log y - 0.2706 = 0.$$

Solving this equation by Newton's method gives $y_1 = 0.4326$ and $y_2 = 1.9261$. Thus we have

$$P(D \leq 2.706) = P(4.326 \leq \chi^2_{(10)} \leq 19.261).$$

Table B4 is not sufficiently detailed to permit accurate evaluation of this probability. However, a computer program for the c.d.f. of the χ^2 distribution gives

$$P\{\chi^2_{(10)} \le 4.326\} = 0.0686;$$

$$P\{\chi^2_{(10)} \le 19.261\} = 0.9629.$$

It follows that, for $n = 5$,

$$P(D \le 2.706) = 0.9629 - 0.0686 = 0.8943.$$

The other entries in Table 11.3.3 may be calculated similarly.

In this example the exact distribution of D depends on n but not θ_0. The approximation (11.3.1) is quite accurate even for n as small as 2 or 3. Here D is a continuous variate, and we do not have to contend with the effects of discreteness as in the preceding two examples.

JUSTIFICATION OF (11.3.2). We conclude this section by sketching a derivation of the χ^2 approximation (11.3.2). A rigorous proof of this result is beyond the scope of the book.

We assume that the data are observed values of n IID random variables $X_1, X_2, \ldots, X_n$. Since the X_i's are independent, it follows by (9.2.3) that the score function $S(\theta)$ can be written as a sum of n independent components. We shall show in Section 11.6 that $S(\theta_0)$ has mean 0 and variance $E\{\mathcal{I}(\theta_0)\}$. By the Central Limit Theorem (6.7.1) we have

$$S(\theta_0)/\sqrt{E\{\mathcal{I}(\theta_0)\}} \approx N(0, 1)$$

for n sufficiently large.

The MLE is the solution of $S(\hat{\theta}) = 0$, and it can be shown using the preceding result that $\hat{\theta}$ tends to θ_0 with probability 1 as $n \to \infty$. It can then be shown that $\mathcal{I}(\hat{\theta})/E\{\mathcal{I}(\theta_0)\} \to 1$ with probability 1 as $n \to \infty$, and hence

$$S(\theta_0)/\sqrt{\mathcal{I}(\hat{\theta})} \approx N(0, 1) \tag{11.3.3}$$

for n sufficiently large.

Since $\hat{\theta}$ will be close to θ_0 when n is large, the normal approximation (9.7.2) gives

$$r(\theta_0) \approx -\tfrac{1}{2}(\hat{\theta} - \theta_0)^2 \mathcal{I}(\hat{\theta}).$$

Table 11.3.3. Exact and Approximate Probabilities for the Likelihood Ratio Statistic in an Exponential Example

n	$P(D \le 2.706)$	$P(D \le 3.841)$	$P(D \le 6.635)$
1	0.874	0.932	0.984
2	0.886	0.941	0.987
3	0.891	0.944	0.988
5	0.894	0.946	0.989
10	0.897	0.948	0.989
χ^2 approx.	0.9	0.95	0.99

Differentiating with respect to θ_0 gives

$$S(\theta_0) \approx (\hat{\theta} - \theta_0)\mathcal{I}(\hat{\theta}).$$

It follows by (11.3.3) that

$$(\hat{\theta} - \theta_0)\sqrt{\mathcal{I}(\hat{\theta})} \approx N(0, 1). \tag{11.3.4}$$

The likelihood ratio statistic is

$$D \equiv -2r(\theta_0) \approx (\hat{\theta} - \theta_0)^2 \mathcal{I}(\hat{\theta}) \approx Z^2$$

where $Z \approx N(0, 1)$ by (11.3.4). It follows by (6.9.8) that D has approximately a $\chi^2_{(1)}$ distribution when n is large. □

PROBLEMS FOR SECTION 11.3

1. Consider the situation described in Example 11.3.2.
   (a) Show that, if $n\mu_0 = 9$, then $r(\mu_0) \ge \log 0.1$ if and only if $4 \le T \le 16$. Hence find the exact coverage probability of the 10% likelihood interval when $n\mu_0 = 9$.
   (b) Investigate the behavior of the coverage probability of the 10% likelihood interval for $9 \le n\mu_0 \le 10$.

2.† Let $X_1, X_2, \ldots, X_n$ be IID random variables with p.d.f.

$$f(x) = 2\lambda x e^{-\lambda x^2} \qquad \text{for } x > 0,$$

where λ is a positive unknown parameter.

   (a) Show that the likelihood ratio statistic is

$$D \equiv -2r(\lambda_0) \equiv T - 2n - 2n \log(T/2n)$$

   where $T \equiv 2\lambda_0 \Sigma X_i^2$.
   (b) Show that $2\lambda_0 X_i^2$ has a χ^2 distribution with two degrees of freedom, and hence that $T \sim \chi^2_{(2n)}$.
   (c) Show that the coverage probability of the $100p\%$ likelihood interval is the same as in Example 11.3.3.

11.4. Confidence Intervals

The random interval $[A, B]$ is called a *confidence interval* (CI) for θ if its coverage probability

$$\mathrm{CP}(\theta_0) = P(A \le \theta_0 \le B | \theta = \theta_0)$$

is the same for all parameter values θ_0. The coverage probability of a confidence interval is called its *confidence coefficient*.

For instance, $[A, B]$ is a 95% confidence interval for θ if

$$P(A \le \theta_0 \le B | \theta = \theta_0) = 0.95$$

for all possible parameter values θ_0. A 95% CI would include the true parameter value θ_0 in 95% of repetitions of the experiment with θ fixed.

In Examples 11.2.2, 11.2.3, and 11.3.3 we found that the coverage probability of the $100p\%$ likelihood interval was the same for all parameter values. In each of these examples, the $100p\%$ LI is a confidence interval. In particular, the confidence coefficient of the 14.7% LI is exactly 0.95 in Examples 11.2.2 and 11.2.3, and is close to 0.95 in Example 11.3.3.

Likelihood intervals are not confidence intervals in Examples 11.2.1, 11.3.1, and 11.3.2 because their coverage probabilities depend on the true parameter value θ_0. In general, when the probability model is discrete, the coverage probability of a random interval $[A, B]$ will be a discontinuous function of θ_0 (see Example 11.2.1). For this reason, it is generally not possible to construct exact confidence intervals in the discrete case. However, the effects of discreteness become less important as the sample size increases. Thus it is often possible to find approximate confidence intervals for which $\mathrm{CP}(\theta_0)$ is nearly constant over those parameter values θ_0 which are of interest.

Because of the χ^2 approximation, likelihood intervals are exact or approximate confidence intervals in most applications. When (11.3.1) applies, the approximate confidence coefficient (coverage probability) of the $100p\%$ likelihood interval is given by

$$\mathrm{CP} \approx P\{\chi^2_{(1)} \leq -2 \log p\}$$

(see Table 11.2.1).

Interpretation

Except in special cases, it is not correct to conclude that a particular observed 95% confidence interval $[a, b]$ has a 95% probability of including the true value of θ. It can happen that $[a, b]$ contains all possible values of θ, and so covers θ_0 with probability 100%. The 95% coverage probability is a theoretical average figure which refers to an imaginary sequence of repetitions of the experiment. It is a property of the method used to construct the interval rather than of the interval calculated in any particular case.

In most applications one has a particular observed data set and wants to know what can be learned from it about the value of θ. If confidence intervals are to be useful in such applications, they must be constructed in such a way that an individual observed interval $[a, b]$ *does* provide a reasonable information summary. Values inside the interval should be in some sense better estimates of θ than values outside the interval.

For this reason, it is recommended that confidence intervals be constructed from the likelihood function. If a 95% confidence interval for θ is desired, then a $100p\%$ likelihood interval is calculated where p is selected to give the desired coverage probability of 0.95. *Intervals constructed in this way will have the desired long-run coverage properties, and in addition, they will provide useful information summaries in particular applications.*

Another method of constructing confidence intervals is by inverting a test of significance (see Section 12.9).

EXAMPLE 11.4.1. In Example 11.2.2 we noted that

$$Z \equiv X - \theta_0 \sim N(0, 1).$$

Since $P\{-1.96 \leq Z \leq 1.96\} = 0.95$, it follows that

$$P(X - 1.96 \leq \theta_0 \leq X + 1.96) = P(-1.96 \leq Z \leq 1.96) = 0.95.$$

The interval $[X - 1.96, X + 1.96]$ has coverage probability 0.95 for all θ_0, and therefore it is a 95% confidence interval for θ. It is also a likelihood interval. Values of θ included by this interval are more likely than the excluded values.

There are plenty of ways to construct confidence intervals in this example. For instance, Table B2 gives

$$P(-2.376 \leq Z \leq 1.751) = 0.95.$$

Thus the interval $[X - 2.376, X + 1.751]$ has coverage probability 0.95 for all θ_0, and is also a 95% confidence interval for θ. Note, however, that this is not a likelihood interval. It includes values of θ at the lower end which are much less likely than values excluded at the upper end. Although this interval would cover the true parameter value 95% of the time in repetitions of the experiment, it would not properly summarize the information available concerning θ in any particular application.

Use of the Normal Approximation

The recommended method for obtaining, say, a 95% confidence interval is to calculate the $100p\%$ likelihood interval, where p is chosen so that the coverage probability is close to 0.95. When (11.3.1) holds, the 14.7% likelihood interval is an exact or approximate 95% confidence interval. It can be found from a graph of $r(\theta)$ as in Section 9.3, or by Newton's method as in Section 9.8.

Since confidence intervals constructed in this way are also likelihood intervals, they will provide proper summaries of the information available in particular cases. A disadvantage of this construction is that a fair bit of arithmetic may be needed to compute the likelihood interval. However, with high-speed computers so widely available, this is not a serious problem.

Sometimes one can avoid most of the arithmetic by using the normal approximation of Section 9.7. By (9.7.3), the interval $\hat{\theta} \pm c/\sqrt{\mathscr{I}(\hat{\theta})}$ is an approximate likelihood interval for θ. Its coverage probability is

$$\begin{aligned}\mathrm{CP}(\theta_0) &= P\{\hat{\theta} - c/\sqrt{\mathscr{I}(\hat{\theta})} \leq \theta_0 \leq \hat{\theta} + c/\sqrt{\mathscr{I}(\hat{\theta})} \mid \theta = \theta_0\} \\ &= P\{-c \leq (\hat{\theta} - \theta_0)\sqrt{\mathscr{I}(\hat{\theta})} \leq c \mid \theta = \theta_0\}.\end{aligned}$$

Now (11.3.4) gives

$$\mathrm{CP}(\theta_0) \approx P(-c \le Z \le c)$$

where $Z \sim N(0, 1)$.

Since $P(-1.96 \le Z \le 1.96) = 0.95$, the interval

$$\hat{\theta} \pm 1.96/\sqrt{\mathscr{I}(\hat{\theta})} \tag{11.4.1}$$

is an approximate 95% confidence interval. Similarly, $\hat{\theta} \pm 1.645/\sqrt{\mathscr{I}(\hat{\theta})}$ is an approximate 90% confidence interval, and $\hat{\theta} \pm 2.576/\sqrt{\mathscr{I}(\hat{\theta})}$ is an approximate 99% confidence interval.

Although we can save arithmetic by using (11.4.1), there are two disadvantages. The more serious is that the normal approximation (9.7.2) may be inaccurate. If this is so, the interval (11.4.1) may exclude some of the plausible parameter values and include some parameter values which are very implausible. The second disadvantage is that the approximation (11.3.4), which was used to evaluate the coverage probability of (11.4.1), is generally less accurate than (11.3.1). Sometimes both of these difficulties can be overcome by making a suitable nonlinear transformation as in Section 9.7. However, in general, it is safer to compute likelihood intervals instead of relying on (11.4.1).

EXAMPLE 11.4.2. Suppose that $x = 17$ successes are observed in $n = 100$ Bernoulli trials with success probability θ. We wish to find an approximate 95% confidence interval for θ.

The MLE is $\hat{\theta} = x/n = 0.17$, and the log RLF is

$$r(\theta) = 17 \log \theta + 83 \log(1 - \theta) + 45.589 \qquad \text{for } 0 < \theta < 1.$$

One can show that $r(\theta) \ge \log 0.147$ for $0.105 \le \theta \le 0.251$. This is a 14.7% likelihood interval and also an approximate 95% confidence interval for θ. Parameter values which belong to this interval are more plausible than parameter values outside the interval. Also, over a series of repetitions of the experiment with θ fixed, intervals constructed in this way would cover the true value of θ about 95% of the time.

From Example 9.1.1, the information function is

$$\mathscr{I}(\theta) = \frac{x}{\theta^2} + \frac{n - x}{(1 - \theta)^2} \qquad \text{for } 0 < \theta < 1.$$

Substituting $\theta = \hat{\theta}$ gives

$$\mathscr{I}(\hat{\theta}) = \frac{x}{\hat{\theta}^2} + \frac{n - x}{(1 - \hat{\theta})^2} = \frac{n}{\hat{\theta}} + \frac{n}{1 - \hat{\theta}} = \frac{n}{\hat{\theta}(1 - \hat{\theta})}.$$

Now (11.4.1) gives

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}} = 0.17 \pm 0.0736,$$

or $0.096 \le \theta \le 0.244$, as the approximate 95% confidence interval for θ. Intervals constructed in this way would cover the true value of θ about 95% of the time in repetitions of the experiment with θ fixed. However, this interval is not a likelihood interval. Its endpoints have relative likelihoods

$$R(0.096) = 0.072; \qquad R(0.244) = 0.200.$$

Thus the interval includes values at the lower end which are much less plausible than values excluded at the upper end. For this reason, the first construction is preferable.

EXAMPLE 11.4.3. Suppose that we have $n = 10$ independent observations from an exponential distribution with mean θ, and that $\hat{\theta} = 28.8$ (see Example 9.7.2). We wish to obtain an approximate 95% confidence interval for θ.

From Example 9.7.2, the log RLF of θ is

$$r(\theta) = -n\left[\frac{\hat{\theta}}{\theta} - 1 - \log\frac{\hat{\theta}}{\theta}\right] = -10\left[\frac{28.8}{\theta} - 1 - \log\frac{28.8}{\theta}\right]$$

for $\theta > 0$. By plotting this function, or by Newton's method, we find that $r(\theta) \ge \log 0.147$ for $16.42 \le \theta \le 57.47$. This is an approximate 95% confidence interval. In fact, from Table 11.3.3, the exact coverage probability of the 14.7% likelihood interval in this situation is 0.948.

Alternatively, we note from Example 9.7.2 that

$$\sqrt{\mathscr{I}(\hat{\theta})} = \sqrt{n}/\hat{\theta} = 0.1098.$$

By (11.4.1), the approximate 95% confidence interval is 28.8 ± 17.85; that is, $10.95 \le \theta \le 46.65$.

It can be shown that, with $n = 10$ independent observations from an exponential distribution, the exact coverage probability of the interval $\hat{\theta} \pm 1.96/\sqrt{\mathscr{I}(\hat{\theta})}$ is 0.9035. The approximation (11.3.4) is not very accurate in this case. More seriously, the interval constructed is symmetric about $\hat{\theta}$ whereas the log RLF is highly skewed (see Figure 9.7.1). The interval includes very implausible values of θ at the lower end, and excludes fairly plausible values at the upper end.

More satisfactory results are obtained if we apply (11.4.1) to the transformed parameter $\lambda = \theta^{-1/3}$. From Example 9.7.2 we have

$$\hat{\lambda} = \hat{\theta}^{-1/3} = 0.3262; \qquad \mathscr{I}_*(\hat{\lambda}) = 9n/\hat{\lambda}^2 = 845.6.$$

By (11.4.1), the approximate 95% confidence interval for λ is

$$\hat{\lambda} \pm 1.96/\sqrt{\mathscr{I}_*(\hat{\lambda})} = 0.3262 \pm 0.0674,$$

or $0.2588 \le \lambda \le 0.3936$. Since $\theta = \lambda^{-3}$, the interval for θ is

$$(0.2588)^{-3} \ge \theta \ge (0.3936)^{-3},$$

which gives $16.39 \le \theta \le 57.66$. This is very nearly a likelihood interval, and it can be shown that the exact coverage probability of intervals constructed in this way is 0.949.

Problems for Section 11.4

1. Suppose that the distribution of the likelihood ratio statistic $D = -2r(\theta_0)$ does not depend upon θ_0. Show that, for all p, the $100p\%$ likelihood interval for θ is a confidence interval.

2. In a poll of 200 randomly chosen voters, 94 indicated that they would vote for the Conservatives if an election were called. Let p be the proportion of all voters who would vote for the Conservatives. Find a likelihood interval which is an approximate 95% confidence interval for p. Is it likely that $p = \frac{1}{2}$?

3.† Five hundred people were chosen at random from a large population and were asked their opinions on capital punishment for murderers of prison guards. Sixty percent of those interviewed were in favor. Let p denote the fraction of the population who favor capital punishment.

   (a) Find likelihood intervals for p which are approximate 95% and 99% confidence intervals.
   (b) Use the normal approximation to construct approximate 95% and 99% confidence intervals, and compare them with the intervals in (a).

4.† The following are the times to failure, measured in hours, of ten electronic components:

2	119	51	77	33	27	14	24	4	37

   Previous experience with similar types of components suggests that the distribution of lifetimes should be exponential. The mean lifetime θ is unknown.

   (a) Find a likelihood interval for θ which is an approximate 90% confidence interval.
   (b) Transform the result in (a) to obtain an approximate 90% confidence interval for p, the proportion of components whose lifetimes exceed 50 hours.

5. The number of accidents per month at a busy intersection has a Poisson distribution with mean μ, and successive months are independent. Over a 10-month period there were 53 accidents altogether.

   (a) Obtain a likelihood interval which is an approximate 97.5% confidence interval for μ.
   (b) Use the normal approximation of Section 9.7 to obtain an approximate 97.5% confidence interval for μ.

6. When an automatic shear is set to cut plates to length μ, the lengths actually produced are normally distributed about μ with standard deviation 1.6 inches. The average length of 15 plates cut at one setting was 125.77 inches. Find three likelihood intervals for μ which are 90%, 95%, and 99% confidence intervals.

7.† In a check of the accuracy of their measurement procedures, fifteen engineers are asked to measure a precisely known distance of 3727 feet between two markers. Their results are as follows:

3727.75	3726.43	3728.04	3729.21	3726.30
3728.15	3724.25	3726.29	3724.90	3727.51
3726.85	3728.50	3725.94	3727.69	3726.09

   Assuming that their measurements are independent $N(3727, \sigma^2)$, obtain a likelihood interval for σ which is an approximate 95% confidence interval.

8. Let $X_1, X_2, \ldots, X_n$ be IID $N(\mu, \sigma^2)$ variates, where μ is known but σ is unknown (see Problem 11.4.7).

   (a) Show that the likelihood ratio statistic is

$$D \equiv -2r(\sigma_0) \equiv T - n - n\log(T/n)$$

   where $T \equiv \Sigma(x_i - \mu)^2/\sigma_0^2$.
   (b) Show that T has a χ^2 distribution with n degrees of freedom.
   (c) Show that the $100p\%$ likelihood interval is a confidence interval, and describe how its exact confidence coefficient can be determined.

9. Let $X_1, X_2, \ldots, X_n$ be IID random variables having a gamma distribution with p.d.f.

$$f(x) \equiv \frac{x}{\theta^2}\exp\left(-\frac{x}{\theta}\right) \qquad \text{for } x > 0$$

   where θ is a positive unknown parameter.

   (a) Show that the likelihood ratio statistic is

$$D \equiv -2r(\theta_0) \equiv T - 4n - 4n\log(T/4n)$$

   where $T \equiv 2\Sigma X_i/\theta_0$.
   (b) Show that $2X_i/\theta_0$ has a χ^2 distribution with 4 degrees of freedom, and hence that $T \sim \chi^2_{(4n)}$.
   (c) The total of $n = 60$ observations was found to be $\Sigma x_i = 71.5$. Find a likelihood interval for θ which is an approximate 90% confidence interval. Will the exact coverage probability be close to 90% in this situation?

10. Let $X_1, X_2, \ldots, X_n$ be IID exponential variates with mean θ, and define $T \equiv \Sigma X_i$. We noted in Example 11.3.3 that the distribution of $2T/\theta_0$ in repetitions of the experiment with $\theta = \theta_0$ is $\chi^2_{(2n)}$. Let a, b be values such that

$$P\{\chi^2_{(2n)} \le a\} = 0.025 = P\{\chi^2_{(2n)} \ge b\}.$$

   (a) Show that the interval

$$\frac{2T}{b} \le \theta \le \frac{2T}{a}$$

   is a 95% confidence interval for θ.
   (b) Let θ_L and θ_u denote the lower and upper endpoints of the confidence interval in (a). Show that

$$r(\theta_u) - r(\theta_L) = n\log\frac{a}{b} + \frac{1}{2}(b - a).$$

   (c) Using tables of the χ^2 distribution, evaluate $r(\theta_u) - r(\theta_L)$ for $n = 1, 5, 10$, and 15. Is the interval in (a) a likelihood interval? What happens as n increases?

11.† Let $X_1, X_2, \ldots, X_n$ be IID variates having a continuous uniform distribution on the interval $[0, \theta]$, where θ is a positive unknown parameter.

(a) Show that the likelihood ratio statistic is

$$D \equiv -2r(\theta_0) \equiv -2n \log(M/\theta_0) \qquad \text{for } M \le \theta_0$$

where M is the largest of the X_i's.

(b) Show that

$$P\{M \le m | \theta = \theta_0\} = (m/\theta_0)^n \qquad \text{for } m \le \theta_0;$$

$$P\{D \le d | \theta = \theta_0\} = 1 - e^{-d/2} \qquad \text{for } d > 0.$$

Thus D is distributed as $\chi^2_{(2)}$. Note that (11.3.1) does not apply here because the range of the X_i's depends on θ.

(c) Show that the $100p\%$ likelihood interval has coverage probability $1 - p$.

(d) Find a likelihood interval for θ which is a 95% confidence interval based on the following sample of size 10.

0.7481	0.7484	0.9537	0.1589	0.3773
0.3345	0.2906	0.8527	0.3479	0.9245

11.5. Results for Two-Parameter Models

Suppose that the probability model for the experiment involves two unknown parameters, α and β. Let $r(\alpha, \beta)$ denote the joint log RLF of α and β as in Section 10.1. The $100p\%$ likelihood region for (α, β) is the set of parameter values such that $r(\alpha, \beta) \ge \log p$ (see Section 10.2).

In Section 10.3, we defined the maximum log RLF of β to be the maximum of $r(\alpha, \beta)$ over α with β fixed:

$$r_{\max}(\beta) = \max_{\alpha} r(\alpha, \beta).$$

The $100p\%$ maximum likelihood interval for β is the set of all β-values such that $r_{\max}(\beta) \ge \log p$. This interval can be found from a graph of $r_{\max}(\beta)$, or from a contour map of $r(\alpha, \beta)$.

Now imagine a series of repetitions of the experiment with (α, β) fixed at (α_0, β_0). We consider two likelihood ratio statistics:

$$D \equiv -2r(\alpha_0, \beta_0);$$

$$D_2 \equiv -2r_{\max}(\beta_0).$$

D is the likelihood ratio statistic for testing the hypothesis $(\alpha, \beta) = (\alpha_0, \beta_0)$, and D_2 is the likelihood ratio statistic for testing the hypothesis $\beta = \beta_0$ (see Sections 12.2 and 12.3).

The values of D and D_2 would vary from one repetition of the experiment to the next depending upon the data obtained. In principle, their exact sampling distributions can be derived from the probability model. In practice, this is usually difficult to do, and so approximations are used.

It can be shown, under conditions similar to those given in Section 11.3, that the distributions of D and D_2 can be approximated by χ^2 distributions in large samples:

$$D \approx \chi^2_{(2)} \qquad \text{and } D_2 \approx \chi^2_{(1)}.$$

The χ^2 approximation has 2 degrees of freedom for D and only one degree of freedom for D_2. See Section 12.3 for a discussion of degrees of freedom.

The true value (α_0, β_0) belongs to the $100p\%$ likelihood region if and only if $r(\alpha_0, \beta_0) \ge \log p$. Thus the coverage probability of the $100p\%$ likelihood region for (α, β) is

$$\begin{aligned} \text{CP}(\alpha_0, \beta_0) &= P(D \le -2 \log p | \alpha = \alpha_0, \beta = \beta_0) \\ &\approx P(\chi^2_{(2)} \le -2 \log p). \end{aligned}$$

The exact coverage probability may depend upon α_0 and β_0, but the approximation does not. Consequently, likelihood regions are approximate confidence regions in large samples.

By (6.9.3), the c.d.f. of $\chi^2_{(2)}$ is

$$P(\chi^2_{(2)} \le d) = 1 - e^{-d/2} \qquad \text{for } d > 0.$$

It follows that

$$\text{CP} \approx P(\chi^2_{(2)} \le -2 \log p) = 1 - e^{\log p} = 1 - p.$$

The $100p\%$ likelihood region for (α, β) is an approximate $100(1-p)\%$ confidence region.

The true value β_0 belongs to the $100p\%$ maximum likelihood interval for β if and only if $r_{\max}(\beta_0) \ge \log p$. Thus the coverage probability of the $100p\%$ maximum likelihood interval for β is

$$\begin{aligned} \text{CP}(\alpha_0, \beta_0) &= P(D_2 \le -2 \log p | \alpha = \alpha_0, \beta = \beta_0) \\ &\approx P(\chi^2_{(1)} \le -2 \log p). \end{aligned}$$

Maximum likelihood intervals are approximate confidence intervals. They have the same approximate coverage probabilities as likelihood intervals in the one-parameter case.

Figures 10.2.2, 10.2.3, and 10.6.1 show both 10% likelihood regions and 10% maximum likelihood intervals for three numerical examples. The 10% likelihood region consists of all points on or within the 10% contour, which is roughly elliptical in shape. This region would include the true values of *both* parameters in about 90% of repetitions of the experiment with both parameters fixed. The broken vertical lines show the 10% maximum likelihood interval for the first parameter. The true value of the first parameter would lie between these lines about 96.8% of the time in repetitions with both parameters fixed. Similarly, the true value of the second parameter would lie between the broken horizontal lines about 96.8% of the time.

In the one-parameter case we considered two normal distribution

examples for which the distribution of the likelihood ratio statistic was exactly $\chi^2_{(1)}$. A two-parameter example is given below in which the distributions of D and D_2 are exactly $\chi^2_{(2)}$ and $\chi^2_{(1)}$, respectively. It can be shown that the same is true in Example 10.1.1.

EXAMPLE 11.5.1. Suppose that an experiment involves taking two measurements x, y which are modeled as observed values of independent variates $X \sim N(\alpha, 1)$ and $Y \sim N(\beta, 1)$. It is easy to show that $\hat{\alpha} = x$, $\hat{\beta} = y$, and

$$r(\alpha, \beta) = -\tfrac{1}{2}(x-\alpha)^2 - \tfrac{1}{2}(y-\beta)^2$$

for $-\infty < \alpha < \infty$ and $-\infty < \beta < \infty$, and that

$$r_{\max}(\beta) = -\tfrac{1}{2}(y-\beta)^2.$$

Imagine a series of repetitions of this experiment with $\alpha = \alpha_0$ and $\beta = \beta_0$. The two likelihood ratio statistics are as follows:

$$D \equiv -2r(\alpha_0, \beta_0) \equiv (X-\alpha_0)^2 + (Y-\beta_0)^2 \equiv Z_1^2 + Z_2^2;$$

$$D_2 \equiv -2r_{\max}(\beta_0) \equiv (Y-\beta_0)^2 \equiv Z_2^2.$$

Here $Z_1 \equiv X - \alpha_0$ and $Z_2 \equiv Y - \beta_0$ are independent $N(0, 1)$ variates. It follows by (6.9.9) and (6.9.8) that $D \sim \chi^2_{(2)}$ and $D_2 \sim \chi^2_{(1)}$.

Use of Normal Approximations

Normal approximations to $r(\alpha, \beta)$ and $r_{\max}(\beta)$ were given in Section 10.4. These results may be used to obtain approximate likelihood regions and maximum likelihood intervals. Their approximate coverage probabilities can be obtained from the χ^2 approximations.

For instance, (10.4.5) gives

$$r_{\max}(\beta) \approx -\tfrac{1}{2}(\hat{\beta} - \beta)^2 / \hat{\mathscr{I}}^{22}$$

where $\hat{\mathscr{I}}^{22}$ is defined in Section 10.4. It follows from this result that

$$\hat{\beta} \pm c\sqrt{\hat{\mathscr{I}}^{22}} \tag{11.5.1}$$

is an approximate $100p\%$ maximum likelihood interval, where $c = \sqrt{-2 \log p}$.

By the χ^2 approximation to D_2, the approximate coverage probability of $100p\%$ maximum likelihood intervals is

$$P(\chi^2_{(1)} \le -2 \log p) = P(\chi^2_{(1)} \le c^2) = P(-c \le Z \le c)$$

where $Z \sim N(0, 1)$. Thus we see that the interval (11.5.1) is an approximate confidence interval with confidence coefficient $P(-c \le Z \le c)$. In particular, (11.5.1) gives an approximate 95% confidence interval when $c = 1.96$.

This procedure will not produce sensible confidence intervals unless

(10.4.5) gives a good approximation to $r_{\max}(\beta)$. A nonlinear parameter transformation may help. See the discussion in Sections 9.7 and 10.4.

PROBLEMS FOR SECTION 11.5

1. Use (11.5.1) to obtain approximate 95% confidence intervals for β and γ in Example 10.5.1.

2. (a) Find approximate 90% confidence intervals for α and β in Problem 10.1.1.
   (b) Use the result of Problem 10.4.4(c) to obtain an approximate 90% confidence interval for the parameter $\gamma = \alpha - \beta$ in Problem 10.1.1.

3.† Use Problem 10.4.4(c) to obtain an approximate 99% confidence interval for the parameter $\gamma = \alpha + 2\beta$ in Example 10.5.1. Transform this interval to obtain an approximate 99% confidence interval for the probability of death at log concentration $d = 2$.

4. (a) Let X and Y be independent Poisson variates with means μ_1 and μ_2, and define $\gamma = \log(\mu_2/\mu_1)$. Derive the information matrix $\mathscr{I}(\hat{\mu}_1, \hat{\mu}_2)$. Then use Problem 10.4.4(c) to show that

$$\log(Y/X) \pm 1.96\sqrt{\frac{1}{X} + \frac{1}{Y}}$$

   is an approximate 95% confidence interval for γ.
   (b) Calculate an approximate 95% confidence interval for $\log \lambda$ in Problem 10.3.2(b), and transform it to obtain an approximate 95% confidence interval for λ.

5. (a) Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent exponentially distributed variates, with $E(X_i) = \theta_1$ and $E(Y_j) = \theta_2$. Define $\gamma = \log(\theta_2/\theta_1)$. Show that

$$\log(\bar{Y}/\bar{X}) \pm 1.96\sqrt{\frac{1}{n} + \frac{1}{m}}$$

   is an approximate 95% confidence interval for γ.
   (b) Use the result in (a) to obtain an approximate 95% confidence for λ in Problem 10.3.3(c).

6.† Find an approximate 95% confidence interval for the median of the Weibull distribution in Example 10.4.2.
Hint: Show that $\log m = \log \theta + (\log \log 2)/\beta$, and use the result in Problem 10.4.4(c).

7. Let $Y_1, Y_2, \ldots, Y_n$ be independent $N(\mu, \sigma^2)$ random variables where μ and σ are unknown, and define $\gamma = \mu/\sigma$. Derive the information matrix $\mathscr{I}(\hat{\mu}, \hat{\sigma})$. Then use Problem 10.4.4(c) and (11.5.1) to show that

$$\hat{\gamma} \pm 1.96\sqrt{\frac{1}{n}(1 + \tfrac{1}{2}\hat{\gamma}^2)}$$

   is an approximate 95% confidence interval for γ.

*11.6. Expected Information and Planning Experiments

Up to this point we have assumed that the experiment had already been selected and performed, and we have considered the problem of extracting information about an unknown parameter θ from the data. Statistical methods are also useful at the planning stage in deciding what experiment should be performed.

From Sections 9.7 and 11.4, the interval

$$\hat{\theta} \pm c/\sqrt{\mathcal{I}(\hat{\theta})}$$

is an approximate likelihood/confidence interval for θ. As we noted in Section 9.7, a large value of $\mathcal{I}(\hat{\theta})$ implies a short interval for θ. Thus, if the aim of the experiment is to estimate θ, we should try to select an experiment for which $\mathcal{I}(\hat{\theta})$ will be large.

In general, $\mathcal{I}(\hat{\theta})$ is a function of $\hat{\theta}$ and possibly other features of the data. Its value will not be known until after the experiment is performed. Consequently, it is usually not possible to use $\mathcal{I}(\hat{\theta})$ in planning the experiment. Instead, one can examine the expected value of the information function,

$$\mathcal{I}_E(\theta) = E\{\mathcal{I}(\theta)\}.$$

This is called the *expected information function* of θ, or *Fisher's measure of expected information*. We shall show at the end of this section that $\mathcal{I}_E(\theta) \geq 0$ for all parameter values.

The expected information is a measure of the average precision that would be attained over a large number of repetitions of the experiment. It is relevant only at the planning stage. Once the experiment has been performed, its actual precision can be assessed by calculating $\mathcal{I}(\hat{\theta})$, or better yet, by examining a graph of $r(\theta)$.

The following two examples illustrate the use of expected information in planning experiments.

EXAMPLE 11.6.1. Suppose that n items are to be tested for a preset length of time T, and the number Y which survive is to be recorded. The lifetimes are assumed to be independent exponential variates with mean θ. If T is chosen to be too small, the experiment may terminate before any failures occur. If T is too large, all items may fail long before the experiment ends. In either of these cases, one would expect to learn very little about θ. How large should T be chosen in order to maximize the expected information about θ?

SOLUTION. Since the lifetime X of an item is assumed to have an exponential distribution with mean θ, the probability that an item survives time T is

*This section may be omitted on first reading.

$$p = P(X > T) = \int_T^\infty \frac{1}{\theta} e^{-x/\theta} dx = e^{-T/\theta}.$$

Then Y, the number surviving, has a binomial (n, p) distribution, and the log likelihood function of θ is

$$l(\theta) = y \log p + (n - y) \log(1 - p)$$

where $p = e^{-T/\theta}$. Differentiating twice with respect to θ gives

$$\mathcal{I}(\theta) = \left[\frac{y}{p^2} + \frac{n-y}{(1-p)^2}\right]\left(\frac{dp}{d\theta}\right)^2 - \left[\frac{y}{p} - \frac{n-y}{1-p}\right]\frac{d^2p}{d\theta^2}.$$

Since $E(Y) = np$, the expected information function is

$$\mathcal{I}_E(\theta) = \frac{n}{p(1-p)}\left(\frac{dp}{d\theta}\right)^2 = \frac{npT^2}{(1-p)\theta^4} = \frac{n}{\theta^2} \cdot h(p)$$

where

$$h(p) = \frac{p(\log p)^2}{1-p} \qquad \text{and } p = e^{-T/\theta}.$$

With the aid of a calculator, one can easily show that $h(p)$ reaches a maximum value of 0.648 for $p = 0.203$. Thus we should try to pick T so that about 20% of the items tested will survive the test period. If we guess T nearly right, the expected information is about $0.64n/\theta^2$; that is, about 64% of the expected information n/θ^2 when all n exponential failure times are observed.

This example is rather artificial because we are not taking costs into account. If a shorter experiment costs less, one might well get "more information per dollar" by choosing a smaller value of T so that more than 20% of items would survive testing. □

EXAMPLE 11.6.2. Two experiments are being considered for obtaining information about a linkage parameter θ, where $0 < \theta < \frac{1}{2}$. Each experiment would involve observing multinomial frequencies X_1, X_2, X_3, X_4 where $\Sigma X_i \equiv n$. For the first experiment, the probabilities are

$$p_1 = p_2 = \theta/2; \qquad p_3 = p_4 = (1 - \theta)/2.$$

For the second experiment, the probabilities are

$$q_1 = (\theta^2 - 2\theta + 3)/4; \qquad q_2 = q_3 = (2\theta - \theta^2)/4; \qquad q_4 = (1 - \theta)^2/4.$$

Which experiment can be expected to yield more information about θ?

SOLUTION. The log likelihood function based on the multinomial distribution is $\Sigma x_i \log p_i$, and differentiating twice with respect to θ gives

$$\mathcal{I}(\theta) = \Sigma \frac{x_i}{p_i^2}\left(\frac{dp_i}{d\theta}\right)^2 - \Sigma \frac{x_i}{p_i}\left(\frac{d^2p_i}{d\theta^2}\right).$$

Since $E(X_i) = np_i$, the expected information function is

$$\mathcal{I}_E(\theta) = n\Sigma \frac{1}{p_i}\left(\frac{dp_i}{d\theta}\right)^2 - n\Sigma \frac{d^2 p_i}{d\theta^2}.$$

The latter sum is zero because $\Sigma p_i = 1$.

Since $dp_i/d\theta = \pm\frac{1}{2}$ and $dq_i/d\theta = \pm(1-\theta)/2$, the expected information functions for the two experiments are

$$\mathcal{I}_1(\theta) = \frac{n}{4}\Sigma\frac{1}{p_i} \qquad \text{and } \mathcal{I}_2(\theta) = \frac{n(1-\theta)^2}{4}\Sigma\frac{1}{q_i}.$$

The ratio of these two functions,

$$\frac{\mathcal{I}_2(\theta)}{\mathcal{I}_1(\theta)} = (1-\theta)^2 \frac{\Sigma 1/q_i}{\Sigma 1/p_i},$$

is called the *expected relative efficiency* for the second experiment versus the first, and is tabulated below:

θ	0.0	0.1	0.2	0.3	0.4	0.5
$\mathcal{I}_2/\mathcal{I}_1$	1	0.88	0.77	0.65	0.55	0.44

For all $\theta > 0$, the first experiment is more efficient (has larger expected information) than the second, and it is considerably more efficient for θ near $\frac{1}{2}$. If costs were equal, the first experiment would be preferable to the second. □

Properties of the Score and Information Functions

We conclude this section by showing that, under suitable regularity conditions, the score function $S(\theta)$ has expected value 0 and variance equal to the expected information $\mathcal{I}_E(\theta) = E\{\mathcal{I}(\theta)\}$. Since variances are non-negative, it then follows that $\mathcal{I}_E(\theta) \geq 0$.

Let X be a random variable or vector of random variables having probability or probability density function $f(x;\theta)$ which depends on a continuous parameter θ. The likelihood function of θ is proportional to $f(x;\theta)$,

$$L(\theta) = c \cdot f(x;\theta),$$

where c is positive and does not depend on θ. The score and information functions are

$$S(\theta) = \frac{\partial \log L}{\partial\theta} = \frac{\partial \log f}{\partial\theta};$$

$$\mathcal{I}(\theta) = -\frac{\partial^2 \log L}{\partial\theta^2} = -\frac{\partial \log S}{\partial\theta} = -\frac{\partial^2 \log f}{\partial\theta^2}.$$

As in Section 11.1, we imagine a series of repetitions of the experiment with θ fixed at a particular value. The value of X will vary from one repetition to the next. S and $\mathcal{I}$ are functions of X, and thus can be considered as random variables.

In what follows, we shall be taking expectations over the distribution of X. Usually these expectations would involve multiple sums or integrals. However, for simplicity, we shall write all expectations as single sums.

For any value of θ, the total probability in the distribution of X is equal to 1, so

$$\sum_x f(x;\theta) = 1 \qquad \text{for all } \theta.$$

Now we differentiate with respect to θ. Assuming that the order of differentiation and summation can be interchanged, we get

$$\sum_x \frac{\partial}{\partial\theta} f(x;\theta) = 0; \qquad \sum_x \frac{\partial^2}{\partial\theta^2} f(x;\theta) = 0.$$

But since

$$S = \frac{\partial \log f}{\partial\theta} = \frac{1}{f}\frac{\partial f}{\partial\theta},$$

we have $\dfrac{\partial f}{\partial\theta} = Sf$, and therefore

$$\sum_x Sf = \sum_x \frac{\partial f}{\partial\theta} = 0.$$

This shows that the expected value of the score function is zero. Also we have

$$\mathcal{I} = -\frac{\partial S}{\partial\theta} = -\frac{\partial}{\partial\theta}\left[\frac{1}{f}\frac{\partial f}{\partial\theta}\right] = \frac{1}{f^2}\left(\frac{\partial f}{\partial\theta}\right)^2 - \frac{1}{f}\frac{\partial^2 f}{\partial\theta^2},$$

from which we obtain

$$\frac{\partial^2 f}{\partial\theta^2} = S^2 f - \mathcal{I} f.$$

It now follows that

$$\sum_x S^2 f - \sum_x \mathcal{I} f = \sum_x \frac{\partial^2 f}{\partial\theta^2} = 0.$$

The first sum is $E(S^2)$, and the second sum is the expected information, so we have shown that

$$E(S^2) = E(\mathcal{I}).$$

Since $E(S) = 0$, it now follows that

$$E(\mathcal{I}) = E(S^2) = \operatorname{var}(S)$$

as required.

PROBLEMS FOR SECTION 11.6

1.† (a) According to the Hardy–Weinberg Law, genotypes AA, Aa, and aa should occur in a population with relative frequencies θ^2, $2\theta(1-\theta)$, and $(1-\theta)^2$, respectively. In an experiment to estimate the gene frequency θ, n randomly chosen individuals are to be examined, and the frequencies Y_1, Y_2, Y_3 with the three genotypes are to be recorded. Find the expected information function of θ.

(b) Suppose that it is very difficult and expensive to distinguish between the Aa and aa genotypes, and that three times as many individuals can be examined if only those with the AA genotype are to be identified. Find the expected information function of θ if $3n$ individuals are to be classified as AA or not AA.

(c) Under what conditions would you recommend doing the experiment in (b) rather than that in (a)?

2. (a) The lifetimes of electronic components are independent and exponentially distributed with mean θ. Suppose that n components are to be tested for a preset time T. The number M which fail and their failure times $Y_1, Y_2, \ldots, Y_M$ are to be recorded. Show that

$$\mathscr{I}_E(\theta) = \frac{n}{\theta^2}(1 - e^{-T/\theta}).$$

Hint: Use the fact that the score function has expected value zero to evaluate $E\{\Sigma Y_i\}$ in terms of $E\{M\}$.

(b) Examine the expected efficiency of the experiment in (a) relative to that described in Example 11.6.1.

(c) For the experiment described in (a), is it better to test $2n$ components for time T, or n components for time $2T$?

3.† Suppose that X, the number of insects on a plant, has a Poisson distribution with mean 100. When an insecticide is applied, each insect on a plant has probability p of surviving, independently of other insects. Two experiments for estimating p are being considered. In the first experiment, both the initial number of insects X_i and the number Y_i which survive the insecticide are to be recorded for n plants. In the second experiment, the initial count is omitted, and only the Y_i's are to be recorded. Show that the expected efficiency of the second experiment relative to the first is $1-p$.

4. Consider a one-to-one parameter transformation from θ to $\lambda = g(\theta)$. Show that the expected information function of λ is given by

$$\left(\frac{d\theta}{d\lambda}\right)^2 \mathscr{I}_E(\theta).$$

5.* In the experiment described in Problem 9.5.5(b), only points of impact on the target are recorded. Show that the expected efficiency of this experiment, relative to one in which all points of impact are recorded, is equal to the probability that a shot misses the target.

*11.7. Bias

Many statistics textbooks suggest that unbiasedness is a desirable property of parameter estimates. Indeed, it is often suggested that one should restrict attention to estimates which are unbiased, and that the "best" estimate is the unbiased estimate having the smallest variance. In this section, unbiased estimates will be defined, and some examples will be given to illustrate their properties.

As in Section 11.1, we suppose that the probability model for an experiment depends upon an unknown parameter θ, and we imagine a series of repetitions of the experiment with θ fixed. Let T be an estimate of θ, such as the maximum likelihood estimate $\hat{\theta}$. The value of T would vary from one repetition to the next, and so we model T as a random variable. Its sampling distribution will depend upon θ, and can be derived from the probability model.

If T has good frequency properties, the values of T obtained in a series of repetitions should be clustered tightly about the value of θ. This means that the sampling distribution of T should be centered near θ, and should have a small spread.

A convenient measure of the "center" of a probability distribution is the mean, $E(T)$. The difference between $E(T)$ and θ is called the *bias* of T:

$$\text{Bias} = E(T) - \theta.$$

T is said to be an *unbiased estimate* of θ if $E(T) = \theta$ for all possible parameter values.

The spread of a probability distribution is usually measured by a second moment. The second moment of the random variable $T-\theta$ is called the *mean squared error* of T:

$$\text{MSE} = E\{(T-\theta)^2\}.$$

If this is small, the estimates obtained in a series of repetitions of the experiment would be clustered about θ. It is sometimes suggested that one should choose an estimate T which minimizes the mean squared error.

If T is unbiased, then $E(T) = \theta$, and hence the mean squared error is equal to the variance of T. An estimate T which is unbiased and has the smallest possible variance is called MVU (minimum variance unbiased).

The following examples illustrate some general results concerning unbiased estimates.

EXAMPLE 11.7.1. If X is the number of successes obtained in n independent trials with success probability θ, the estimate of θ obtained by the method of maximum likelihood is $T \equiv X/n$ (Example 9.1.1). Since $E(X) = n\theta$, we have

*This section may be omitted on first reading.

$$E(T) = \frac{1}{n}E(X) = \frac{1}{n}(n\theta) = \theta,$$

and hence T is an unbiased estimate of θ.

By the invariance property, the maximum likelihood estimate of θ^2 is $T^2 \equiv X^2/n^2$, with expected value

$$E(T^2) = E(X^2)/n^2 = [\text{var}(X) + E(X)^2]/n^2$$

by (5.2.3). Since $\text{var}(X) = n\theta(1-\theta)$, we have

$$E(T^2) = [n\theta(1-\theta) + n^2\theta^2]/n^2 = \theta^2 + \frac{\theta(1-\theta)}{n}.$$

Hence T^2 is not an unbiased estimate of θ^2. The bias is

$$E(T^2) - \theta^2 = \frac{\theta(1-\theta)}{n}$$

which is positive for $0 < \theta < 1$, and tends to zero as $n \to \infty$.

EXAMPLE 11.7.2. If $X_1, X_2, \ldots, X_n$ are independent Poisson variates with mean μ, the maximum likelihood estimate of μ is $\bar{X} \equiv \Sigma X_i/n$ (Example 9.1.2). Since $E(X_i) = \mu$, we have

$$E(\bar{X}) = \frac{1}{n}\Sigma E(X_i) = \frac{1}{n}(n\mu) = \mu,$$

and hence $\bar{X}$ is an unbiased estimate of μ.

By the invariance property, the maximum likelihood estimate of $\beta = e^{-\mu}$ is $e^{-\bar{X}}$. Since $T \equiv \Sigma X_i$ has a Poisson distribution with mean $n\mu$ (corollary to Example 4.6.1), the expected value of $e^{-\bar{X}}$ is

$$E(e^{-T/n}) = \sum_{t=0}^{\infty} e^{-t/n}(n\mu)^t e^{-n\mu}/t! = e^{-n\mu}\Sigma(n\mu e^{-1/n})^t/t!$$

$$= e^{-n\mu} \cdot e^{n\mu e^{-1/n}} = e^{-\mu(n-ne^{-1/n})} = \beta^{n(1-e^{-1/n})}.$$

Hence $e^{-\bar{X}}$ is not an unbiased estimate of β. The bias is

$$\beta^{n(1-e^{-1/n})} - \beta$$

which is always positive, and tends to zero as $n \to \infty$.

EXAMPLE 11.7.3. Suppose that $X_1, X_2, \ldots, X_n$ are independent variates with the same mean μ and variance σ^2. A *linear estimate* of μ is a linear combination of the X_i's, $T \equiv \Sigma a_i X_i$, where the a_i's are constants. Show that the sample mean $\bar{X}$ is the unique MVU linear estimate of μ.

SOLUTION. By (5.5.5) and (5.5.7), the mean and variance of T are

$$E(T) = \mu\Sigma a_i; \qquad \text{var}(T) = \sigma^2\Sigma a_i^2.$$

Thus T is unbiased for μ if and only if $\Sigma a_i = 1$. Now it is easy to show

$$\Sigma a_i^2 = \Sigma(a_i - \bar{a})^2 + n\bar{a}^2$$

that where $\bar{a} = \frac{1}{n}\Sigma a_i$. If $\Sigma a_i = 1$, then $\bar{a} = \frac{1}{n}$ and Σa_i^2 is minimized for $a_1 = a_2 = \cdots = a_n = 1/n$. Hence the unbiased linear estimate of μ with smallest variance is

$$T \equiv \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n \equiv \frac{1}{n}\Sigma X_i \equiv \bar{X}. \qquad \square$$

EXAMPLE 11.7.4. Let X be the number of successes before the first failure in independent trials with success probability θ. Define $T(x) = 0$ for $x = 0$, and $T(x) = 1$ for $x \geq 1$. Show that T is the unique unbiased estimate of θ.

SOLUTION. The distribution of X is geometric:

$$f(x) = \theta^x(1-\theta) \qquad \text{for } x = 0, 1, 2, \ldots.$$

The expected value of T is

$$E(T) = 0 \cdot P(T=0) + 1 \cdot P(T=1)$$

$$= P(X \geq 1) = 1 - f(0) = \theta.$$

Hence T is an unbiased estimate of θ.

Now suppose that T' is another unbiased estimate, and define $U(x) = T'(x) - T(x)$ for $x = 0, 1, \ldots$. Then

$$E(U) = E(T') - E(T) = \theta - \theta = 0 \qquad \text{for all } \theta.$$

Also by (5.1.3) we have

$$E(U) = \Sigma U(x)\theta^x(1-\theta).$$

It follows that

$$U(0) + U(1)\theta + U(2)\theta^2 + U(3)\theta^3 + \cdots = 0$$

for all values of θ between 0 and 1. This will be true if and only if $U(0) = U(1) = U(2) = \cdots = 0$. Hence $T(x) = T'(x)$ for $x = 0, 1, 2, \ldots$, and T is the unique unbiased estimate of θ. $\square$

Discussion

Examples 11.7.1 and 11.7.2 show that the criterion of unbiasedness is not invariant under one-to-one parameter changes. If T is an unbiased estimate of θ, then $g(T)$ will generally not be an unbiased estimate of $g(\theta)$ unless g is a linear transformation. It is not possible to require both unbiasedness and invariance.

If an invariant estimation procedure is used, it does not matter whether we use θ, $1/\theta$, or some other one-to-one function of θ to label the distributions

which make up the probability model. Often the choice of parameter is largely arbitrary, and therefore invariance would seem to be a highly desirable property (see Section 9.6). However, if unbiased estimates are required, a nonlinear parameter transformation completely changes the estimation problem.

Since maximum likelihood estimates are invariant under one-to-one parameter transformations, they will generally be biased. Usually the bias is small and tends to zero as the number of independent observations per unknown parameter increases.

The only way to achieve unbiasedness in Example 11.7.4 is to estimate P(success) by 1 or 0 according to whether the first trial gives success or failure. This is not a very sensible estimation procedure. It ignores information that might be gained in trials beyond the first one, and it either overestimates or underestimates θ in every particular application. This seems a high price to pay in order to achieve the correct long-run average value in a series of repetitions that will never take place!

It may be sensible to require an estimate with small bias in some situations. However, the requirement of unbiasedness is too strong, and as Example 11.7.4 shows, this requirement may eliminate all "sensible" estimation procedures. The choice of statistical terminology is rather unfortunate. No one likes to be accused of bias, but in parameter estimation, a little bias may be a good thing.

There is an extensive literature on the theory of unbiased estimation. Although this is of some mathematical interest, it does not seem terribly relevant from a practical point of view.

The suggestion that estimates should be chosen to minimize mean squared error or the variance of the sampling distribution also deserves some comment. The variance or mean squared error of an estimate in a hypothetical series of repetitions does not necessarily indicate the precision of the estimate in any particular application. In maximum likelihood estimation it is the observed information $\mathscr{I}(\hat{\theta})$, and not the variance or mean squared error of $\hat{\theta}$, which measures precision. There are special situations where the variance of the sampling distribution is an appropriate measure of precision, but this is not true in general.

Problems for Section 11.7

1. Let T be an estimate of θ. Show that the mean squared error of T is equal to its variance plus the square of its bias.

2.† Suppose that $Y \sim N(\mu, 1)$, and consider the following estimates of μ^2:

$$T_1 \equiv Y^2; \qquad T_2 \equiv Y^2 - 1.$$

Show that T_2 is unbiased and has smaller mean squared error than T_1. Why would T_2 be unsatisfactory as an estimate of μ^2?

3. In Example 11.7.4, show that the bias of the maximum likelihood estimate $\hat{\theta}$ is

$$(1-\theta)\left[1+\frac{1}{\theta}\log(1-\theta)\right].$$

Plot the bias, and show that it tends to zero as $\theta \to 0$ and as $\theta \to 1$.

4. Suppose that Y has a binomial (n, θ) distribution. Show that $T \equiv \dfrac{Y(Y-1)}{n(n-1)}$ is an unbiased estimate of μ^2, and more generally, that $Y^{(k)}/n^{(k)}$ is an unbiased estimate of μ^k for $k = 1, 2, \ldots$.

5. Let $Y_1, Y_2, \ldots, Y_n$ be independent Poisson variates with mean μ, and define $S \equiv Y_1 + Y_2 + \cdots + Y_n$. Show that $S^{(k)}/n^{(k)}$ is an unbiased estimate of μ^k for $k = 1, 2, \ldots$.

6. Suppose that $X_1, X_2, \ldots, X_n$ are independent variates with the same mean μ and variance σ^2. Show that

$$\Sigma(X_i - \bar{X})^2 \equiv \Sigma X_i^2 - n\bar{X}^2$$

and hence verify that

$$S^2 \equiv \frac{1}{n-1}\Sigma(X_i - \bar{X})^2$$

is an unbiased estimate of σ^2. Is S an unbiased estimate of σ?

7.† Suppose that $X_1, X_2, \ldots, X_n$ are independent variates with the same mean μ but with different variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. Find the minimum variance linear unbiased estimate of μ.

CHAPTER 12

Tests of Significance

A test of significance is a procedure for evaluating the strength of the evidence provided by the data against an hypothesis. Section 1 gives a general introduction to significance tests and their interpretation, and defines test statistics and significance levels.

In many applications, the hypothesis of interest can be formulated as an hypothesis concerning the values of unknown parameters in the probability model. It is then possible to derive a test statistic, called the likelihood ratio statistic, from the log likelihood function. Likelihood ratio tests are described in Sections 2 and 3.

Sections 4, 5, 6, and 8 give applications of significance tests to examples involving frequency data, where the basic model for the experiment is binomial or multinomial. In particular, Section 5 discusses goodness of fit tests for multinomial data, and Section 6 describes tests for independence in contingency tables. Section 7 is concerned with the importance of controlled experiments and randomization in establishing cause and effect.

Significance intervals or regions are defined in Section 9, and their coverage probabilities are determined. Also, the connection between significance intervals and likelihood regions is investigated.

The power of a test statistic against an alternative hypothesis is defined in Section 10. Power is sometimes useful in a theoretical comparison of two or more possible test statistics, or in selecting the sample size for an experiment.

12.1. Introduction

A test of significance is a procedure for measuring the strength of the evidence provided by the data against an hypothesis H. It is similar to a proof by contradiction in mathematics. In each case we assume the hypothesis to be

true and then check whether this assumption leads to an inconsistency. If a contradiction is found, the hypothesis is disproved. If no contradiction is found, the method of proof fails and the hypothesis could be either true or false.

For instance, to prove by contradiction that there is no largest prime number, we first formulate the hypothesis

H: there is a largest prime number

which is the opposite to what we want to prove. Assuming H to be true, there are finitely many prime numbers $p_1 < p_2 < \cdots < p_n$. If this is so, every number larger than p_n is composite, and is divisible by at least one of $p_1, p_2, \ldots, p_n$. However, $p = 1 + p_1 p_2 \cdots p_n$ is larger than p_n and is not divisible by any of the p_i's. This is a contradiction, and therefore H is false.

In a mathematical proof by contradiction, we look for a logical inconsistency, but in statistical applications there will rarely be a logical inconsistency between data and hypothesis. Even if we observed 100 heads in 100 tosses of a coin, we could not prove mathematically that the coin was biased, because this result could have arisen from 100 tosses of a balanced coin. Nevertheless, we would be quite sure that the coin was biased, because the probability of obtaining such an extreme result with a balanced coin is extremely small.

In a significance test, we compute the probability of observing such an extreme result when the hypothesis is true. The smaller the probability, the stronger the evidence that the hypothesis is false.

EXAMPLE 12.1.1. Let X be the number of heads in 100 tosses of a coin. We assume that tosses are independent, and that θ, the probability of heads, is the same at all trials. We observe a value of X, and we wish to test the hypothesis H: $\theta = \frac{1}{2}$.

Under the hypothesis, X has a binomial distribution with probability function

$$f(x) = \binom{100}{x} (\tfrac{1}{2})^{100} \qquad \text{for } x = 0, 1, \ldots, 100.$$

If H is true we expect to observe a value of X near 50. The quantity $D \equiv |X - 50|$ measures how closely the observation agrees with the hypothesis. If D is close to 0, then X is in good agreement with H: $\theta = \frac{1}{2}$. A large value of D indicates poor agreement between the data and hypothesis.

Suppose that we observe $X = 35$, so that the observed value of D is $|35 - 50| = 15$. The probability of getting such poor agreement with H (i.e. such a large value of D) is

$$P\{D \geq 15\} = P\{|X - 50| \geq 15\} \approx 0.0027$$

(see below). If H were true, a result as extreme as $X = 35$ would occur very rarely. Thus we have strong evidence that H is false and the coin is biased.

On the other hand, if we observe $X = 45$, the observed value of D is

$|45 - 50| = 5$, and the probability of such poor agreement with H is

$$P\{D \geq 5\} = P\{|X - 50| \geq 5\} \approx 0.32.$$

Results as extreme as $X = 45$ would occur fairly often with a balanced coin, and we do not have evidence that H is false. We have not shown that H is true either! There are plenty of other values of θ, such as $\theta = 0.45$, which could have produced $X = 45$. A large probability means simply that no contradiction has been found. The method of proof fails, and H could be either true or false.

The above probabilities could be calculated exactly by summing $f(x)$ over the appropriate values of X. Instead, the normal approximation to the binomial distribution was used (see Section 6.8). Under H: $p = \frac{1}{2}$ we have

$$X \sim \text{bin}(100, \tfrac{1}{2}) \approx N(50, 25),$$

so that $(X - 50)/5$ is approximately $N(0, 1)$. Thus

$$P\{|X - 50| \geq 15\} = P\left\{\left|\frac{X - 50}{5}\right| \geq 3\right\} \approx P\{|Z| \geq 3\}$$

where $Z \sim N(0, 1)$. Now Table B2 gives

$$P\{|Z| \geq 3\} = 2(1 - 0.998650) = 0.0027.$$

Test Statistics and Significance Levels

For a test of significance we require a ranking of possible outcomes according to how closely they agree with the hypothesis. This ranking is usually specified by defining a *test statistic* D, also called the test criterion or discrepancy measure. A small value of D shows close agreement between the outcome and the hypothesis, and a large value of D indicates poor agreement.

The test statistic is to be chosen before the data are examined, and the choice will reflect the type of departure from the hypothesis that we wish to detect. A general method for constructing test statistics from the likelihood function will be described in the next two sections. Power comparisons may help in choosing among several possible test statistics (see Section 12.10).

When the experiment has been performed and data have been obtained, we can compute the observed value of D. Then, assuming H to be true, we compute the probability of obtaining a value of D at least as great as that observed. This probability is called the *significance level* (SL), or P-value, of the data in relation to the hypothesis:

$$\text{SL} = P\{D \geq D_{\text{obs}} | H \text{ is true}\}.$$

The significance level is the probability of observing such poor agreement between the hypothesis and data if the hypothesis is true.

If SL is very small, then such poor agreement would almost never occur if

the hypothesis were true, and we have evidence that H is false. The smaller the significance level, the stronger the evidence against the hypothesis. A large SL indicates only a lack of evidence against the hypothesis. Even a significance level of 90% or 100% does not imply that the hypothesis is "probably true". The probability statement refers to the data, not the hypothesis.

Conventionally, 0.05 is taken to be the dividing line between "small" and "large" significance levels. If $\text{SL} \leq 0.05$, the hypothesis is said to be contradicted by the data (at the 5% level), whereas if $\text{SL} > 0.05$, the hypothesis is said to be consistent or compatible with the data (at the 5% level). Of course, this convention should not be taken too seriously. Significance levels 0.049 and 0.051 are on opposite sides of 0.05, but they imply about the same strength of evidence against the hypothesis.

EXAMPLE 12.1.2 (Test for ESP). Consider a possible experiment for detecting ESP (extra-sensory perception) in a human subject. Four cards labeled A, B, C, and D are shuffled and placed face down on a table. The subject attempts to match the hidden letters to envelopes marked a, b, c, and d, and the number of correct matches is recorded. The experiment is to be repeated 50 times altogether.

Even if the subject has no special powers, some correct matches will occur by chance. A subject with ESP should be able to achieve more correct matches than would occur by chance alone. To determine whether there is evidence for ESP, we compare the results obtained with what would be expected under H, the hypothesis that the subject has no ESP. If the observed results are in reasonable agreement with H, then we cannot claim to have proof of ESP.

Let T denote the total number of correct guesses in 50 repetitions. We shall show below that, under the hypothesis of no ESP, T has approximately a normal distribution with mean 50 and variance 50. Large values of T will be interpreted as evidence against H and in favor of ESP, so we take the test statistic to be $D \equiv T$. The significance level is then

$$\text{SL} = P(T \geq T_{\text{obs}}) \approx P\left(Z \geq \frac{T_{\text{obs}} - 50}{\sqrt{50}}\right)$$

where Z has a standardized normal distribution.

For instance, suppose that such an experiment produced the following data:

No. of correct matches	0	1	2	4	Total
Frequency observed	17	18	9	6	50

The total number of correct matches is

$$T_{\text{obs}} = 0 \times 17 + 1 \times 18 + 2 \times 9 + 4 \times 6 = 60,$$

and hence the significance level is

$$P(T \geq 60) \approx P(Z \geq 1.41) = 0.079.$$

If the subject has no ESP, there is about an 8% chance of getting such a large number of correct guesses. Thus we cannot claim to have proved that the subject has ESP. Nevertheless the results are encouraging, and one might wish to collect additional data for this subject.

To complete the example, we show that the distribution of T under the hypothesis is approximately normal with mean 50 and variance 50. Let X_i be the number of correct guesses in the ith repetition, so that

$$T \equiv X_1 + X_2 + \cdots + X_{50}.$$

Since the cards are randomly rearranged in each repetition, the X_i's are independent and identically distributed under H, and the distribution of X_i is as follows:

x = Number correct	0	1	2	4	Total
$f(x)$ = Probability	9/24	8/24	6/24	1/24	1

(see Problem 1.3.1). The mean and variance of this distribution are

$$E(X_i) = \Sigma x f(x) = 1;$$

$$\text{var}(X_i) = \Sigma (x-1)^2 f(x) = 1.$$

The mean of T is then $\Sigma E(X_i) = 50$. Since the X_i's are independent, the variance of T is $\Sigma \text{ var}(X_i) = 50$. The distribution of T is approximately normal by the Central Limit Theorem.

Note. All that we needed in this example was the distribution of T under the hypothesis H of no ESP. We didn't need to know what the distribution of T would be if H were false. A model which incorporates the possibility of ESP is likely to be quite complex. The test of significance tells us that so far we don't have conclusive evidence that ESP even exists, so we'd probably be wasting our time if we tried to model it at this stage. One of the most important uses of significance tests is in helping us to avoid wasting time with complicated models when a simple one will do.

Tests Suggested by the Data

Examination of a set of data will often reveal an interesting pattern which had not been anticipated before the experiment was performed. There is then a temptation to design a test of significance which will "prove" that this pattern could not have arisen by chance.

For example, upon examining the data in Example 12.1.2, we see that all four cards were correctly matched in 6 cases out of 50. Under the hypothesis of no ESP, the probability of matching all four is only 1/24, and the

probability that all four are correctly matched y times is

$$g(y) = \binom{50}{y}\left(\frac{1}{24}\right)^y\left(\frac{23}{24}\right)^{50-y} \qquad \text{for } y = 0, 1, \ldots, 50.$$

The probability of at least 6 correct matches is

$$P(Y \geq 6) = 1 - g(0) - g(1) - \cdots - g(5) = 0.017.$$

This looks like fairly strong evidence that the subject has ESP.

Of course, since the test was performed because we had noticed a large value of Y, we should not be surprised that a small significance level was obtained! One can find something unusual about almost any set of data, and then devise a test to produce a small significance level. In such situations, a small significance level proves nothing.

For a valid statistical proof, one must specify the type of discrepancy being sought *before* the data are examined. Tests suggested by the data may be useful for indicating questions to be investigated in future experiments, but they do not "prove" anything by themselves.

Detection Versus Estimation

A small significance level shows that there is a "real" departure from the hypothesis; that is, a departure which cannot readily be explained by chance. A small significance level does not mean that the departure is necessarily large or important. With a large amount of data, very small departures of no practical importance may be detected by the test. With a small sample, large and important departures may go undetected.

Merely reporting a small significance level is not enough. We also need to describe the type of departure observed and estimate its size. For instance, in Example 12.1.1, one should not report just the significance level from the test $H: \theta = \frac{1}{2}$. One should also give the MLE and a likelihood or confidence interval for θ.

Problems for Section 12.1

1.† Of 100 peas planted in a genetics experiment, 65 produced tall plants and 35 produced short plants. According to genetic theory, plants are independent and the probability of a tall plant is $\frac{3}{4}$. Carry out a test of significance to investigate whether the theory is consistent with the data.

2. In a particular Ontario county, a very large number of people are eligible for jury duty, and half of these are women. The judge is supposed to prepare a jury list by randomly selecting individuals from all those eligible. In an important 1974 murder trial, the jury list of 82 people contained 58 men and 24 women. Could such an extreme imbalance in the sexes reasonably have occurred by chance?

3.†In research on drugs to counteract the intoxicating effect of alcohol, twenty subjects were used to compare the relative merits of benzedrine and caffeine. Each subject received the drugs in a random order on two different occasions far enough apart to eliminate carry-over effects. Benzedrine brought about the more rapid recovery in 14 subjects, while caffeine was judged better in the other 6 cases. Are these results consistent with the hypothesis that the two drugs are equally good?

4. In January, party A won 53% of a very large number of votes in an election. Six months later, a poll of 200 randomly selected voters showed that only 48% would vote for party A if another election were called. Could these results reasonably be due to chance, or is there evidence of a real change in the support for party A?

5.†A seed merchant states that 80% of the seeds of a certain plant will germinate. Each of 4 customers buys one packet and sows 100 seeds from it. The numbers of seedlings appearing are 73, 76, 74, and 77.

(a) Discuss whether any customer, on the basis of his observation only, has adequate cause to claim that the stated germination rate is erroneous.
(b) If the 4 packets are from a homogeneous stock of seed, is the total germination record consistent with the stated rate?

6. (a) Each of 25 individuals was given two similar glasses, one of Pepsi and one of Coke, and was asked to pick the one he preferred. Sixty percent of them picked Coke. Is this result consistent with the hypothesis that there is no detectable difference between Pepsi and Coke?
(b) Repeat (a) if 250 individuals were tested and 60% of them picked Coke.

7. Determine the approximate significance level of each of the following observations in relation to the hypothesis that p, the probability of a male birth, is equal to $\frac{1}{2}$. Find an approximate 95% confidence interval for p in each case.

(a) 293 girls and 299 boys in 592 births;
(b) 2930 girls and 2990 boys in 5920 births;
(c) 29300 girls and 29900 boys in 59200 births.

8. The table below summarizes the data from 100 repetitions of the ESP experiment described in Example 12.1.2.

No. of correct matches	0	1	2	4	Total
Observed frequency	28	33	31	8	100

Test whether the total number of correct matches is significantly greater than would be expected if cards were turned over in a random order.

9.†In an experiment studying the relationship between color perception and order, a psychologist asks young children to place 6 similar blocks in a row. Four of the blocks are red and two are green, but otherwise the blocks are identical. The number of red blocks between the two green blocks is recorded, and the observed frequencies in 100 repetitions of the experiment are as follows:

Number of red blocks	0	1	2	3	4	Total
Frequency observed	28	22	22	18	10	100

(a) Suppose that the blocks are placed in a random order. Tabulate the probability function of X, the number of red blocks placed between the two green blocks, and find the mean and variance of X.
(b) Let $\bar{X} \equiv (X_1 + X_2 + \cdots + X_{100})/100$ be the average number of red blocks between green blocks in 100 replications of the experiment. Is the observed $\bar{X}$ significantly different from what one would expect under random placement?

10. In an experiment on human behavior, a sociologist asks four men and four women to enter a room and sit wherever they wish at a rectangular table. There are three chairs at each side of the table and one at each end. The two end seats are considered to be special in that people sitting there have more dominant positions at the table.

(a) Find the mean and variance of X, the number of men occupying end seats, under the assumption that people choose their seats at random.
(b) The seating experiment was repeated 84 times, and altogether there were 98 men in the end seats and only 70 women. Do these results differ significantly from what one would expect under random seating?

11.†Under normal conditions, the mean number of personal calls handled by a company switchboard was 7.2 per hour. The manager sent a letter to all employees requesting that the number of personal calls be reduced. During five one-hour periods the following week, the numbers of personal calls were 4, 2, 7, 5, and 3. Do these observations give strong evidence that the mean number of personal calls per hour has been reduced?

12. A food taster is given 12 samples of natural flavoring and 12 samples of synthetic flavoring in a random order. He is asked to identify the 12 samples of natural flavoring, and manages to get 8 of them right. Test the hypothesis that the taster is unable to distinguish between the two flavors.

13. An experiment is carried out to investigate whether subjects can tell the difference between butter and margarine. Each subject is blindfolded and receives two samples of butter and two of margarine in a random order. The subject is asked to identify the two samples of butter. The following table shows the number of correct butter identifications in 100 independent replications of the experiment.

Number correct	0	1	2
Frequency observed	18	62	20

Altogether there were 102 correct identifications. Using a test statistic based on the total number correct, test the hypothesis that subjects are unable to distinguish between butter and margarine.

12.2. Likelihood Ratio Tests for Simple Hypotheses

In many applications, the hypothesis to be tested can be formulated as an hypothesis concerning the values of unknown parameters in the probability model. A test statistic D, called the *likelihood ratio statistic*, can then be

derived from the log likelihood function. A significance test in which the likelihood ratio statistic is used as test statistic is called a *likelihood ratio test.*

In this section we shall restrict the discussion to simple hypotheses. H is called a *simple hypothesis* if it specifies numerical values for all of the unknown parameters in the model. A simple hypothesis reduces the number of unknown parameters in the model to zero. A *composite hypothesis* is one which reduces the number of unknown parameters in the model, but not to zero. Thus, under a composite hypothesis, there will still be one or more parameters which require estimation from the data. Likelihood ratio tests for composite hypotheses will be discussed in the next section.

One-Parameter Case

First suppose that the probability model involves a single unknown parameter θ. We wish to test the hypothesis $H: \theta = \theta_0$, where θ_0 is a particular numerical value. For instance, in Example 12.1.1, θ was the success probability in Bernoulli trials, and the hypothesized value was $\theta_0 = \frac{1}{2}$. H is a simple hypothesis because it specifies a numerical value for the only unknown parameter.

Let $l(\theta)$ denote the log likelihood function and $\hat{\theta}$ the MLE of θ under the model. The maximum log likelihood under the model is $l(\hat{\theta})$. The log likelihood under the hypothesis is $l(\theta_0)$. The likelihood ratio statistic for testing $H: \theta = \theta_0$ is defined to be twice the difference between these two log likelihoods,

$$D = 2[l(\hat{\theta}) - l(\theta_0)] = -2r(\theta_0), \tag{12.2.1}$$

where $r(\theta)$ is the log relative likelihood function of θ.

Since $\hat{\theta}$ maximizes $l(\theta)$, we have $l(\theta) \leq l(\hat{\theta})$ for all values of θ, and therefore $D \geq 0$ (see Section 11.1). A small value of D means that the outcome of the experiment is such that θ_0 is a likely parameter value. A large value of D means that the outcome is such that θ_0 is unlikely. Thus D ranks possible outcomes of the experiment according to how well they agree with $H: \theta = \theta_0$.

Taking D as the test statistic, we have

$$\begin{aligned} \text{SL} &= P\{D \geq D_{\text{obs}} | H \text{ is true}\} \\ &= P\{D \geq D_{\text{obs}} | \theta = \theta_0\}, \end{aligned} \tag{12.2.2}$$

where D_{obs} is the observed value of D. The significance level is calculated from the (sampling) distribution of the likelihood ratio statistic when $\theta = \theta_0$. If we imagine a series of repetitions of the experiment with θ fixed at θ_0, SL is the fraction of the time that the test statistic D would be greater than or equal to the observed value D_{obs}.

Upon comparing (12.2.2) with (11.2.2), we see that there will be a close connection between significance levels in likelihood ratio tests and coverage probabilities of likelihood intervals. The relationship between significance tests and likelihood/confidence intervals will be considered in Section 12.9.

In some simple examples, it is possible to derive the exact sampling distribution of D when $\theta = \theta_0$ (see Section 11.1). More often, derivation of the exact distribution is too difficult, and an approximation is used. We noted in Section 11.3 that, under suitable conditions, the distribution of D when $\theta = \theta_0$ is well approximated by a χ^2 distribution with one degree of freedom. When this approximation is applicable, we have

$$\text{SL} = P\{D \geq D_{\text{obs}} | \theta = \theta_0\} \approx P\{\chi^2_{(1)} \geq D_{\text{obs}}\}. \tag{12.2.3}$$

We can find approximate significance levels by using Table B4, or by using (6.9.8) and Table B2.

See Section 11.3 for a discussion of the conditions under which the χ^2 approximation applies, and for some examples in which the accuracy of this approximation is investigated.

EXAMPLE 12.2.1. Suppose that we observe X, the number of successes in n Bernoulli trials with success probability θ. We want to test the hypothesis $H: \theta = \theta_0$ where θ_0 is a particular numerical value such as $\frac{1}{2}$.

The distribution of X is binomial (n, θ), and the log likelihood function of θ is

$$l(\theta) = x \log \theta + (n - x) \log (1 - \theta)$$

for $0 < \theta < 1$. The MLE is $\hat{\theta} = x/n$, and the maximum log likelihood is

$$l(\hat{\theta}) = x \log \frac{x}{n} + (n - x) \log \frac{n - x}{n}.$$

The log likelihood under H is

$$l(\theta_0) = x \log \theta_0 + (n - x) \log (1 - \theta_0),$$

and so the likelihood ratio statistic for testing $H: \theta = \theta_0$ is

$$\begin{aligned} D &= 2[l(\hat{\theta}) - l(\theta_0)] = -2r(\theta_0) \\ &= 2x \log \frac{x}{n\theta_0} + 2(n - x) \log \frac{n - x}{n(1 - \theta_0)}. \end{aligned}$$

If n is large, the distribution of D is approximately χ^2 with one degree of freedom (see Example 11.3.1). If θ_0 is near $\frac{1}{2}$, the approximation gives fairly accurate results for $n = 20$. However, a much larger value of n is needed when θ_0 is close to 0 or 1.

For instance, suppose that we observe $X = 35$ in $n = 100$ trials and wish to test $H: \theta = \frac{1}{2}$ as in Example 12.1.1. The likelihood ratio statistic for testing $H: \theta = \frac{1}{2}$ is

$$D = 2x \log \frac{x}{50} + 2(100 - x) \log \frac{100 - x}{50},$$

and its observed value is

$$D_{\text{obs}} = 70 \log \frac{35}{50} + 130 \log \frac{65}{50} = 9.14.$$

The significance level is

$$\text{SL} = P\{D \geq 9.14 | \theta = \tfrac{1}{2}\}$$
$$\approx P\{\chi_{(1)} \geq 9.14\} = 0.0025$$

from Table B4. If θ were equal to $\frac{1}{2}$, a result as extreme as $X = 35$ would very rarely occur, and so there is strong evidence that $\theta \neq \frac{1}{2}$.

Any other hypothesized value for θ can be tested in a similar way. For instance, the LR statistic for testing the hypothesis $H: \theta = 0.4$ is

$$D = 2x \log \frac{x}{40} + 2(100 - x) \log \frac{100 - x}{60}$$

and its observed value is

$$D_{\text{obs}} = 70 \log \frac{35}{40} + 130 \log \frac{65}{60} = 1.06.$$

Table B4 gives

$$\text{SL} \approx P\{\chi^2_{(1)} \geq 1.06\} = 0.303,$$

so the hypothesis $\theta = 0.4$ is compatible with the data.

To find the exact significance level, it is necessary to add up the binomial probabilities of all outcomes x such that $D \geq D_{\text{obs}}$. Taking $\theta = 0.4$, we find that $D \geq 1.06$ for $x \leq 35$ and for $x \geq 46$. Thus we have

$$\text{SL} = P\{D \geq D_{\text{obs}}\} = P(X \leq 35) + P(X \geq 46) = 1 - P(36 \leq X \leq 45).$$

Under $H: \theta = 0.4$, the distribution of X is binomial (100, 0.4), and therefore

$$\text{SL} = 1 - \sum_{x=36}^{45} \binom{100}{x} (0.4)^x (0.6)^{100-x} = 0.311.$$

Similarly, the exact significance level in relation to the hypothesis $\theta = \frac{1}{2}$ is found to be

$$\text{SL} = 1 - \sum_{x=36}^{64} \binom{100}{x} (0.5)^x (0.5)^{100-x} = 0.0035.$$

There is good agreement between the approximate and exact results.

Two or More Parameters

Suppose that the probability model depends on a vector of unknown parameters θ. Let θ_0 be a vector of numerical values, one for each component of θ. Then $H: \theta = \theta_0$ is a simple hypothesis because it specifies a numerical value for each of the unknown parameters.

The likelihood ratio statistic for testing $H: \theta = \theta_0$ is twice the difference between the maximum log likelihood and the log likelihood under H. Thus we have

$$D = 2[l(\hat{\theta}) - l(\theta_0)] = -2r(\theta_0)$$

where now $\hat{\theta}$ is a vector of MLE's and $r(\theta)$ is the joint log relative likelihood function. The exact significance level in the likelihood ratio test of $H: \theta = \theta_0$ is again given by (12.2.2).

It can be shown that, under conditions similar to those described in Section 11.3, the distribution of D when $\theta = \theta_0$ can be approximated by a χ^2 distribution. *The degrees of freedom for the χ^2 approximation is equal to the number of functionally independent unknown parameters in the model* (see the example below). It follows that

$$\text{SL} = P\{D \geq D_{\text{obs}} | \theta = \theta_0\} \approx P\{\chi^2_{(k)} \geq D_{\text{obs}}\} \tag{12.2.4}$$

where k is the number of functionally independent unknown parameters in the model.

The case of two unknown parameters, $\theta = (\alpha, \beta)$, was considered previously in Section 11.5. The likelihood ratio statistic for testing $H: \alpha = \alpha_0$ and $\beta = \beta_0$ is

$$D = 2[l(\hat{\alpha}, \hat{\beta}) - l(\alpha_0, \beta_0)] = -2r(\alpha_0, \beta_0).$$

We noted in Section 11.5 that the distribution of D is approximately $\chi^2_{(2)}$. D has exactly a $\chi^2_{(2)}$ distribution in the normal distribution Examples 11.5.1 and 10.1.1.

EXAMPLE 12.2.2. The following are the observed frequencies of the six faces in 100 rolls of a die from Example 1.4.1:

Face j	1	2	3	4	5	6	Total
Obs.freq. f_j	16	15	14	20	22	13	100

Are these observations consistent with the hypothesis that the die is balanced?

SOLUTION. Assuming that rolls of the die are independent, the distribution of the f_j's is multinomial, with joint probability function

$$(f_1 f_2^n \ldots f_6) p_1^{f_1} p_2^{f_2} \ldots p_6^{f_6}$$

where $\Sigma p_j = 1$ and $\Sigma f_j = n = 100$. The hypothesis to be tested is

$$H: p_1 = p_2 = \cdots = p_6 = \tfrac{1}{6}.$$

This is a simple hypothesis because it assigns a numerical value to each of the unknown parameters $p_1, p_2, \ldots, p_6$.

The log likelihood function is

$$l(p_1, p_2, \ldots, p_6) = \Sigma f_j \log p_j$$

where the p_j's are non-negative, and $\Sigma p_j = 1$. It can be shown that the MLE's are given by $\hat{p}_j = f_j/n$ (see Section 12.5). Hence the maximum log likelihood is

$$l(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_6) = \Sigma f_j \log \hat{p}_j = \Sigma f_j \log (f_j/n).$$

Under H, the log likelihood is

$$l(\tfrac{1}{6}, \tfrac{1}{6}, \ldots, \tfrac{1}{6}) = \Sigma f_j \log \tfrac{1}{6}.$$

The likelihood ratio statistic for testing H is

$$D = 2[l(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_6) - l(\tfrac{1}{6}, \tfrac{1}{6}, \ldots, \tfrac{1}{6})].$$

D is large whenever the f_j's are such that the set of hypothesized parameter values $(\frac{1}{6}, \frac{1}{6}, \ldots, \frac{1}{6})$ is unlikely.

Substituting the observed f_j's from above gives

$$D_{\text{obs}} = 2[-177.326 + 179.176] = 3.70.$$

By (12.2.4), the significance level is

$$\text{SL} = P\{D \geq 3.70 | H \text{ is true}\} \approx \{\chi^2_{(k)} \geq 3.70\}$$

where k is the number of functionally independent parameters in the model.

There are six unknown parameters $p_1, p_2, \ldots, p_6$. However, since $\Sigma p_j = 1$, these are not functionally independent. Only five of the p_j's are free to vary, and then the sixth is determined by the condition $\Sigma p_j = 1$. Thus there are just five functionally independent parameters, and the χ^2 approximation will have $k = 5$ degrees of freedom.

Table B4 now gives

$$\text{SL} \approx P\{\chi^2_{(5)} \geq 3.70\} \geq 0.5.$$

The observed value of D is certainly not unusually large, and hence there is no evidence against the hypothesis that the die is balanced.

The exact significance level is a sum of multinomial probabilities:

$$\text{SL} = P\{D \geq 3.70 | H \text{ is true}\}$$
$$= \Sigma \binom{100}{f_1 f_2 \ldots f_6} \left(\frac{1}{6}\right)^{100}.$$

The sum is taken over all sets of frequencies $\{f_j\}$ with $\Sigma f_j = 100$ such that $D \geq 3.70$. Much arithmetic is needed to determine the appropriate sets of frequencies $\{f_j\}$, although the calculations are certainly feasible on a high-speed computer.

Alternatively, one could simulate the experiment a large number of times on a computer and determine the fraction of the time that D is greater than or equal to 3.70. This gives an estimate of SL which can be made as precise as desired by increasing the number of simulations.

Neither of these procedures is really necessary in this example. The χ^2 approximation is quite accurate, and the conclusion will not be affected by a small change in the computed value of the significance level. □

Problems for Section 12.2

1.† Let θ be the probability of a tall plant in the genetics experiment of Problem 12.1.1. Perform an approximate likelihood ratio test of the hypothesis $\theta = 3/4$.

2. Suppose that $X = 5$ successes are observed in $n = 10$ Bernoulli trials with success probability θ. Perform an exact likelihood ratio test of the hypothesis $\theta = 3/4$. Would the same significance level be obtained if $|X - 7.5|$ were used as the test statistic?

3. Consider a sequence of Bernoulli trials with success probability θ. Trials are continued until the 5th success has occurred, and it is observed that altogether 10 trials are needed. Carry out an exact likelihood ratio test of the hypothesis $\theta = 3/4$.

4. The seating experiment of Problem 12.1.10 is repeated 28 times using new subjects each time. The following table shows the numbers of times that the two end seats were occupied by two men, by two women, and by a man and a woman:

Occupants of end seats	MM	FF	MF or FM
Frequency observed	10	4	14

Test the hypothesis that the probabilities for the three classes are $\frac{3}{14}, \frac{3}{14}$, and $\frac{8}{14}$.

5.† Seeds from a variety of pea plant are classified as round or angular, and as green or yellow, so that there are four possible seed types: RY, RG, AY, and AG. The following are the observed frequencies of the four types in 556 seeds:

Pea type	RY	RG	AY	AG
Frequency	315	108	101	32

Test the hypothesis that the probabilities of the four types are $\frac{9}{16}, \frac{3}{16}, \frac{3}{16}$, and $\frac{1}{16}$, respectively, as predicted by Mendelian theory.

6. In a long-term study of heart disease in a large group of men, it was noted that 65 men who had no previous record of heart problems died suddenly of heart attacks. The following table shows the number of such deaths recorded on each day of the week.

Day of week	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Sun.
No. of deaths	22	7	6	13	5	4	6

Test the significance of these data in relation to the hypothesis that deaths are equally likely to occur on any day of the week.

7. (a) Let $X_1, X_2, \ldots, X_n$ be IID Poisson variates with mean μ. Derive the likelihood ratio statistic for testing $H: \mu = \mu_0$.
   (b) Prior to the installation of a traffic signal, there were 6 accidents per month (on the average) at a busy intersection. In the first year following the installation there were 53 accidents. Using an approximate likelihood ratio test, determine whether there is evidence that the accident rate has changed.

8.†(a) Let $X_1, X_2, \ldots, X_n$ be independent exponential variates with mean θ. Derive the likelihood ratio statistic for testing $H: \theta = \theta_0$.
   (b) Survival times for patients treated for a certain disease may be assumed to be exponentially distributed. Under the standard treatment, the expected survival is 37.4 months. Ten patients receiving a new treatment survived for the following times (in months):

99	8	30	6	53
60	44	12	105	17

   (i) Are these data consistent with a mean survival time of 37.4 months?
   (ii) The doctor who developed the new treatment claims that it gives a 50% increase in mean survival time. Are the data consistent with this claim?
   (iii) Obtain a likelihood interval which is an approximate 95% confidence interval for the mean survival time under the new treatment.

9. (a) Let $X_1, X_2, \ldots, X_n$ be IID normal variates with known standard deviation σ and unknown mean μ. Derive the likelihood ratio statistic for testing the hypothesis $H: \mu = \mu_0$.
   (b) The measurement errors associated with a set of scales are independent normal with known standard deviation $\sigma = 1.3$ grams. Ten weighings of an unknown mass μ give the following results (in grams):

227.1	226.8	224.8	228.2	225.6
229.7	228.4	228.8	225.9	229.6

   (i) Perform likelihood ratio tests of the hypothesis $\mu = 226$, and the hypothesis $\mu = 229$.
   (ii) For which parameter values μ_0 does a likelihood ratio test of $H: \mu = \mu_0$ give a significance level of 5% or more?

10. Let $X_1, X_2, \ldots, X_n$ be independent normal variates with known variances $v_1, v_2, \ldots, v_n$ and the same unknown mean μ. Show that the likelihood ratio statistic for testing $H: \mu = \mu_0$ is

$$D \equiv (\hat{\mu} - \mu_0)^2 \Sigma v_i^{-1},$$

where $\hat{\mu} \equiv (\Sigma X_i v_i^{-1})/\Sigma v_i^{-1}$. Show that, if H is true, the distribution of D is exactly $\chi^2_{(1)}$.

12.3. Likelihood Ratio Tests for Composite Hypotheses

In this section we extend the discussion of likelihood ratio tests to include composite hypotheses as well as simple hypotheses.

Suppose that the basic probability model for the experiment depends upon a vector of unknown parameters θ, and consider an hypothesis H concerning the value of θ. Together, the basic model and hypothesis determine the hypothesized model.

Let k denote the number of functionally independent unknown parameters in the basic probability model, and let q denote the number of functionally independent unknown parameters which remain in the hypothesized model. In general, it is not possible to test an hypothesis H unless it produces a real simplification in the model, so that $q < k$.

A simple hypothesis specifies numerical values for all of the unknown parameters in the basic probability model. Thus there are no unknown parameters in the hypothesized model, and so $q = 0$ for a simple hypothesis. A composite hypothesis does not completely eliminate the unknown parameters, and so $q > 0$ for a composite hypothesis.

Let $l(\theta)$ denote the log likelihood function of θ under the basic model. Let $\hat{\theta}$ be the MLE under the basic model, so that $l(\hat{\theta}) \geq l(\theta)$ for all possible values of θ. The maximum log likelihood under the basic model is $l(\hat{\theta})$.

Next let $\tilde{\theta}$ denote the MLE of θ under the hypothesized model. The maximum log likelihood under the hypothesis is $l(\tilde{\theta})$. Since $l(\hat{\theta}) \geq l(\theta)$ for all possible values of θ, we have $l(\hat{\theta}) \geq l(\tilde{\theta})$. The restricted maximum of $l(\theta)$ under the hypothesis cannot exceed the unrestricted maximum of $l(\theta)$.

The *likelihood ratio statistic* for testing the hypothesis H is defined to be twice the difference between these two maximum log likelihoods,

$$D = 2[l(\hat{\theta}) - l(\tilde{\theta})]. \tag{12.3.1}$$

Note that D is twice the natural logarithm of a ratio of likelihoods,

$$D = 2 \log [L(\hat{\theta})/L(\tilde{\theta})],$$

and this explains its name.

Since $l(\hat{\theta}) \geq l(\tilde{\theta})$, D is non-negative. If D is small, then the maximum probability of the data is nearly as great under the hypothesis as it is under the basic model, and therefore the data are in good agreement with the hypothesis. A large value of D means that the data are much less probable under the hypothesis, and therefore the agreement is poor. Thus D ranks possible outcomes of the experiment according to how closely they agree with the hypothesis.

A simple hypothesis has the form $H: \theta = \theta_0$, where θ_0 is a vector of numerical values. Under H there is only one possible parameter value θ_0. Thus we have $\tilde{\theta} = \theta_0$, and the maximum log likelihood under H is $l(\theta_0)$. Hence (12.3.1) is the same as (12.2.1) when H is a simple hypothesis.

Calculation of the Significance Level

The significance level in a likelihood ratio test of the hypothesis H is given by

$$\text{SL} = P\{D \geq D_{\text{obs}} | H \text{ is true}\},$$

where D is the likelihood ratio statistic for testing H, and D_{obs} is the observed value of D.

Calculation of the exact significance level is possible in some examples, but in general there are both theoretical and computational difficulties. If H is composite, the exact significance level may well depend on the values of the q unknown parameters in the hypothesized model. Sometimes this problem can be avoided by using a suitable conditional distribution to calculate the significance level, but then the calculations required may become unmanageable. See Chapter 15 for further discussion of conditional tests.

Usually it is satisfactory to calculate an approximate significance level using the χ^2 approximation to the distribution of the likelihood ratio statistic D. It can be shown that, under conditions similar to those described in Section 11.3, the distribution of D when H is true is approximately χ^2 with $k-q$ degrees of freedom. When this approximation applies, we have

$$\text{SL} \approx P\{\chi^2_{(k-q)} \geq D_{\text{obs}}\}, \tag{12.3.2}$$

which can be evaluated using Table B4.

The χ^2 approximation will generally be quite accurate whenever the number of independent observations in the experiment is large in comparison with k, the number of parameters in the basic model. It is unwise to trust (12.3.2) whenever $\hat{\theta}$ or $\tilde{\theta}$ is on or near the boundary of the parameter space.

Note that the degrees of freedom for the χ^2 approximation is equal to $k-q$, where k and q are the numbers of functionally independent unknown parameters in the basic model and hypothesized model, respectively. Thus *the degrees of freedom for testing H is equal to the number of unknown parameters which are eliminated by H.*

To conclude this section, we give two examples of likelihood ratio tests for composite hypotheses. Many additional examples will be found in the following sections.

Testing H: $\beta = \beta_0$ when α is Unknown

Suppose that the probability model involves two unknown parameters, $\theta = (\alpha, \beta)$, so that $k = 2$. Consider the hypothesis H: $\beta = \beta_0$ where β is a particular numerical value. This is a composite hypothesis because no value is given for α. The hypothesized model involves the unknown parameter α, so that $q = 1$.

Let $l(\alpha, \beta)$ be the joint log likelihood function of α and β under the model. Let S_1 and S_2 be the two components of the score function as in Section 10.1.

Usually we can find the MLE's $\hat{\alpha}$ and $\hat{\beta}$ by solving the simultaneous equations

$$S_1(\alpha, \beta) = 0; \qquad S_2(\alpha, \beta) = 0.$$

We can find $\hat{\alpha}(\beta_0)$, the MLE of α given that $\beta = \beta_0$, by solving the equation

$$S_1(\alpha, \beta_0) = 0.$$

The maximum log likelihood under the model is $l(\hat{\alpha}, \hat{\beta})$. The maximum log likelihood under the hypothesis H: $\beta = \beta_0$ is $l(\hat{\alpha}(\beta_0), \beta_0)$. Hence the likelihood ratio statistic for testing H: $\beta = \beta_0$ is

$$D = 2[l(\hat{\alpha}, \hat{\beta}) - l(\hat{\alpha}(\beta_0), \beta_0)].$$

Note that, by (10.3.1),

$$D = -2r_{\max}(\beta_0)$$

where $r_{\max}(\beta)$ is the maximum log relative likelihood function of β. We considered this likelihood ratio statistic in Section 11.5, and noted that its distribution when $\beta = \beta_0$ is approximately χ^2 with $k-q=1$ degrees of freedom. Thus we have

$$\text{SL} = P\{D \geq D_{\text{obs}} | \beta = \beta_0\} \approx P\{\chi^2_{(1)} \geq D_{\text{obs}}\}.$$

There is one degree of freedom for testing H: $\beta = \beta_0$, because it reduces the number of unknown parameters by one.

EXAMPLE 12.3.1. In Example 10.1.2 we considered the lifetimes $x_1, x_2, \ldots, x_n$ of $n = 23$ deep-groove ball bearings. These were assumed to be independent observations from a Weibull distribution with probability density function

$$f(x) = \lambda\beta x^{\beta-1} \exp\{-\lambda x^\beta\} \qquad \text{for } 0 < x < \infty.$$

There are two unknown parameters, $\lambda > 0$ and $\beta > 0$.

We noted in Example 10.2.2 that the value $\beta = 1$ is of special importance, because when $\beta = 1$ the Weibull distribution simplifies to an exponential distribution. Under an exponential distribution model, there is a constant risk of failure, and no deterioration or improvement with age. Thus we wish to know whether the 23 observed lifetimes are consistent with the hypothesis $\beta = 1$.

To test H: $\beta = 1$, we shall compute the observed value of the likelihood ratio statistic and then use the χ^2 approximation. Since H reduces the number of unknown parameters by one, there is one degree of freedom for the test.

From Example 10.1.2, the joint log likelihood function is

$$l(\lambda, \beta) = n \log \lambda + n \log \beta + (\beta - 1)\Sigma \log x_i - \lambda \Sigma x_i^\beta,$$

and the MLE's are

$$\hat{\lambda} = 9.515 \times 10^{-5}; \qquad \hat{\beta} = 2.1021.$$

The maximum log likelihood under the model is

$$l(\hat{\lambda}, \hat{\beta}) = -113.691.$$

Also from Example 10.1.2, the MLE of λ given β is

$$\hat{\lambda}(\beta) = n/\Sigma x_i^{\beta}.$$

Thus the MLE of λ under the hypothesis $\beta = 1$ is

$$\hat{\lambda}(1) = n/\Sigma x_i = 23/1661 = 0.001385,$$

and the maximum log likelihood under H: $\beta = 1$ is

$$l(0.001385, 1) = -121.433.$$

The observed value of the likelihood ratio statistic for testing H: $\beta = 1$ is twice the difference between these maximum log likelihoods:

$$D_{\text{obs}} = 2[-113.691 + 121.433] = 15.48.$$

This result could also have been obtained from the expression for $r_{\max}(\beta)$ in Example 10.3.2. The χ^2 approximation gives

$$\text{SL} \approx P\{\chi^2_{(1)} \geq 15.48\} < 0.001$$

from Table B4. There is very strong evidence against the hypothesis $\beta = 1$. The observations are not compatible with the simpler exponential distribution model.

Tests for Homogeneity

Suppose that two or more independent experiments give information about the same unknown parameter θ. If the experiments are in reasonable agreement with one another, we can pool or combine the information about θ by adding log likelihood functions (see Section 9.2 and Example 9.3.2). If, on the other hand, the experiments contradict one another, it would not be appropriate to combine them. Instead we would estimate θ separately for each experiment, and try to discover why the experiments produced dissimilar results.

Suppose that there are k independent experiments, and initially let us suppose that we have a different parameter θ_i for each experiment. Let $l_i(\theta)$ and $\hat{\theta}_i$ denote the log likelihood function and MLE for the ith experiment. $(i = 1, 2, \ldots, k)$. The overall log likelihood function is

$$l(\theta_1, \theta_2, \ldots, \theta_k) = l_1(\theta_1) + l_2(\theta_2) + \cdots + l_k(\theta_k) = \Sigma l_i(\theta_i),$$

and its maximum value is $\Sigma l_i(\hat{\theta}_i)$.

Now consider the hypothesis of homogeneity,

$$H\colon \theta_1 = \theta_2 = \cdots = \theta_k,$$

and let θ denote the unknown common value of the θ_i's. Under H, the log likelihood function is $\Sigma l_i(\theta)$, and we maximize this to obtain the combined or pooled MLE $\tilde{\theta}$, say. The maximum of the log likelihood under H is $\Sigma l_i(\tilde{\theta})$, and

by (12.3.1), the likelihood ratio statistic for testing H is

$$D = 2[\Sigma l_i(\hat{\theta}_i) - \Sigma l_i(\tilde{\theta})] = -2\Sigma r_i(\tilde{\theta})$$

where r_i is the log RLF from the ith experiment:

$$r_i(\theta_i) = l_i(\theta_i) - l_i(\hat{\theta}_i).$$

If D is large, there is no parameter value which is reasonably plausible in all experiments, and hence the experiments give conflicting information about θ.

There are $k - 1$ degrees of freedom for testing H because it reduces the number of unknown parameters from k to 1. Hence (12.3.2) gives

$$\text{SL} \approx P\{\chi^2_{(k-1)} \geq D_{\text{obs}}\}.$$

A small significance level is evidence that the homogeneity hypothesis is false, and that the information from the k experiments should not be pooled.

EXAMPLE 12.3.2. In Examples 9.2.2 and 9.3.2 we considered data from $k = 2$ experiments with test tubes containing river water. The parameter of interest is μ, the expected number of bacteria per ml of river water. For the first experiment the log RLF is

$$r_1(\mu) = -280\mu + 12 \log(1 - e^{-10\mu}) + 24.43,$$

and for the second experiment we have

$$r_2(\mu) = -37\mu + 3 \log(1 - e^{-\mu}) + 10.66.$$

The pooled MLE of μ based on the data from both experiments was found to be $\tilde{\mu} = 0.04005$. Hence the observed value of the likelihood ratio statistic for testing homogeneity is

$$D_{\text{obs}} = -2[r_1(\tilde{\mu}) + r_2(\tilde{\mu})] = 1.24.$$

There is just one degree of freedom for testing H, and thus

$$\text{SL} \approx P\{\chi^2_{(1)} \geq 1.24\} > 0.25.$$

There is no evidence against the homogeneity hypothesis, and it is reasonable to pool information about μ as we did in Example 9.3.2.

PROBLEMS FOR SECTION 12.3

1.† Suppose that X_1, X_2, and X_3 have a trinomial distribution with index n and probability parameters p_1, p_2, p_3 where $\Sigma p_j = 1$. The log likelihood function is

$$l(p_1, p_2, p_3) = \Sigma X_j \log p_j,$$

and the observed values of the X_j's are 32, 46, and 22 (see Problem 9.1.4).

(a) Find the maximum log likelihood when

(i) p_j is estimated as X_j/n for $j = 1, 2, 3$;

(ii) the p_j's satisfy the hypothesis

$$H: p_1 = \theta^2, \qquad p_2 = 2\theta(1-\theta), \qquad p_3 = (1-\theta)^2;$$

(iii) the p_j's satisfy H and, in addition, $\theta = \frac{1}{2}$.

(b) Use the results from (i) and (ii) above to test the hypothesis H.
(c) Use the results from (ii) and (iii) to test whether $\theta = \frac{1}{2}$, assuming H to be true.
(d) Use the results from (i) and (iii) to test the hypothesis $p_1 = p_2 = \frac{1}{4}$, $p_3 = \frac{1}{2}$. Note that the likelihood ratio statistic and degrees of freedom are the totals of those in (b) and (c).

2. A genetics experiment yields observations X_1, X_2, X_3, X_4 with multinomial probability

$$\binom{n}{X_1 X_2 X_3 X_4}\left(\frac{p}{2}\right)^{X_1+X_4}\left(\frac{1-p}{2}\right)^{X_2+X_3}$$

where $\Sigma X_i = n$. The following are the results from three independent repetitions of the experiment:

	X_1	X_2	X_3	X_4
Repetition 1	26	7	9	22
Repetition 2	24	9	9	22
Repetition 3	23	9	12	20

Test the hypothesis that the value of p is the same in all three repetitions.

3. (a) Let $Y_1, Y_2, \ldots, Y_k$ be independent Poisson variates with means $\mu_1, \mu_2, \ldots, \mu_k$. Show that the likelihood ratio statistic for testing $H: \mu_1 = \mu_2 = \cdots = \mu_k$ is given by

$$D = 2\Sigma Y_i \log(Y_i/\bar{Y}).$$

(b) The numbers of cancer cells surviving a treatment in each of three replications of an experiment were 235, 184, and 189. Test the hypothesis that these three observations come from the same Poisson distribution.

4. (a) Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent Poisson variates with means $\mu_1, \mu_2, \ldots, \mu_n$. Let $P_1, P_2, \ldots, P_n$ be known constants. Consider the hypothesis

$$H: \mu_1 = \lambda P_1, \mu_2 = \lambda P_2, \ldots, \mu_n = \lambda P_n$$

where λ is unknown. Show that the likelihood ratio statistic for testing H is

$$D = 2\Sigma Y_i \log(Y_i/\tilde{\mu}_i)$$

where $\tilde{\mu}_i = \tilde{\lambda} P_i$ and $\tilde{\lambda} = (\Sigma Y_i)/(\Sigma P_i)$.

(b) In Problem 9.2.2(b), test the hypothesis that the death rates for the 10 regions are proportional to the populations of the regions.

5.†(a) Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent Poisson variates. The X_i's have expected value μ_1, and the Y_i's have expected value μ_2. Derive the likelihood ratio statistic for testing the hypothesis $\mu_1 = \mu_2$.

(b) Bacteria counts were made for 27 volumes of river water, each of unit volume.

The results were as follows:

Location 1:	0	2	0	1	1	2	2	0	2	0	0	1			
Location 2:	3	1	2	1	3	2	3	3	1	2	2	1	3	3	1

The bacteria are assumed to be randomly and uniformly distributed throughout the river water, with μ_1 per unit volume at location 1, and μ_2 per unit volume at location 2. Test the hypothesis $\mu_1 = \mu_2$.

6. Let $X_1, X_2, \ldots, X_n$ be independent exponential variates with mean θ_1, and let $Y_1, Y_2, \ldots, Y_m$ be independent exponential variates with mean θ_2. Show that the likelihood ratio statistic for testing $H: \theta_1 = \theta_2$ depends only on n, m, and $\bar{X}/\bar{Y}$.

7. Suppose that k independent experiments give log RLF's $r_1, r_2, \ldots, r_k$ and MLE's $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ for the same unknown parameter θ. Furthermore, suppose that the normal approximation applies to each of the r_i's:

$$r_i(\theta) \approx -\tfrac{1}{2}(\theta - \hat{\theta}_i)^2 c_i,$$

where $c_i = \mathcal{I}_i(\hat{\theta}_i)$.

(a) Show that the MLE of θ based on all k experiments is approximately equal to $\tilde{\theta}$, where

$$\tilde{\theta} = (\Sigma c_i \hat{\theta}_i)/(\Sigma c_i).$$

(b) Show that the likelihood ratio statistic for testing $H: \theta = \theta_0$ is approximately $(\tilde{\theta} - \theta_0)^2 \Sigma c_i$.
(c) Show that the likelihood ratio statistic for testing the homogeneity hypothesis $H: \theta_1 = \theta_2 = \cdots = \theta_k$ is approximately $\Sigma(\hat{\theta}_i - \tilde{\theta})^2 c_i$.
(d) What are the approximate distributions of the likelihood ratio statistics in (b) and (c)?

8. Continuation of Problem 7. Seven different dilution series experiments were used to estimate a parameter h, called the "hit number". The MLE $\hat{h}$ and observed information $\hat{\mathcal{I}}$ are given below for each of the seven experiments.

$\hat{h}$	2.028	2.108	1.912	1.675	1.730	1.808	1.889
$\hat{\mathcal{I}}$	19.63	25.18	32.34	70.54	64.88	67.63	36.58

In each case, the likelihood function was approximately normal in h.

(a) Are these results consistent with a common value of h in all seven experiments?
(b) Are the combined results consistent with the theoretical value $h = 2$?

9.†Continuation of Problem 7. Suppose that three independent experiments give likelihood functions that are approximately normal in θ, with the following summary statistics:

$$\hat{\theta}_1 = 9.74 \qquad \hat{\theta}_2 = 8.35 \qquad \hat{\theta}_3 = 10.27$$
$$\mathcal{I}_1(\hat{\theta}_1) = 0.563 \qquad \mathcal{I}_2(\hat{\theta}_2) = 0.345 \qquad \mathcal{I}_3(\hat{\theta}_3) = 0.695$$

(a) Test the hypothesis that the value of θ is the same in all three experiments.
(b) Obtain four approximate 95% confidence intervals for θ, one from each

experiment taken separately, and one from the combined results of all three experiments.

10. Consider the situation described in Problem 10.1.5. Testing stops when there have been m failures with each treatment. Let $X_1, X_2, \ldots, X_m$ be the numbers of successes with treatment A, and let $Y_1, Y_2, \ldots, Y_m$ be the numbers of successes with treatment B. Derive the likelihood ratio statistic for testing the hypothesis $\alpha = \beta$.

11. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent normal variates, all with the same known variance σ^2. The X_i's have mean μ_1 and the Y_i's have mean μ_2.

(a) Show that the likelihood ratio statistic for testing $H: \mu_1 = \mu_2$ is

$$D = \frac{1}{\sigma^2}[n(\bar{X} - \tilde{\mu})^2 + m(\bar{Y} - \tilde{\mu})^2] = \frac{1}{\sigma^2}\frac{nm}{n+m}(\bar{X} - \bar{Y})^2,$$

where $\tilde{\mu} = (n\bar{X} + m\bar{Y})/(n+m)$.

(b) Find the distribution of $\bar{X} - \bar{Y}$. Hence show that the distribution of D is exactly $\chi^2_{(1)}$.

12.4. Tests for Binomial Probabilities

Suppose that k different treatments are to be compared on the basis of success/failure data. The first treatment is given to n_1 subjects and Y_1 successes are observed. The second treatment is given to n_2 different subjects and Y_2 successes are observed. The results of an experiment with k treatments can be summarized in a table as follows:

Treatment no.	1	2	...	k
No. of successes	Y_1	Y_2	...	Y_k
No. of failures	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_k - Y_k$
Total	n_1	n_2	...	n_k

We wish to make inferences about the success probabilities $p_1, p_2, \ldots, p_k$ on the basis of the observed results.

We assume that Y_i, the number of successes with treatment i, has a binomial (n_i, p_i) distribution, and that the Y_i's are independent. The basic model involves a vector of k different unknown parameters, $p = (p_1, p_2, \ldots, p_k)$, where p_i is the success probability for the ith treatment. The log likelihood function is

$$l(p) = \Sigma y_i \log p_i + \Sigma(n_i - y_i)\log(1 - p_i).$$

The MLE of p_i is $\hat{p}_i = y_i/n_i$, and the maximum of the log likelihood under the basic model is

$$l(\hat{p}) = \Sigma y_i \log\frac{y_i}{n_i} + \Sigma(n_i - y_i)\log\frac{n_i - y_i}{n_i}.$$

Now suppose that we wish to test an hypothesis H about the p_i's. For instance, we may wish to test that they are equal:

$$H_1: p_1 = p_2 = \cdots = p_k.$$

Their common value is not given, so there is one unknown parameter under H_1. Alternatively, if the k treatments are different doses $d_1, d_2, \ldots, d_k$ of a drug, we might wish to test the hypothesis

$$H_2: p_i = 1 - (1 + e^{\alpha + \beta d_i})^{-1} \qquad \text{for } i = 1, 2, \ldots, k,$$

which states that the response probability is related to the dose via the logistic model (10.5.1). There are two unknown parameters under H_2.

Assuming H to be true, we can rewrite the log likelihood as a function of the q remaining unknown parameters and find their MLE's. From these we can compute $\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_k$, the MLE's of the original probability parameters under H. The maximum of the log likelihood is then

$$l(\tilde{p}) = \Sigma y_i \log \tilde{p}_i + \Sigma(n_i - y_i)\log(1 - \tilde{p}_i).$$

By (12.3.1), the likelihood ratio statistic for testing H is

$$D = 2[l(\hat{p}) - l(\tilde{p})]$$

$$= 2\Sigma y_i \log\frac{y_i}{n_i\tilde{p}_i} + 2\Sigma(n_i - y_i)\log\frac{n_i - y_i}{n_i(1 - \tilde{p}_i)}.$$

Note that $n_i\tilde{p}_i$ and $n_i(1 - \tilde{p}_i)$ are the expected numbers of successes and failures for the ith treatment under H, whereas y_i and $n_i - y_i$ are the observed frequencies. Thus we can write

$$D = 2\Sigma(\text{obs freq}) \cdot \log\frac{\text{obs freq}}{\text{exp freq}} \tag{12.4.1}$$

where the sum extends over all $2k$ classes (successes and failures).

The degrees of freedom for testing H is $k - q$, where q is the number of unknown parameters which remain under H. By (12.3.2) we have

$$\text{SL} \approx P\{\chi^2_{(k-q)} \geq D_{\text{obs}}\}.$$

The approximation will be accurate provided that all of the expected frequencies $n_i\tilde{p}_i$ and $n_i(1 - \tilde{p}_i)$ are fairly large.

EXAMPLE 12.4.1. The food additive "Red Dye Number 2" was fed to 44 rats at a low dose and to 44 rats at a high dose. Later the rats were examined for tumors, and the results were as follows:

Treatment	Low dose	High dose
Tumor present	4(9)	14(9)
No tumor	40(35)	30(35)
Total	44	44

Note that 47% developed tumors at the high dose, and only 9% developed tumors at the low dose. Could these results have arisen by chance, or is there evidence of a real dose effect?

Let Y_1 and Y_2 be the numbers of rats with tumors at the low and high doses, respectively. We assume that Y_1 and Y_2 are independent, with $Y_1 \sim$ binomial (n_1, p_1) and $Y_2 \sim$ binomial (n_2, p_2), where $n_1 = n_2 = 44$. We wish to know whether there is conclusive evidence against $H: p_1 = p_2$.

Let p denote the unknown common value of p_1 and p_2 under the hypothesis H. From Example 9.2.2, the MLE of p is

$$\tilde{p} = \frac{y_1 + y_2}{n_1 + n_2} = \frac{4 + 14}{44 + 44} = \frac{9}{44}.$$

Under H we have $\tilde{p}_1 = \tilde{p}_2 = \frac{9}{44}$, and the expected frequencies are

$$n_1\tilde{p}_1 = 9; \qquad n_2\tilde{p}_2 = 9;$$
$$n_1(1 - \tilde{p}_1) = 35; \qquad n_2(1 - \tilde{p}_2) = 35.$$

The table above shows these values in parentheses. By (12.4.1), the observed value of the LR statistic is

$$D_{\text{obs}} = 2\left[4 \log \frac{4}{9} + 14 \log \frac{14}{9} + 40 \log \frac{40}{35} + 30 \log \frac{30}{35}\right] = 7.32.$$

Since H reduces the number of unknown parameters from 2 to 1 there is one degree of freedom for testing H, and

$$\text{SL} \approx P\{\chi^2_{(1)} \geq 7.32\} < 0.01.$$

Results as extreme as those observed would rarely occur if p_1 and p_2 were equal, and therefore we have strong evidence against $H: p_1 = p_2$. The incidence of tumors is greater at the high dose than at the low dose, and the difference is too large to be attributed to chance.

EXAMPLE 12.4.2. Table 10.5.1 shows the data from an experiment in which an insecticide was administered in $k = 5$ doses. We assume that Y_i, the number killed at dose d_i, has a binomial (n_i, p_i) distribution, and that results for different doses are independent. We wish to determine whether the logistic dose-response model (10.5.1) is compatible with these data. Thus the hypothesis of interest is

$$H: p_i = 1 - (1 + e^{\alpha + \beta d_i})^{-1} \qquad \text{for } i = 1, 2, \ldots, 5$$

where α and β are unknown parameters.

We showed in Example 10.5.1 that the MLE's of α and β are $\tilde{\alpha} = -4.8869$, $\tilde{\beta} = 3.1035$. Using these values, we computed estimated probabilities $\tilde{p}_i$ and then found the expected frequencies $n_i\tilde{p}_i$, $n_i(1 - \tilde{p}_i)$ (see Table 10.5.2). Now, by (12.4.1), the observed value of the LR statistic for testing H is

$$D_{\text{obs}} = 2\left[6 \log \frac{6}{6.39} + 44 \log \frac{44}{43.61} + \cdots + 6 \log \frac{6}{4.47}\right] = 1.42.$$

Since H reduces the number of unknown parameters from 5 to 2, there are three degrees of freedom for the test, and

$$\text{SL} \approx P\{\chi^2_{(3)} \geq 1.42\} > 0.5.$$

The observed value of D is not unusually large, and hence there is no evidence against the hypothesis of a logistic dose-response curve.

We concluded previously, after informal inspections of Figure 10.5.2 and Table 10.5.2, that the logistic model fits the data well. The LR test just performed provides a more formal justification of this conclusion. The test tells us whether the observed discrepancies can be attributed to chance variations. Tables and graphs tell us what kinds of departures have occurred and how large they are. Both significance tests and less formal methods are useful in assessing the fit of the model.

PROBLEMS FOR SECTION 12.4

1. Two hundred volunteers participated in an experiment to examine the effectiveness of vitamin C in preventing the common cold. One hundred of them were selected at random to receive a daily dose of vitamin C, and the others received a placebo. None of the volunteers knew which group they were in. During the test period, 20 of those taking vitamin C and 35 of those receiving the placebo caught colds. Test the hypothesis that the probability of catching a cold is the same for both groups.

2. A seed dealer claims that his sweet pea seeds have a germination rate of 80%. A customer purchased 4 packages of sweet pea seeds, one package of each of four colors. He planted 100 seeds from each package. The numbers of seeds germinating within one month were as follows:

	Red	White	Blue	Yellow
Germination	75	66	81	74
No germination	25	34	19	26

   (a) Test the hypothesis that the germination rate is 80% for all four colors.
   (b) Test the hypothesis that the germination rate is the same for all four colors (but not necessarily 80%).
   (c) Assuming that the germination rate is the same for all four colors, test the hypothesis that it is 80%.
   (d) How are the likelihood ratio statistics in (a), (b), and (c) related?

3.† Four hundred patients took part in a study to compare the effectiveness of three similar drugs. Each drug was given to 100 patients, and the remaining 100 patients received a placebo. It was then observed whether or not there was improvement in the condition of each patient. The results were as follows:

	Drug A	Drug B	Drug C	Placebo
Improvement	24	19	29	10
No improvement	76	81	71	90

(a) Test the hypothesis that the probability of improvement is the same in all four groups.
(b) Test the hypothesis that the three drugs are equally effective.
(c) Assuming that the three drugs are equally effective, test the hypothesis that the success rate is the same for those receiving a drug as for those receiving the placebo.
(d) How are the likelihood ratio statistics in (a), (b), and (c) related?

4. An experiment involved exposing a large number of cancer cells to a treatment and then observing how many survived. There were two treatments, each of which was applied to two different groups of cells. The results were as follows:

Treatment	A	A	B	B
Number of cells	48000	48000	192000	192000
Number surviving	7	9	49	39

Assume that cells respond independently, and that the survival probabilities for the four groups are α_1, α_2, β_1, and β_2, respectively.

(a) Test the hypothesis $H: \alpha_1 = \alpha_2, \beta_1 = \beta_2$.
(b) Assuming that the hypothesis in (a) is true, test the hypothesis that the survival probability is the same for both treatments.

5. Test the hypothesis of a logistic dose-response model in Problem 10.5.1.

6.† Test the hypothesis $p = e^{\alpha+\beta d}$ in Problem 10.5.2.

7. An interviewer in a shopping plaza asks individuals who pass by if they are willing to fill in a questionnaire. He keeps asking people until 30 agree. The following are the numbers of refusals he receives on each of six days.

Day	1	2	3	4	5	6
Number refusing	70	67	80	62	100	112

Assume that individuals respond independently, and that each individual questioned on the ith day has probability p_i of responding. Test the hypothesis $p_1 = p_2 = \cdots = p_6$.

Note: Since the distribution of the number refusing is negative binomial rather than binomial, you will need to derive the likelihood ratio statistic from first principles.

12.5. Tests for Multinomial Probabilities

Suppose that we have data from n independent repetitions of an experiment, and that we wish to assess how well the data agree with an hypothesized probability model for the experiment. One way of doing this is to construct a table of observed frequencies, which are then compared with expected frequencies under the hypothesized model (see Section 1.4). A test of significance may be used to determine whether the discrepancy between the observed and expected frequencies is too great to be attributed to chance.

To construct a frequency table, we partition the sample space S for a single repetition into k mutually exclusive classes or events, $S = A_1 \cup A_2 \cup \cdots \cup A_k$. Let p_j be the probability of event A_j, and let f_j be the number of times that A_j occurs in the n repetitions. Exactly one of the events must occur in each repetition, so $\Sigma p_j = 1$ and $\Sigma f_j = n$.

Under the assumption of independent repetitions, the distribution of the f_j's is multinomial with joint probability function

$$\binom{n}{f_1 f_2 \cdots f_k} p_1^{f_1} p_2^{f_2} \cdots p_k^{f_k}.$$

The log likelihood function is

$$l(p) = l(p_1, p_2, \ldots, p_k) = \Sigma f_j \log p_j$$

where $\Sigma p_j = 1$. It can be shown that, subject to this condition, $l(p)$ is maximized for $\hat{p}_j = y_j/n$. Hence the maximum log likelihood under the basic multinomial model is

$$l(\hat{p}) = \Sigma f_j \log (f_j/n).$$

The hypothesized model will determine the p_j's numerically or as functions of unknown parameters. We find the MLE's of any unknown parameters and use these to compute $\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_k$, the MLE's of the p_j's under the hypothesized model. The maximum of the log likelihood is then

$$l(\tilde{p}) = \Sigma f_j \log \tilde{p}_j.$$

By (12.3.1), the likelihood ratio statistic for testing the model is

$$D = 2[l(\hat{p}) - l(\tilde{p})] = 2\Sigma f_j \log (f_j/e_j), \tag{12.5.1}$$

where $e_j = n\tilde{p}_j$ is the estimated expected frequency for the jth class under the hypothesized model. Note that (12.5.1) has the same form as (12.4.1), but the sum is now taken over k classes rather than $2k$ classes.

Since the k probabilities $p_1, p_2, \ldots, p_k$ must sum to 1, there are only $k-1$ functionally independent parameters in the basic multinomial model. Let q be the number of unknown parameters in the hypothesized model. Then there are $(k-1)-q$ degrees of freedom for the χ^2 approximation, and (12.3.2) gives

$$\text{SL} \approx P\{\chi^2_{(k-1-q)} \geq D_{\text{obs}}\}.$$

Classes for which $e_j \approx 0$ but $f_j \geq 1$ will have a big effect on D_{obs}, and the χ^2 approximation should not be trusted when the e_j's are small. The usual rule of thumb is that the e_j's should all be at least 5, but an occasional smaller value is not too harmful.

Another test statistic which may be used with multinomial or binomial data is the *Pearson goodness of fit statistic*,

$$D = \Sigma (f_j - e_j)^2/e_j. \tag{12.5.2}$$

The observed value of this statistic will be very nearly equal to that of the likelihood ratio statistic (12.5.1) or (12.4.1) when the e_j's are very large, and the same χ^2 approximation can be used.

Significance tests for multinomial data using test statistic (12.5.1) or (12.5.2) are often called *goodness of fit tests.*

EXAMPLE 12.5.1. In Example 12.2.2 the basic model was multinomial with $k = 6$ classes, and we carried out a likelihood ratio test of the hypothesis

$$H: p_1 = p_2 = \cdots = p_6 = \tfrac{1}{6}.$$

The data were the observed frequencies 16, 15, 14, 20, 22, 13 from 100 rolls of a die. This analysis can be simplified by using formula (12.5.1) to compute the observed value of the likelihood ratio statistic. Under H, each class has expected frequency $e_j = 100(\frac{1}{6}) = 16.67$. Now (12.5.1) gives

$$D_{\text{obs}} = 2\left[16 \log \frac{16}{16.67} + 15 \log \frac{15}{16.67} + \cdots + 13 \log \frac{13}{16.67}\right] = 3.70,$$

which agrees with the result in Example 12.2.2. Since H reduces the number of unknown parameters from $6 - 1 = 5$ to 0, there are 5 degrees of freedom for the test, and

$$\text{SL} \approx P\{\chi^2_{(5)} \geq 3.70\} \geq 0.5$$

as before. There is no evidence against the hypothesis of a balanced die.

EXAMPLE 12.5.2. The following are the observed frequencies from the ESP experiment in Example 12.1.2:

No. correct j	0	1	2	4	Total
Obs freq f_j	17	18	19	6	50
Exp freq e_j	18.75	16.67	12.50	2.08	50

Under the basic model, the f_j's come from a multinomial distribution with $k = 4$ classes. If there is no ESP, the four classes have probabilities $\frac{9}{24}$, $\frac{8}{24}$, $\frac{6}{24}$, and $\frac{1}{24}$, so the hypothesis of interest is

$$H: p_1 = \frac{9}{24}, p_2 = \frac{8}{24}, p_3 = \frac{6}{24}, p_4 = \frac{1}{24}.$$

We multiply these four probabilities by 50 to get the expected frequencies under H.

By (12.5.1), the observed value of the LR statistic is

$$D_{\text{obs}} = 2\left[17 \log \frac{17}{18.75} + \cdots + 6 \log \frac{6}{2.08}\right] = 6.22.$$

H reduces the number of unknown parameters from $4 - 1 = 3$ to 0, so

$$\text{SL} \approx P\{\chi^2_{(3)} \geq 6.22\} \approx 0.10.$$

If H were true, one would obtain $D \geq 6.22$ in about 10% of repetitions of the experiment. Therefore we do not have conclusive evidence against the hypothesis.

This test and the one in Example 12.1.2 give about the same significance level for these data, but in other examples they may give quite different results. For instance, suppose that

$$f_0 = 25, f_1 = 10, f_2 = 7, f_4 = 8.$$

Then $T_{\text{obs}} = 56$, and the test of Example 12.1.3 gives $\text{SL} \approx 0.2$. However, the likelihood ratio test gives

$$\text{SL} \approx P\{\chi^2_{(3)} \geq 17.58\} < 0.001.$$

The total number of correct guesses is not far from the expected number under H, but the observed frequencies are not at all like what we'd expect under H.

The likelihood ratio statistic (12.5.1) is a "general purpose" measure which does not look for any specific type of departure from H. The test statistic used in Example 12.1.2 was designed to detect a particular type of departure — an excess of correct guesses. It is more sensitive to departures of the type anticipated, but it may fail to detect substantial departures of other kinds.

EXAMPLE 12.5.3. In Example 4.4.3 we considered the distribution of flying-bomb hits over 576 regions of equal area in south London. The following table shows the number of regions f_j which suffered exactly j hits ($j = 0, 1, 2, \ldots$):

No. of hits j	0	1	2	3	4	≥ 5	Total
Obs freq f_j	229	211	93	35	7	1	576
Exp freq e_j	226.74	211.39	98.54	30.62	7.14	1.57	576

One region received 7 hits, and the total number of hits observed on all 576 regions is

$$\Sigma j f_j = 229 \times 0 + 211 \times 1 + \cdots + 7 \times 4 + 1 \times 7 = 537.$$

Under the basic model, the f_j's come from a multinomial distribution with $k = 6$ classes.

If points of impact are randomly and uniformly distributed over the study region, the number of hits in an area should have a Poisson distribution. Thus we consider the hypothesis

$$H: p_j = \mu^j e^{-\mu}/j! \qquad \text{for } j = 0, 1, 2, \ldots$$

where μ is an unknown parameter.

Under H, the log likelihood function is

$$\Sigma f_j \log p_j = \Sigma j f_j \log \mu - \mu \Sigma f_j - \Sigma f_j \log j!$$

from which the MLE is found to be

$$\tilde{\mu} = \Sigma j f_j / \Sigma f_j = 537/576 = 0.9323.$$

(This is not quite right — see the note at the end of this example.) Using this estimate, we can find $\tilde{p}_j$ and $e_j = 576\tilde{p}_j$ for $j = 0, 1, \ldots, 4$. The expected frequency for the last class is then obtained by subtraction from the total (see Example 4.4.3).

The observed value of the LR statistic is

$$D_{\text{obs}} = 2\Sigma f_j \log \frac{f_j}{e_j} = 1.18.$$

The hypothesis reduces the number of unknown parameters from $k - 1 = 5$ to 1, so

$$\text{SL} \approx P\{\chi^2_{(4)} \geq 1.18\} \approx 0.9.$$

There is no evidence against the hypothesis. The observed frequencies are in close agreement with the expected frequencies from a Poisson distribution.

The expected frequency in the last class is only 1.57, and we might therefore have some concern about the adequacy of the χ^2 approximation. To check this, we could combine the last two classes into a single class (≥ 4) with $f = 7 + 1 = 8$ and $e = 7.14 + 1.57 = 8.71$. Summing over the $k = 5$ classes gives $D_{\text{obs}} = 1.00$ with $(5 - 1) - 1 = 3$ degrees of freedom, and

$$\text{SL} \approx P\{\chi^2_{(3)} \geq 1.00\} \approx 0.8.$$

The conclusion is the same as before.

Note. In calculating $\tilde{\mu}$ above, we used the fact that the observation in class ≥ 5 was "7". Strictly speaking, $\tilde{\mu}$ should be obtained using only the information in the frequency table. For the first test with $k = 6$ we have $p_j = \mu^j e^{-\mu}/j!$ for $0 \leq j \leq 4$, and

$$p_{\geq 5} = 1 - e^{-\mu}(1 + \mu + \mu^2/2! + \mu^3/3! + \mu^4/4!)$$

so the appropriate log likelihood function is

$$\sum_{j=0}^{4} f_j \log p_j + f_{\geq 5} \log[1 - e^{-\mu}(1 + \mu + \cdots + \mu^4/4!)].$$

Maximizing this by Newton's method or trial and error gives $\tilde{\mu} = 0.9291$, and this is the estimate which should be used in computing the e_j's. The result is a slightly better fit ($D_{\text{obs}} = 1.17$ rather than 1.18), but no change in the conclusion.

Similarly, when we combine the last two classes, the appropriate log likelihood function is

$$\sum_{j=0}^{3} f_j \log p_j + f_{\geq 4} \log[1 - e^{-\mu}(1 + \mu + \mu^2/2! + \mu^3/3!)]$$

which is maximized for $\tilde{\mu} = 0.9300$. Recomputing expected frequencies with this value of μ gives $D_{\text{obs}} = 0.99$ rather than 1.00.

In general, if the value of μ used in the calculations is not the "true" MLE, D_{obs} will be too large. However, unless there is a substantial amount of grouping, the difference will usually be too small to matter.

EXAMPLE 12.5.4. Consider the set of 109 waiting times between mining accidents which we discussed in Sections 1.2 and 1.4. If accidents occur randomly in time at the constant rate of λ per day, the time T between successive accidents has an exponential distribution with mean $\theta = 1/\lambda$ (see Section 6.5). Here we have $n = 109$ observations $t_1, t_2, \ldots, t_n$, and we wish to determine whether an exponential distribution model is satisfactory.

One way to examine the fit of the model is to group the data into k classes $[a_{j-1}, a_j)$ and prepare a frequency table (see Example 1.2.1). For the exponential distribution we have

$$P(a_{j-1} \leq T < a_j) = \exp(-a_{j-1}/\theta) - \exp(-a_j/\theta)$$

and so the hypothesis of interest is

$$H: p_j = \exp(-a_{j-1}/\theta) - \exp(-a_j/\theta) \qquad \text{for } j = 1, 2, \ldots, k.$$

There will be $k - 2$ degrees of freedom for testing H because it reduces the number of unknown parameters from $k - 1$ to 1.

Table 12.5.1 is obtained from Table 1.4.1 by combining the last two classes. The e_j's were computed using $\theta = \bar{t} = 241$, which is the MLE based on the original set of 109 measurements. Now (12.5.1) gives

$$D_{\text{obs}} = 2\Sigma f_j \log(f_j/e_j) = 18.79.$$

Since there are $k = 11$ classes, we have

$$\text{SL} \approx P\{\chi^2_{(9)} \geq 18.79\} \approx 0.025.$$

Thus there is some evidence against the exponential distribution model.

The expected frequency for the last class is only 1.72, and we might be tempted to combine the last two classes as we did in Example 12.5.3. We

Table 12.5.1. Observed and Expected Frequencies for the Mining Accident Data of Example 1.2.1

Class	f_j	e_j	Class	f_j	e_j
[0, 50)	25	20.42	[300, 350)	11	5.88
[50, 100)	19	16.60	[350, 400)	6	4.78
[100, 150)	11	13.49	[400, 600)	5	11.69
[150, 200)	8	10.96	[600, 1000)	3	7.32
[200, 250)	9	8.91	[1000, ∞)	5	1.72
[250, 300)	7	7.24	Total	109	109.01

would then obtain $D_{obs} = 11.51$ with 8 degrees of freedom, and SL ≈ 0.2. The result is now quite different because the deviations in the last two classes were in opposite directions and have cancelled one another. Combining these classes is not a good idea because it obscures the difficulties in the right hand tail of the distribution. In fact, the model is not appropriate for these data because the accident rate λ is not constant over time (see Example 1.4.2).

A difficulty with the above analysis is that the results obtained will depend to some extent upon the arbitrary grouping used to produce the frequency table. It is a good idea to try three or four different groupings and check that similar results are obtained. Alternatively, the fit of the model can be checked via informal graphical procedures (see Example 6.3.1).

The note following Example 12.5.3 applies to this example as well. The likelihood function based on the frequency table is

$$\sum_{j=1}^{11} f_j \log p_j \qquad \text{where } p_j = \exp(-a_{j-1}/\theta) - \exp(-a_j/\theta),$$

and maximizing this function gives $\tilde{\theta} = 232.5$. This value, not $\theta = 241$, should properly have been used in computing expected frequencies. We would then have obtained $D_{obs} = 18.65$ instead of $D_{obs} = 18.79$, a change which is too small to be of any practical importance.

Problems for Section 12.5

1. In Problem 12.1.9, carry out a goodness of fit test of the hypothesis that blocks are placed in a random order.

2.† Twelve dice were rolled 26306 times. Each time, the number of dice showing 5 or 6 uppermost was recorded. The results are summarized in the following table:

No. of 5's and 6's	0	1	2	3	4	5	6
Frequency observed	185	1149	3265	5475	6114	5194	3067

No. of 5's and 6's	7	8	9	10	11	12	Total
Frequency observed	1331	403	105	14	4	0	26306

Compute expected frequencies under the assumption that trials are independent and the dice are balanced. Test for consistency, and give a possible explanation for the poor agreement.

3. Mass-produced items are packed in cartons of 10 as they come off the assembly line. The items from 250 cartons are inspected for defects, with the following results:

Number defective	0	1	2	3	4	5	≥ 6
Frequency observed	103	81	39	19	6	2	0

Test the hypothesis that the number of defective items per carton has a binomial distribution. Can you suggest a reason that the distribution might not be binomial?

4. Test the goodness of fit of a Poisson distribution model to the data of Example 4.4.2.

5. In a biological experiment, a square millimeter of yeast culture was subdivided into 400 equal-sized squares, and the number of yeast cells in each small square was recorded. The results are summarized in the following frequency table:

Number of cells	0	1	2	3	4	5	6	≥ 7
Frequency observed	129	137	83	38	10	2	1	0

If yeast cells are randomly and uniformly distributed over the area examined, the number of yeast cells per square should have a Poisson distribution. Test whether a Poisson model is consistent with the data.

6.† According to genetic theory, blood types MM, MN, and NN should occur in a very large population with relative frequencies θ^2, $2\theta(1-\theta)$, and $(1-\theta)^2$ where θ is the (unknown) gene frequency.

(a) The observed frequencies in a sample of size 100 from the population were 33, 44, and 23, respectively. Test the goodness of fit of the model to these data.

(b) Suppose that the observed frequencies in a sample of size 400 were exactly four times those given in (a). Carry out a goodness of fit test and explain why it gives a different result than that in (a).

7. Test the goodness of fit of the model in Problem 10.1.1.

8. Test the goodness of fit of the exponential distribution model in Problem 9.6.3.

9. Test the goodness of fit of the model in Problem 9.1.10(b).

10.†(a) A city police department kept track of the number of traffic accidents involving personal injury on sixty week-day mornings. The results were as follows:

Number of accidents	0	1	2	3	4	5	≥ 6
Frequency observed	17	17	16	7	2	1	0

Is a Poisson distribution model consistent with these data?

(b) The police department also recorded the number of persons injured in traffic accidents for the same sixty mornings, with the following results:

Number injured	0	1	2	3	4	5	6	7	≥ 8
Frequency observed	17	8	9	8	10	4	2	2	0

If injuries were randomly and uniformly distributed over time, the number of injuries per morning would have a Poisson distribution. Show that this model is contradicted by the data, and indicate which of the assumptions for a Poisson process is violated.

(c) Of the 83 accidents recorded in (a), 22 occurred on Mondays, 13 on Tuesdays, 11 on Wednesdays, 12 on Thursdays, and 25 on Fridays. Are these results consistent with the hypothesis that accidents are equally likely to occur on any day of the week?

11. The following results were obtained in 150 rolls of a 6-sided die:

Side	1	2	3	4	5	6
Frequency observed	29	26	31	24	19	21

(a) For a brick-shaped die (Example 1.3.2), the face probabilities are

$$p_1 = p_2 = p_3 = p_4 = \tfrac{1}{6} + \theta; \qquad p_5 = p_6 = \tfrac{1}{6} - 2\theta.$$

Compute expected frequencies under this model, and test the goodness of fit.

(b) Assuming the model in (a) to be correct, test the hypothesis $\theta = 0$.

(c) Carry out a likelihood ratio test of the hypothesis $p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$. How is the likelihood ratio statistic related to those in (a) and (b)?

12. Fifty specimens of plastic are repeatedly struck with a hammer until they fracture. The data are summarized in the following table:

No. of blows	1	2	3	4	5	6	≥ 7
Observed frequency	23	13	8	4	1	1	0

It is thought that the number of blows needed to fracture a specimen is geometrically distributed, with probability function

$$f(x) = \theta(1-\theta)^{x-1}; \qquad x = 1, 2, \ldots; \qquad 0 < \theta < 1.$$

Find the maximum likelihood estimate of θ, and test the goodness of fit of the model to the data.

13.† A long sequence of digits (0, 1, ..., 9) produced by a random number generator was examined. There were 51 zeroes altogether, giving 50 pairs of successive zeroes. For each such pair, the number of nonzero digits between the two zeroes was determined. The results were as follows:

1	1	6	8	10	22	12	15	0	0
2	26	1	20	4	2	0	10	4	19
2	3	0	5	2	8	1	6	14	2
2	2	21	4	3	0	0	7	2	4
4	7	16	18	2	13	22	7	3	5

Describe an appropriate probability model for these counts if the random number generator is actually producing random digits. Construct a frequency table and test the goodness of fit of the model.

14. Consider the multinomial log likelihood function

$$l = f_1 \log p_1 + f_2 \log p_2 + \cdots + f_k \log p_k$$

where $\Sigma p_j = 1$ and $\Sigma f_j = n$. Show that l is maximized when $p_j = f_j/n$ for $j = 1, 2, \ldots, k$.

15. Show that, for both the likelihood ratio statistic (12.5.1) and the Pearson goodness of fit statistic (12.5.2), the effect of doubling all observed and expected frequencies is to double the value of D.

16.† n items were examined from the output of three similar machines in a factory. Ten percent were defective for the first machine, five percent for the second machine, and twelve percent for the third machine. A likelihood ratio test of the hypothesis that the probability of a defective is the same for all three machines gave a significance level of 5%. How large was n?

17. (a) Let f_1, f_2, e_1, and e_2 be any positive real numbers. Prove that

$$(f_1 + f_2) \log \frac{f_1 + f_2}{e_1 + e_2} \leq f_1 \log \frac{f_1}{e_1} + f_2 \log \frac{f_2}{e_2}.$$

Hint: Consider the function

$$l(\lambda_1, \lambda_2) = f_1 \log \lambda_1 + f_2 \log \lambda_2 - (e_1\lambda_1 + e_2\lambda_2).$$

Its restricted maximum subject to $\lambda_1 = \lambda_2$ cannot exceed its unrestricted maximum over all λ_1, λ_2.

(b) Suppose that (12.5.1) is calculated over k classes. The first two classes are then combined into a single class with observed frequency $f_1 + f_2$ and expected frequency $e_1 + e_2$ and D is recalculated. Show that the value of D cannot increase.

(c) Show that the Pearson goodness of fit statistic (12.5.2) has a similar property.

18.†(a) A population consists of a_i families with exactly i children, for $i = 0, 1, \ldots, k$. There are $\Sigma i a_i$ children in the population, and from these n children are chosen at random. Find the expected number of children in the sample who have exactly j siblings (brothers and sisters).

(b) The 1931 Canada census yields the following data

i	a_i	i	a_i	i	a_i
1	207756	5	32080	9	2859
2	156111	6	18128	10	1353
3	95779	7	10511	11	575
4	56275	8	5624	12	326

Find the expected number of children with exactly j siblings ($j = 0, 1, \ldots, 11$) in a random sample of 242 children from this population.

(c) A sociologist asked each of 242 alcoholics how many siblings he had, and the results were as follows:

No. of siblings	0	1	2	3	4	5	6	7	8	9	10	11	Total
Observed frequency	21	32	40	47	29	23	20	11	10	3	3	2	242

Test the hypothesis that the distribution of family size for alcoholics is as indicated by the 1931 census.

(d) Another possible procedure would be to compare observed values $f'_j = f_j/(j+1)$ with the population values $e'_j = 242a_j/\Sigma a_j$ using (12.5.1) or (12.5.2). Explain why this procedure is incorrect.

19.† The following table records 292 litters of mice classified according to litter size and number of females in the litter.

		Number of females				
		0	1	2	3	4
Litter size	1	8	12			
	2	23	44	13		
	3	10	25	48	13	
	4	5	30	34	22	5

Suppose that the number of females in a litter of size i is binomially distributed with parameters (i, p_i).

(a) Test the hypothesis $p_1 = p_2 = p_3 = p_4$.
(b) Assuming the hypothesis in (a) to be true, test the significance of deviations from the binomial distribution model.
(c) How would the test in (b) be affected if equality of the p_i's was not assumed?

20. Test the goodness of fit of the exponential distribution model in Problem 9.4.1(b). First, compute expected frequencies using the MLE of θ based on the original 27 measurements. Then repeat the test using the MLE of θ based on just the frequency table.

21. In Problem 10.1.11, test the hypothesis that the length of the gestation period is normally distributed. Use the approximate MLE's from Problem 10.1.11(a) in computing the expected frequencies. How would the value of the likelihood ratio statistic change if one used the exact MLE's in calculating expected frequencies?

12.6. Tests for Independence in Contingency Tables

Many interesting statistical applications involve the analysis of cross-classified frequency data. For instance, in a study to evaluate three treatments for cancer, one might classify each of n patients according to the treatment received, and also according to whether or not the patient survived a five-year period. The results could be displayed in a 3×2 array, with one row for each treatment category and one column for each survival category. The body of the table would give the number of patients in each of the six classes. A cross-tabulation of frequency data such as this is called a *contingency table.*

In the example just described, we have a two-way or two-dimensional table. If we also classified patients by sex, we would have a three-way ($3 \times 2 \times 2$) contingency table containing 12 frequencies. We shall restrict the discussion here to two-way tables only. Many examples of higher dimensional contingency tables may be found in Bishop, Fienberg, and Holland, *Discrete Multivariate Analysis*, MIT Press (1975).

A question of interest in the cancer study would be whether there is a connection or association between the column classification (survival) and

the row classification (treatment). This can be investigated by testing the hypothesis that the row and column classifications are independent. If this hypothesis is contradicted by the data, then there is evidence of an association between the two classifications.

$a \times b$ Contingency Table

As in the preceding section, we consider n independent repetitions of an experiment. However now we suppose that the outcome of each repetition is classified in two ways: according to which of the events $A_1, A_2, \ldots, A_a$ occurs, and according to which of the events $B_1, B_2, \ldots, B_b$ occurs. We assume that the A_i's (and similarly the B_j's) are mutually exclusive and exhaustive, so that each outcome belongs to exactly one of them. Altogether there are $k = ab$ possible classes A_iB_j. Let p_{ij} be the probability of class A_iB_j, and let f_{ij} be the observed frequency for this class in the n repetitions. The frequencies can be arranged in an $a \times b$ table as shown in Figure 12.6.1. We denote the ith row total by r_i and the jth column total by c_j. Note that

$$\Sigma\Sigma p_{ij} = 1; \qquad \Sigma\Sigma f_{ij} = n = \Sigma r_i = \Sigma c_j.$$

Under the assumption of independent repetitions, the distribution of the f_{ij}'s is multinomial with $k = ab$ classes. The joint probability function is

$$\binom{n}{f_{11} f_{12} \cdots f_{ab}} p_{11}^{f_{11}} p_{12}^{f_{12}} \cdots p_{ab}^{f_{ab}},$$

and the log likelihood function is

$$l(p) = \Sigma\Sigma f_{ij} \log p_{ij}$$

where $\Sigma\Sigma p_{ij} = 1$. The situation is the same as in Section 12.5 except that now we are using double subscripts. It follows from (12.5.1) that the likelihood ratio statistic for testing an hypothesis H concerning the p_{ij}'s is

$$D = 2\Sigma\Sigma f_{ij} \log(f_{ij}/e_{ij}) \tag{12.6.1}$$

where $e_{ij} = n\tilde{p}_{ij}$ is the expected frequency for class A_iB_j under the hypothesis. The degrees of freedom for testing H will be $(k-1) - q = ab - 1 - q$, where q is the number of unknown parameters which remain under H.

	B_1	B_2	...	B_b	Total
A_1	f_{11}	f_{12}	...	f_{1b}	r_1
A_2	f_{21}	f_{22}	...	f_{2b}	r_2
...	...	...	...	...	...
A_a	f_{a1}	f_{a2}	...	f_{ab}	r_a
Total	c_1	c_2	...	c_b	n

Figure 12.6.1. $a \times b$ contingency table.

Hypothesis of Independence

A question which is often of interest is whether there is any connection or association between the row and column classifications. To investigate this, we consider the hypothesis of independence,

$$H\colon P(A_iB_j) = P(A_i)P(B_j) \qquad \text{for all } i, j$$

which states that each row event A_i is independent of every column event B_j. Using the definition of conditional probability (3.4.1), we can rewrite this as

$$H\colon P(B_j|A_i) = P(B_j) \qquad \text{for all } i, j$$

which states that the probability of obtaining an observation in the jth column is the same for every row. Evidence against the independence hypothesis is evidence in favor of an association between the row and column classifications.

Under H, the unknown parameters are

$$\alpha_i = P(A_i) \qquad \text{for } i = 1, 2, \ldots, a;$$

$$\beta_j = P(B_j) \qquad \text{for } j = 1, 2, \ldots, b.$$

Since $\Sigma\alpha_i = 1$ and $\Sigma\beta_j = 1$, the number of functionally independent parameters under H is $q = (a-1) + (b-1)$, and the degrees of freedom for testing H is

$$(k-1) - q = ab - 1 - (a-1) - (b-1) = (a-1)(b-1).$$

Since $p_{ij} = \alpha_i\beta_j$, the log likelihood function is

$$\begin{aligned}\Sigma\Sigma f_{ij} \log p_{ij} &= \Sigma\Sigma f_{ij}(\log \alpha_i + \log \beta_j) \\ &= \sum_i (\log \alpha_i) \sum_j f_{ij} + \sum_j (\log \beta_j) \sum_i f_{ij} \\ &= \sum_i r_i \log \alpha_i + \sum_j c_j \log \beta_j.\end{aligned}$$

Maximizing this function subject to $\Sigma\alpha_i = 1$ and $\Sigma\beta_j = 1$ gives

$$\tilde{\alpha}_i = r_i/n; \qquad \tilde{\beta}_j = c_j/n.$$

Hence the expected frequency for class A_iB_j is

$$e_{ij} = n\tilde{p}_{ij} = n\tilde{\alpha}_i\tilde{\beta}_j = r_ic_j/n. \tag{12.6.2}$$

To obtain expected frequencies under the independence hypothesis, we multiply row totals by column totals and divide by the grand total.

Note that the e_{ij}'s have the same row and column totals as the f_{ij}'s:

$$\sum_j e_{ij} = r_i(\Sigma c_j)/n = r_i; \qquad \sum_i e_{ij} = (\Sigma r_i)c_j/n = c_j.$$

If we compute all of the expected frequencies in the upper left hand $(a-1) \times (b-1)$ subtable, we can get the expected frequencies for the last row and column of the table by subtraction from the marginal totals.

EXAMPLE 12.6.1. (R.A. Fisher, *Smoking and the Cancer Controversy*, Oliver and Boyd, 1959). Seventy-one pairs of twins were examined with respect to their smoking habits. For each pair, it was ascertained whether they were identical twins (A_1) or fraternal twins (A_2), and whether their smoking habits were alike (B_1) or unlike (B_2). The results are shown in the following 2×2 contingency table:

	Like habits	Unlike habits	Total
Identical twins	44 (39.56)	9 (13.44)	53
Fraternal twins	9 (13.44)	9 (4.56)	18
Total	53	18	71

Note that 83% of identical twin pairs have like habits, but only 50% of fraternal twins have like habits. Could such a large difference reasonably have occurred by chance, or is the probability of like habits different for the two types of twins?

We wish to know whether the probability of B_1 (like habits) could be the same for identical twins (A_1) and fraternal twins (A_2); that is, we wish to examine the hypothesis

$$H\colon P(B_1|A_1) = P(B_1|A_2) = P(B_1).$$

This is the independence hypothesis. Under H, the expected frequency for the (1, 1) cell is

$$e_{11} = r_1c_1/n = 53 \times 53/71 = 39.56.$$

The remaining expected frequencies can be found in a similar way, or by subtraction from the marginal totals. The observed value of the LR statistic (12.6.1) is

$$D_{\text{obs}} = 2\left[44 \log \frac{44}{39.56} + 9 \log \frac{9}{13.44} + \cdots\right] = 7.15.$$

The degrees of freedom for the test is $(a-1)(b-1) = 1$, and

$$\text{SL} \approx P\{\chi^2_{(1)} \geq 7.15\} < 0.01.$$

It is not reasonable to attribute the observed discrepancies to chance, and hence there is strong evidence against the independence hypothesis. The probability of like smoking habits is greater for identical twins than for fraternal twins.

Note. The main difference between the situations in Examples 12.4.1 and 12.6.1 is that the column totals $c_1 = c_2 = 44$ were fixed in advance in the former, whereas only the grand total $n = 71$ was fixed in the latter. Thus the basic model for Example 12.4.1 is a pair of independent binomial distributions, but the basic model for Example 12.6.1 is a single multinomial

distribution. If we had applied the analysis from Example 12.4.1 to the current example, we would have obtained exactly the same expected frequencies, LR statistic, degrees of freedom, and significance level. For a test of the independence hypothesis it makes no difference whether the marginal totals are random variables, or are fixed in advance by the experimental design.

EXAMPLE 12.6.2. Twenty-seven of the pairs of identical twins considered in Example 12.6.1 had been separated at birth, whereas the other 26 pairs had been raised together. The frequencies of like and unlike smoking habits for the two groups are as follows:

	Like habits	Unlike habits	Total
Separated	23 (22.42)	4 (4.58)	27
Not separated	21 (21.58)	5 (4.42)	26
Total	44	9	53

The figures in parentheses are the expected frequencies under the assumption that the two classifications are independent. We do not need a formal test of significance to tell us that the agreement is extremely good. There is no evidence that the probability of like smoking habits is different for the two groups.

The greater similarity between smoking habits of identical twins (Example 12.6.1) could be accounted for in two ways. Firstly, it could be due to the fact that identical twins have the same genotype, whereas fraternal twins are no more alike genetically than ordinary brothers and sisters. Secondly, it could be due to greater social pressures on identical twins to conform in their habits. If the latter were the case, one would expect to find less similarity in the smoking habits of identical twins who had been separated at birth. Since this is not the case, it appears that genetic factors are primarily responsible for the similarity of smoking habits.

The possibility that genetic factors may influence smoking habits has interesting implications for the smoking and cancer controversy, since these same genetic factors might also produce an increased susceptibility to cancer. See Fisher's pamphlet for further discussion.

EXAMPLE 12.6.3. In a study to determine whether laterality of hand is associated with laterality of eye (measured by astigmatism, acuity of vision, etc.), 413 subjects were classified with respect to these two characteristics. The results were as follows:

	Left-eyed	Ambiocular	Right-eyed	Total
Left-handed	34 (35.43)	62 (58.55)	28 (30.02)	124
Ambidextrous	27 (21.43)	28 (35.41)	20 (18.16)	75
Right-handed	57 (61.14)	105 (101.04)	52 (51.82)	214
Total	118	195	100	413

Is there any evidence that laterality of eye is related to laterality of hand?

Assuming that the two classifications are independent, the expected frequencies for the four classes in the upper left hand corner are

$$118 \times 124/413 = 35.43 \qquad 195 \times 124/413 = 58.55$$

$$118 \times 75/413 = 21.43 \qquad 195 \times 75/413 = 35.41$$

The remaining expected frequencies are obtained by subtraction from the marginal totals, and are shown in parentheses above. The observed value of the LR statistic (12.6.1) is $D_{\text{obs}} = 4.03$, and there are $(a-1)(b-1) = 4$ degrees of freedom for the test. Thus

$$\text{SL} \approx P\{\chi^2_{(4)} \geq 4.03\} > 0.25.$$

The hypothesis of independence is compatible with the data, and there is no evidence of an association between laterality of hand and laterality of eye.

EXAMPLE 12.6.4. Nine hundred and fifty school children were classified according to their nutritional habits and intelligence quotients, with the following results:

	Intelligence Quotient				
	<80	80–89	90–99	≥100	Total
Good nutrition	245 (252.5)	228 (233.3)	177 (173.8)	219 (209.4)	869
Poor nutrition	31 (23.5)	27 (21.7)	13 (16.2)	10 (19.6)	81
Total	276	255	190	229	950

If there is no relationship, the row and column classifications are independent, and the expected frequencies are as shown. The observed value of the LR statistic is 10.51 with $(a-1)(b-1) = 3$ degrees of freedom, giving

$$\text{SL} \approx P\{\chi^2_{(3)} \geq 10.51\} \approx 0.02.$$

The data provide reasonably strong evidence against the hypothesis of independence. Poor nutrition and a low IQ tend to occur together.

PROBLEMS FOR SECTION 12.6

1.†In December 1897 there was an outbreak of plague in a jail in Bombay. Of 127 persons who were uninoculated, 10 contracted the plague. Of 147 persons who had been inoculated, 3 contracted the disease. Test the hypothesis that contraction of the plague is independent of inoculation status.

2. It was noticed that married undergraduates seemed to do better academically than single students. Accordingly, the following observations were made: of 1500 engineering students, 297 had failed their last set of examinations; 157 of them were married, of whom only 14 had failed. Are these observations consistent with the hypothesis of a common failure rate for single and married students? Under

what conditions would the information that there were more married students in 3rd and 4th years than in 1st and 2nd years affect your conclusion?

3. Six hundred and four adult patients in a large hospital were classified according to whether or not they had cancer, and according to whether or not they were smokers. The results were as follows:

	Cancer patient	Other
Smoker	70	397
Non-smoker	12	125

Test the hypothesis that the disease classification is independent of the smoking classification.

4. A total of 1376 father–daughter pairs were classified as SS, ST, TS, or TT where S stands for short and T for tall. Heights were divided at 68″ for fathers and 64″ for daughters. The proportion of short daughters among short fathers is 522/726 while among tall fathers the proportion is 206/650. Do the data indicate any association between the heights of fathers and daughters?

5.† In a series of autopsies, indications of hypertension were found in 37% of 200 heavy smokers, in 40% of 290 moderate smokers, in 45.3% of 150 light smokers, and in 51.3% of 160 non-smokers. Test the hypothesis that the probability of hypertension is independent of the smoking category.

6. The following table classifies 5816 births by day of the week. Row 1 classifies the first 2000 births in *Who's Who* for 1970 (average year of birth 1907). Row 2 classifies the 3816 births which were announced in *The Times* during one year ending in August 1976.

	M	T	W	Th	F	Sa	Su	Total
Who's Who	262	307	270	307	272	280	302	2000
The Times	572	585	594	594	582	498	391	3816

(a) Test the hypothesis that, for the sample from *Who's Who*, births are uniformly distributed over the days of the week.
(b) Show that the distributions of births are significantly different for the two samples. In what way are they different? Can you suggest an explanation for this?

7.† Gregor Mendel grew 529 pea plants using seed from a single source, and classified them according to seed shape (round, round and wrinkled, wrinkled) and color (yellow, yellow and green, green). He obtained the following data:

38 round, yellow
65 round, yellow and green
60 round and wrinkled, yellow
138 round and wrinkled, yellow and green
28 wrinkled, yellow
68 wrinkled, yellow and green
35 round, green
67 round and wrinkled, green
30 wrinkled, green

(a) Test the hypothesis that the shape and color classifications are independent.
(b) According to Mendel's theory, the frequencies of yellow, yellow and green, and green seeds should be in the ratio 1:2:1. Test whether this hypothesis is consistent with the data.

8. In the following table, 64 sets of triplets are classified according to the age of their mother at their birth and their sex distribution:

	3 boys	2 boys	2 girls	3 girls	Total
Mother under 30	5	8	9	7	29
Mother over 30	6	10	13	6	35
Total	11	18	22	13	64

(a) Is there any evidence of an association between the sex distribution and the age of the mother?
(b) Suppose that the probability of a male birth is 0.5, and that the sexes of triplets are determined independently. Find the probability that there are x boys in a set of triplets ($x = 0, 1, 2, 3$), and test whether the column totals are consistent with this distribution.

9. 1398 school children with tonsils present were classified according to tonsil size and absence or presence of the carrier for streptococcus pyogenes. The results were as follows:

	Normal	Enlarged	Much enlarged
Carrier present	19	29	24
Carrier absent	497	560	269

Is there evidence of an association between the two classifications?

10. The following data on heights of 205 married couples were presented by Yule in 1900.

	Tall wife	Medium wife	Short wife
Tall husband	18	28	19
Medium husband	20	51	28
Short husband	12	25	9

Test the hypothesis that the heights of husbands and wives are independent.

11.† A study was undertaken to determine whether there is an association between the birth weights of infants and the smoking habits of their parents. Out of 50 infants of above average weight, 9 had parents who both smoked, 6 had mothers who smoked but fathers who did not, 12 had fathers who smoked but mothers who did not, and 23 had parents of whom neither smoked. The corresponding results for 50 infants of below average weight were 21, 10, 6, and 13, respectively.

(a) Test whether these results are consistent with the hypothesis that birth weight is independent of parental smoking habits.
(b) Are these data consistent with the hypothesis that, given the smoking habits of the mother, the smoking habits of the father are not related to birth weight?

12. In a California study, data were collected on some features of motorcycle accidents. As part of the study, questionnaires were sent to individuals who had been involved in motorcycle collisions. One question of interest was the possible relationship between the occurrence of head injury and helmet use. The following data were reported on 626 *injured* male drivers who responded to the questionnaire.

	No head injury	Minor head injury	Serious head injury
Helmet used	165	20	33
No helmet	262	53	93

(a) Is there any evidence of a difference in relative frequencies of the different injury types between the two groups (helmet versus no helmet)?
(b) Of those who received a head injury, is there any evidence of a difference in the frequency of serious versus minor head injuries for the two groups?
(c) From these data we see that, of all the injured drivers, only 218 out of 626 (35%) wore helmets. Is there any evidence in these data that wearing helmets reduces the chance of an injury in an accident?

13.† In an experiment to detect a tendency of a certain species of insect to aggregate, 12 insects were released near two adjacent leaf areas, A and B, and after a certain period of time the number of insects that had settled on each was counted. The process was repeated 10 times, using the same two leaf areas. The observations are set out below.

Trial number	1	2	3	4	5	6	7	8	9	10
Number on A	7	3	3	9	0	0	5	5	7	4
Number on B	3	5	6	1	10	8	2	5	4	6

Do the observations suggest that insects tend to aggregate, or that they distribute themselves at random over the two areas?

14. Consider a two-way cross-classification of counts f_{ij}, where $1 \le i \le a$ and $1 \le j \le b$. Assume that the f_{ij}'s are independent, and that f_{ij} has a Poisson distribution with mean μ_{ij}. Under this assumption, the total count $n = \Sigma\Sigma f_{ij}$ is a random variable. Consider the hypothesis

$$H: \mu_{ij} = \alpha_i \beta_j \gamma \qquad \text{for } 1 \le i \le a, 1 \le j \le b,$$

where the α_i's, β_j's, and γ are unknown parameters and $\Sigma\alpha_i = \Sigma\beta_j = 1$. This hypothesis says that the expected counts in any two rows of the table are proportional:

$$\mu_{1j}/\mu_{2j} = \alpha_1/\alpha_2 \qquad \text{for } 1 \le j \le b.$$

Show that the expected frequencies under H are given by (12.6.2), and the likelihood ratio statistic for testing H is given by (12.6.1).

15. Continuation of Problem 14. An experiment was carried out to determine whether two concentrations of a virus would produce different effects on tobacco plants. Half a leaf of a tobacco plant was rubbed with cheesecloth soaked in one

preparation of the virus extract, and the other half was rubbed with the second extract. The following table shows the number of lesions appearing on the half leaf.

Leaf no.	1	2	3	4	5	Total
Extract 1	31	20	18	17	9	95
Extract 2	18	17	14	11	10	70

Test the hypothesis that the proportion of lesions produced by Extract 1 is the same on all leaves.

12.7. Cause and Effect

A small significance level in a test for independence implies that the observed frequencies would have been unlikely to occur if the row and column classifications were independent of one another. Thus the data indicate some connection or association between the two classifications. However, the fact that an association has been detected does not imply that there is necessarily a direct or causative relationship between the two classifications.

The statement "A causes B" means that, by manipulating the cause A, we can control the effect B. If we make A happen, we increase the probability that B will occur (within some reasonable time limit). If we prevent A from happening, we decrease the probability that B will subsequently occur.

The statement "A and B are associated" means that A and B tend to occur together. However, there is no guarantee that forcing A to occur will have any effect on the occurrence of B. In fact, there are three possible cause-and-effect relationships which could produce the association:

(i) A causes B;
(ii) B causes A;
(iii) some other factor C causes both A and B.

We cannot claim to have proof that A causes B unless the data have been collected in such a way that (ii) and (iii) can be ruled out.

For instance, in Example 12.6.4, low IQ's were found more often in children with poor nutrition than in children with good nutrition. The significance test tells us that the observed association cannot reasonably be attributed to chance. However, it would not be valid to conclude that poor nutrition causes low IQ, or that low IQ causes poor nutrition. There could be a third factor such as poor home environment which is responsible for both poor nutrition and low IQ.

Rigorous evidence of cause-and-effect can be obtained only from a controlled experiment in which the experimenter demonstrates that by manipulating the cause A, he can control the effect B. Randomization is an

important component of the experiment. If the subjects who received A were chosen at random, then we know what caused A, and neither (ii) nor (iii) above is a possible explanation.

For instance, suppose that we wish to demonstrate that aspirin causes a reduction in the probability of a second attack for heart attack victims. The experimental subjects should be assigned at random to either the treatment group which receives aspirin, or the control group which receives a placebo (a pill similar to aspirin in appearance and flavor, but with no active ingredients). If we allowed the patients or their doctors to choose their treatments, we could not be sure that any reduction in second attacks was actually due to the aspirin rather than to the way in which the treatments were assigned.

The following example shows the sort of difficulty that can arise when treatments are not randomly assigned.

EXAMPLE 12.7.1. In order to compare two possible treatments for the same disease, hospital records for 1100 applications of each treatment were examined. The treatment was classified as a success or failure in each application and the observed frequencies were as follows:

	Treatment 1	Treatment 2
Success	595	905
Failure	505	195
Total	1100	1100

The success rate was 82% for treatment 2 versus only 54% for treatment 1, and a significance test shows that this difference is far too great to be attributed to chance.

One might be tempted at this point to assume that the relationship was causal, and that the overall success rate could be improved if treatment 2 were always used. This isn't necessarily so! We do not know that patients receiving the two treatments were similar in other ways, and therefore we cannot rule out possibility (iii) above.

For instance, it might be that treatment 1 was given primarily to patients who were seriously ill, while treatment 2 was usually given to those less seriously ill. The breakdown into serious cases and less serious cases might be as follows:

	Less serious cases		More serious cases	
	Trt 1	Trt 2	Trt 1	Trt 2
Success	95	900	500	5
Failure	5	100	500	95
Total	100	1000	1000	100

Treatment 1 has a higher success rate for both the less serious and more

serious cases. The larger total number of successes with treatment 2 is due to a third factor, illness severity, which was not considered in the original table.

Of course, there may be additional factors, such as the sex and age of the patient, which also affect success rates. A further breakdown of the data according to these factors may change the picture again. If just one important factor is overlooked, conclusions about the relative merits of the two treatments may be incorrect.

In a designed experiment, patients would first be grouped according to any factors such as disease severity which were expected to influence the success rate. The patients in a group would then be assigned to treatments at random.

Under random assignment one would expect that any important factor which had been overlooked would be reasonably well balanced over the treatment groups. An imbalance could still occur by chance. However, if the experiment is repeated, it is very unlikely that we will again obtain an imbalance of the same sort. Thus, by randomly assigning subjects to treatment groups, and by repeating the experiment, one can guard against the presence of unsuspected factors which might invalidate the conclusions.

PROBLEMS FOR SECTION 12.7

1. Explain why it is that studies such as those in Problems 12.6.3 and 12.6.5 cannot be used to establish cause-and-effect relationships.

2. In an Ontario study, 50267 live births were classified according to the baby's weight (less than or greater than 2.5 kg) and according to the mother's smoking habits (non-smoker, 1–20 cigarettes per day, or more than 20 cigarettes per day). The results were as follows:

No. of cigarettes	0	1–20	>20
Weight ≤2.5	1322	1186	793
Weight >2.5	27036	14142	5788

   (a) Test the hypothesis that birth weight is independent of the mother's smoking habits.
   (b) Explain why it is that these results do not prove that birth weights would increase if mothers stopped smoking during pregnancy. How should a study to obtain such proof be designed?
   (c) A similar, though weaker, association exists between birth weight and the amount smoked by the father. Explain why this is to be expected even if the father's smoking habits are irrelevant.

3.† 918 freshman mathematics students were classified according to their first term average on six subjects, and according to whether or not they had written a high school mathematics competition. The results were as follows:

First term average	<50	50–59	60–69	≥70
Wrote competition	10	46	128	289
Did not write	41	89	146	169

(a) Test the hypothesis that first term average is independent of competition status.
(b) Explain why it is incorrect to conclude that high school students can improve their prospects for first term at university by writing the competition.

4. One hundred and fifty Statistics students took part in a study to evaluate computer-assisted instruction (CAI). Seventy-five received the standard lecture course while the other 75 received some CAI. All 150 students then wrote the same examination. Fifteen students in the standard course and 29 of those in the CAI group received a mark over 80%.

(a) Are these results consistent with the hypothesis that the probability of achieving a mark over 80% is the same for both groups?
(b) Based on these results, the instructor concluded that CAI increases the chances of a mark over 80%. How should the study have been carried out in order for this conclusion to be valid?

5.† (a) The following data were collected in a study of possible sex bias in graduate admissions at a large university:

	Admitted	Not admitted
Male applicants	3738	4704
Female applicants	1494	2827

Test the hypothesis that admission status is independent of sex. Do these data indicate a lower admission rate for females?

(b) The following table shows the numbers of male and female applicants and the percentages admitted for the six largest graduate programs in (a):

	Men		Women	
Program	Applicants	% Admitted	Applicants	% Admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	373	35
E	191	28	393	24
F	373	6	341	7

Test the independence of admission status and sex for each program. Does any of the programs show evidence of a bias against female applicants?

(c) Why is it that the totals in (a) seem to indicate a bias against women, but the results for individual programs in (b) do not?

12.8. Testing for Marginal Homogeneity

Although the hypothesis of independence will often be of interest, one should not assume that every contingency table automatically calls for such a test. Contingency tables can arise in many ways, and the hypothesis of interest will

depend upon the situation. To illustrate this point, we give an example where one is interested in comparing the marginal probabilities $P(A_i)$ and $P(B_i)$ rather than in testing independence.

Two drugs are compared to see which of them is less likely to produce unpleasant side effects. Each of 100 subjects is given the two drugs on different occasions, and is classified according to whether or not the drugs upset his stomach. The results can be summarized in a 2×2 contingency table as follows:

	Nausea with drug B	No nausea with B	Total
Nausea with drug A	38	2	40
No nausea with A	10	50	60
Total	48	52	100

Drug B produced nausea in 48% of subjects, but drug A produced nausea in only 40% of subjects. Could this discrepancy reasonably be ascribed to chance, or is there evidence of a real difference between the two drugs?

We assume that results for different subjects are independent, so that we have $n = 100$ independent repetitions of an experiment with $k = 4$ possible outcomes. The basic model for the experiment is multinomial as in Section 12.6. However, the hypothesis of independence is not of interest in this example. One would expect a patient who experiences nausea from one drug to be more susceptible to nausea from the other drug as well. Indeed, 88 of the 100 subjects reacted in the same way to both drugs. The row and column classifications are certainly not independent, and there would be no point in testing the hypothesis of independence.

The question of interest in this example is whether the probability of nausea is the same for both drugs. Thus we consider the *hypothesis of marginal homogeneity*,

$$H: P(A) = P(B).$$

Since $P(A) = p_{11} + p_{12}$ and $P(B) = p_{11} + p_{21}$, this hypothesis is equivalent to

$$H: p_{12} = p_{21}$$

which states that the 2×2 table of probabilities (p_{ij}) is symmetric. There will be one degree of freedom for testing H because it reduces the number of unknown parameters by 1.

Under H, the log likelihood function is

$$l = \Sigma\Sigma f_{ij} \log p_{ij} = f_{11} \log p_{11} + (f_{12} + f_{21}) \log p + f_{22} \log p_{22}$$

where p is the common value of p_{12} and p_{21}, and $p_{11} + 2p + p_{22} = 1$. Maximizing l subject to this restriction gives

$$\tilde{p}_{11} = f_{11}/n; \qquad \tilde{p}_{22} = f_{22}/n; \qquad \tilde{p}_{12} = \tilde{p}_{21} = \tilde{p} = \tfrac{1}{2}(f_{12} + f_{21})/n,$$

and hence the expected frequencies are

$$e_{11} = f_{11}; \qquad e_{22} = f_{22}; \qquad e_{12} = e_{21} = \tfrac{1}{2}(f_{12} + f_{21}).$$

The following are the observed and expected frequencies for the present example:

38 (38)	2 (6)
10 (6)	50 (50)

The observed value of the LR statistic is

$$D_{\text{obs}} = 2\left[38 \log \frac{38}{38} + 2 \log \frac{2}{6} + 10 \log \frac{10}{6} + 50 \log \frac{50}{50}\right] = 5.82,$$

and therefore

$$\text{SL} \approx P\{\chi^2_{(1)} \geq 5.82\} \approx 0.02.$$

There is fairly strong evidence against the hypothesis of marginal homogeneity. The chance of nausea is significantly less with drug A than with drug B.

Note that, since $\log 1 = 0$, the diagonal terms contribute nothing to the LR statistic. The hypothesis says nothing about the probabilities on the diagonal, and so the test does not depend upon the diagonal frequencies. The hypothesis says only that the off-diagonal probabilities are equal, and so the total frequency $f_{12} + f_{21}$ for these cells should be divided equally between them.

An Alternative (Incorrect) Analysis

The marginal totals from the table above can be arranged to form a new 2×2 table:

Treatment	Drug A	Drug B
Nausea	40 (44)	48 (44)
No nausea	60 (56)	52 (56)
Total	100	100

The situation now appears similar to that in Example 12.4.1. The number of subjects Y_1 who experience nausea with drug A will have a binomial $(100, p_1)$ distribution, and the number Y_2 with drug B will have a binomial $(100, p_2)$ distribution. Under $H: p_1 = p_2$, the estimated probability of nausea is

$$\tilde{p} = \frac{40 + 48}{100 + 100} = 0.44,$$

and using this estimate we obtain the expected frequencies shown in parentheses. The observed value of the LR statistic (12.4.1) is $D_{\text{obs}} = 1.30$, and

hence

$$\text{SL} \approx P\{\chi^2_{(1)} \geq 1.30\} > 0.25.$$

According to this analysis, there is no evidence of a real difference between the two drugs.

This alternative analysis would be correct if Y_1 and Y_2 were independent. In Example 12.4.1 there were 88 different rats, 44 for each column of the table, and so it was reasonable to assume the independence of Y_1 and Y_2. However, in the present example, the same 100 subjects received both drugs. A subject who experiences nausea with drug A is likely to be affected in a similar way by drug B, and so the results in the second column of the table are not independent of those in the first column. Thus the alternative analysis is not valid in this example.

In general, it is not valid to assume that repeated observations on the same subject are independent, and care must be taken that the analysis does not depend upon the independence assumption. See Section 13.7 for further discussion of this point.

Testing Marginal Homogeneity in Larger Tables

The hypothesis of marginal homogeneity in an $a \times a$ contingency table is

$$H: P(A_i) = P(B_i) \qquad \text{for } i = 1, 2, \ldots, a.$$

This hypothesis implies that the matrix of p_{ij}'s is symmetric for $a = 2$, but not for $a > 2$. Numerical methods are required to determine the expected frequencies under H when $a > 2$. Once these have been obtained, the hypothesis may be tested using the LR statistic (12.6.1). Since H reduces the number of unknown parameters by $a - 1$, there will be $a - 1$ degrees of freedom for the test.

Problems for Section 12.8

1. Consider the function

$$l(\alpha, \beta, \gamma) = a \log \alpha + b \log \beta + c \log \gamma,$$

where $\alpha + 2\beta + \gamma = 1$. Show that this function is maximized for $\alpha = a/n$, $\beta = b/2n$, $\gamma = c/n$, where $n = a + b + c$.

2. (a) A random sample of 10000 people was taken from the Canadian labor force. Of these, 523 were unemployed. Obtain an approximate 95% confidence interval for the proportion of unemployed in the Canadian labor force.
   (b) A second random sample of 10000 people was taken from the Canadian labor force one year later. This time 577 were found to be unemployed. Is there conclusive evidence that the overall unemployment rate has changed?
   (c) Suppose that, instead of choosing a second random sample, the same 10000 people had been re-interviewed one year later. Why would the test in (b) no longer be appropriate?

3.† Of 400 randomly chosen electors in a riding, 212 said that they supported government policy and 188 were opposed. Soon after this a new budget was introduced and the same 400 electors were re-interviewed. There were found to be 196 who now supported government policy, including 17 who had previously been opposed.

(a) Explain why it would not be valid to carry out a test for independence in the following table:

	Support	Opposed
Before budget	212	188
After budget	196	204

(b) Another way to tabulate the data is as follows:

	Support after	Opposed after
Support before	179	33
Support after	17	171

Carry out a test for independence in this table, and carefully explain what the result means.

(c) Test the hypothesis that the proportion of voters who support government policy is the same after the budget as it was before the budget.

12.9. Significance Regions

In Section 11.4 we defined confidence intervals and suggested that they be constructed from the likelihood function. In this section we consider another construction based on a test of significance.

Suppose that the model involves a single unknown parameter θ, and that we have a test of significance for the hypothesis $\theta = \theta_0$. The significance level will depend upon which parameter value is tested, so we can think of SL as a function of θ. If $\mathrm{SL}(\theta_0)$ is near 1, H: $\theta = \theta_0$ is in good agreement with the data and θ_0 is a "reasonable" parameter value. If $\mathrm{SL}(\theta_0)$ is near 0, H: $\theta = \theta_0$ is strongly contradicted by the data, and θ_0 is not a reasonable parameter value. The significance level, considered as a function of θ, gives a ranking of possible parameter values.

Intervals or regions of parameter values can be obtained from $\mathrm{SL}(\theta)$ in the same way that likelihood regions are obtained from $R(\theta)$. The set of parameter values such that $\mathrm{SL}(\theta) \geq p$ is called a $100p\%$ *significance region for* θ. Significance regions are also called consonance regions. See Kempthorne and Folks, *Probability, Statistics, and Data Analysis*, Iowa State University Press (1971).

The 5% significance region for θ consists of all parameter values θ_0 such that $\mathrm{SL}(\theta_0) \geq 0.05$. Usually this will be an interval. Any parameter value θ_0 inside this region is compatible with the data at the 5% level because a test of H: $\theta = \theta_0$ gives a significance level of 5% or more. Any parameter value θ_1 outside this region is contradicted by the data at the 5% level because a test of H: $\theta = \theta_1$ gives a significance level less than 5%.

EXAMPLE 12.9.1. Suppose that X has a binomial (n, θ) distribution. The expected value of X under H: $\theta = \theta_0$, is $n\theta_0$, and so we might choose $D = |X - n\theta_0|$ as the test statistic (see Example 12.1.1). Given an observed value x, the significance level is

$$\mathrm{SL}(\theta_0) = P\{|X - n\theta_0| \geq |x - n\theta_0|\}.$$

For n large, the normal approximation to the binomial distribution gives

$$(X - n\theta_0)/\sqrt{n\theta_0(1-\theta_0)} \approx N(0, 1),$$

and hence the square of this quantity is approximately $\chi^2_{(1)}$. It follows that

$$\mathrm{SL}(\theta_0) \approx P\left\{\chi^2_{(1)} \geq \frac{(x - n\theta_0)^2}{n\theta_0(1-\theta_0)}\right\}.$$

The approximate 5% significance interval for θ based on this test contains all parameter values θ_0 such that $\mathrm{SL}(\theta_0) \geq 0.05$. Since $P\{\chi^2_{(1)} \geq 3.841\} = 0.05$, we have $\mathrm{SL} \geq 0.05$ if and only if

$$(x - n\theta_0)^2/n\theta_0(1-\theta_0) \leq 3.841$$

$$\leftrightarrow (\hat{\theta} - \theta_0)^2 \leq 3.841\theta_0(1-\theta_0)/n$$

where $\hat{\theta} = x/n$. The endpoints of the interval are thus the roots of a quadratic equation

$$(\hat{\theta} - \theta)^2 = 3.841\theta(1-\theta)/n.$$

For instance, suppose that we observe $X = 35$ in $n = 100$ trials as in Example 12.2.1. Then $\hat{\theta} = 0.35$, and the equation is

$$(0.35 - \theta)^2 = 0.03841\theta(1-\theta).$$

Its roots are $\theta = 0.2636$ and $\theta = 0.4474$, and so the approximate 5% significance interval for θ based on the above test is $0.2636 \leq \theta \leq 0.4474$.

Alternatively, we could use the likelihood ratio statistic

$$D' = -2r(\theta_0) = 2x \log\frac{x}{n\theta_0} + 2(n-x)\log\frac{n-x}{n(1-\theta_0)}$$

as in Example 12.2.1. Since $D' \approx \chi^2_{(1)}$ for n large, we have

$$\mathrm{SL}(\theta_0) \approx P\{\chi^2_{(1)} \geq -2r(\theta_0)\}.$$

We then have $SL \geq 0.05$ if and only if

$$-2r(\theta_0) \leq 3.841.$$

Solving for $n = 100$, $x = 35$ gives $0.2612 \leq \theta \leq 0.4464$ as the approximate 5% significance interval for θ. This is also a 14.7% likelihood interval and an approximate 95% confidence interval (see Section 11.4).

In order to find an exact 5% significance interval, we would need to evaluate $SL(\theta_0)$ by summing binomial (n, θ_0) probabilities as in Example 12.2.1. We would repeat this calculation for several values of θ_0, and by trial and error find the range of parameter values θ_0 such that $SL(\theta_0) \geq 0.05$. The two test statistics D, D' will give slightly different intervals.

Coverage Probabilities of Significance Regions

As in Section 11.1, we imagine a very large number of repetitions of the experiment, with θ having the same unknown value in all repetitions. At each repetition we compute a 5% significance region for θ using a significance test of $H: \theta = \theta_0$ with test statistic D, say. The coverage probability CP is the proportion of these regions which contain the true value of θ.

The true value θ_0, say, belongs to the 5% significance region if and only if a test of $H: \theta = \theta_0$ gives $SL \geq 0.05$.

If D is a continuous variate, there exists a variate value d such that $P(D \geq d) = 0.05$. The significance level will be 5% or more if and only if the observed value of D is at most d. Thus we have

$$CP(\theta_0) = P(SL \geq 0.05) = P(D \leq d) = 1 - P(D > d).$$

Since D is continuous, $P(D = d)$ is zero, and therefore

$$CP(\theta_0) = 1 - P(D \geq d) = 0.95.$$

The coverage probability is exactly 95% for all parameter values θ_0. Therefore the 5% significance region is a 95% confidence region in the continuous case.

If the probability model for the experiment is discrete then D will be a discrete variate, and there usually will not exist a variate value d such that $P(D \geq d) = 0.05$. Instead we take d to be the variate value such that $P(D \geq d) \geq 0.05$, and $P(D > d) < 0.05$. The significance level will be 5% or more if and only if the observed value of D is at most d. Thus

$$CP(\theta_0) = P(SL \geq 0.05) = P(D \leq d) = 1 - P(D > d).$$

It follows that $CP(\theta_0) > 0.95$ for all parameter values θ_0. The exact coverage probability of the 5% significance region will usually depend upon θ_0 in the discrete case, and it is always greater than 95%.

In general, the coverage probability of a $100p\%$ significance region for θ is exactly $1 - p$ if D is continuous, and greater than $1 - p$ if D is discrete.

Construction of Confidence Regions

In Section 11.4 we constructed confidence regions from the likelihood function. For instance, the 95% confidence region was taken to be the $100p\%$ likelihood region where p was chosen to give coverage probability 0.95. Regions constructed in this way have the property that each parameter value inside the region is more likely than each parameter value outside the region.

The above results on coverage probability suggest a second method of construction. We begin with a test of the hypothesis $\theta = \theta_0$ with test statistic D, say. Using this test, we determine the 5% significance region for θ and take this as the 95% confidence region. Regions constructed in this way have the property that parameter values included are compatible with the data at the 5% level, while values excluded are contradicted by the data at this level.

Significance levels, and hence significance regions, depend upon the particular test statistic D which is used in the significance test. If a different test statistic is used, the 5% significance region for θ will generally change. Likelihood regions do not depend upon the choice of a test statistic. For this reason, it seems preferable to take confidence regions directly from the likelihood function as suggested in Section 11.4. However the second construction using a significance test is widely used.

Significance Regions from Likelihood Ratio Tests

We now have two methods for obtaining confidence regions from the likelihood function. The first method is to take a likelihood region with the desired coverage probability. The second method is to obtain a significance region from the likelihood ratio test of $H: \theta = \theta_0$. Under what conditions will these two constructions produce the same region?

The likelihood ratio statistic for testing $H: \theta = \theta_0$ is

$$D = 2[l(\hat{\theta}) - l(\theta_0)] = -2r(\theta_0).$$

Let $d_p = d_p(\theta_0)$ be the largest value of D such that

$$P\{D \geq d_p(\theta_0) | \theta_0 \text{ is the true value}\} \geq p.$$

Then θ_0 belongs to the $100p\%$ significance region if and only if the observed value of D is at most $d_p(\theta_0)$. Thus the $100p\%$ significance region is given by the inequality

$$-2r(\theta_0) \leq d_p(\theta_0).$$

This defines a likelihood region if and only if $d_p(\theta_0)$ does not depend upon θ_0.

If the distribution of D is the same for all θ_0, then d_p does not depend upon θ_0. Every significance region obtained from the likelihood ratio test will be a likelihood region, and the two constructions will agree. This is the case in large samples when $D \approx \chi^2_{(1)}$ for all θ_0 (see Example 12.9.1).

If the distribution of D depends upon θ_0, as it usually will in examples with discrete distributions, then d_p will generally depend upon θ_0. Significance regions obtained from the likelihood ratio test need not be likelihood regions, and the two constructions will usually give slightly different results.

EXAMPLE 12.9.1 (continued). Consider again the binomial distribution example with $n = 100$ and X observed to be 35. We shall show that the exact 5.4% significance interval obtained from the likelihood ratio test of $H: \theta = \theta_0$ is not a likelihood interval.

The 14.7% likelihood interval (approximate 95% confidence interval) for θ is given by

$$-2r(\theta) \leq 3.841$$

and solving gives $0.26117 \leq \theta \leq 0.44642$. The two endpoints of this interval have equal relative likelihoods.

An exact likelihood ratio test of $H: \theta = 0.26117$ can be carried out as in Example 12.2.1, and the significance level is found to be 0.052. Similarly, an exact test of $H: \theta = 0.44642$ gives $\mathrm{SL} = 0.056$. Thus the exact 5.4% significance interval for θ will contain $\theta = 0.44642$, but it won't contain the equally likely value $\theta = 0.26117$. It follows that the 5.4% significance interval is not a likelihood interval.

*12.10. Power

This section briefly introduces a theory of test statistics. This theory is based on the concept of the power or sensitivity of a test statistic against an alternative hypothesis. Power comparisons may be helpful in a theoretical comparison of several possible test statistics to determine which of them is more likely to detect departures of a particular type.

Consider a test of the simple hypothesis H_0, with test statistic D. The significance level of outcome x in relation to H_0 is

$$\mathrm{SL} = P(D \geq d | H_0 \text{ is true})$$

where $d = D(x)$ is the observed value of D.

If D is a continuous variate, it is possible to obtain any significance level between 0 and 1. However if D is discrete there will be only a discrete set of possible significance levels corresponding to the possible values of D. If there exists a variate value d_α such that $P(D \geq d_\alpha | H) = \alpha$, then α is called an *achievable significance level*. Two test statistics are called *comparable* if they have the same set of achievable significance levels.

The *size α critical region* of a test is the set C_α of outcomes x for which

*This section may be omitted on first reading.

$\mathrm{SL} \leq \alpha$. If α is achievable, then $x \in C_\alpha$ if and only if $D(x) \geq d_\alpha$. It follows that, for any achievable α,

$$P(X \in C_\alpha | H_0 \text{ is true}) = P(D \geq d_\alpha | H_0) = \alpha.$$

Note the similarity with the results on coverage probabilities in Section 12.8.

Now let H_1 denote another hypothesis which is chosen to represent the kind of departure from H_0 that we wish to detect. H_0 and H_1 are called the *null hypothesis* and the *alternative hypothesis*, respectively. Initially we assume that both H_0 and H_1 are simple hypotheses, so that the probability of any outcome x can be computed numerically under H_0 and under H_1.

The *size α power* (or sensitivity) of a test statistic D with respect to the simple alternative hypothesis H_1 is

$$K_\alpha = P\{\mathrm{SL} \leq \alpha | H_1 \text{ is true}\} = P\{X \in C_\alpha | H_1\}.$$

For instance, $K_{0.05}$ is the probability that a test of H_0 using D will produce a significance level of 5% or less if in fact H_1 is true. If $K_{0.05}$ is near 1, the test statistic D is said to be powerful or sensitive against H_1, because if H_1 were true the test would almost surely give evidence that H_0 is false.

Now let D, D' be two comparable statistics for testing H_0 with power K_α, K'_α against H_1. D is said to be *more powerful* than D' against H_1 if $K_\alpha \geq K'_\alpha$ for all achievable significance levels α. A statistic D is called *most powerful* for testing H_0 against H_1 if it is more powerful than every comparable statistic D'.

EXAMPLE 12.10.1. Let $X \sim N(\mu, 1)$, and consider a test of $H_0: \mu = 0$ against $H_1: \mu = 2$. Two possible statistics for testing $\mu = 0$ are $D \equiv X$ and $D' \equiv |X|$. With test statistic D, only large positive values of X are considered to be in poor agreement with $\mu = 0$, whereas with D' both large positive and large negative values of X are considered as evidence against $\mu = 0$. Both D and D' are continuous variates, and therefore all significance levels are achievable for both statistics.

The size α critical region for D has the form $X \geq d_\alpha$ where d_α is chosen so that

$$P\{X \geq d_\alpha | H_0 \text{ is true}\} = \alpha.$$

Since X is $N(0, 1)$ under H_0, d_α is the value such that $F(d_\alpha) = 1 - \alpha$ where F is the standardized normal c.d.f. The size α power of D with respect to H_1 is

$$\begin{aligned} K_\alpha &= P\{X \geq d_\alpha | H_1 \text{ is true}\} = P\{X \geq d_\alpha | X \sim N(2, 1)\} \\ &= P\{Z \geq d_\alpha - 2 | Z \sim N(0, 1)\} = 1 - F(d_\alpha - 2). \end{aligned}$$

The size α critical region for D' has the form $|X| \geq d'_\alpha$ where d'_α is chosen so that

$$P\{|X| \geq d'_\alpha | H_0 \text{ is true}\} = \alpha.$$

Since X is $N(0, 1)$ under H_0, d'_α is the value such that $F(d'_\alpha) = 1 - \frac{\alpha}{2}$. The size α

power of D' with respect to H_1 is

$$K'_\alpha = P\{|X| \geq d'_\alpha | H_1 \text{ is true}\} = 1 - P\{|X| < d'_\alpha | X \sim N(2, 1)\}$$
$$= 1 - P\{-d'_\alpha - 2 \leq Z \leq d'_\alpha - 2 | Z \sim N(0, 1)\}$$
$$= 1 - F(d'_\alpha - 2) + F(-d'_\alpha - 2).$$

For $\alpha = 0.05$ we find from Table B2 that $d_\alpha = 1.645$ and $d'_\alpha = 1.960$. Thus we have

$$K_{0.05} = 1 - F(-0.355) = F(0.355) = 0.64;$$
$$K'_{0.05} = 1 - F(-0.004) + F(-3.960) = 0.48.$$

If $\mu = 2$ and we test the hypothesis $\mu = 0$, the probability of getting SL ≤ 0.05 is 0.64 with statistic D, but only 0.48 with statistic D'. Thus D gives us a better chance of obtaining evidence against H: $\mu = 0$ when in fact $\mu = 2$.

It can be shown that $K_\alpha \geq K'_\alpha$ for all values of α, so that D is more powerful than D' for testing H_0: $\mu = 0$ versus H_1: $\mu = 2$. In fact, it follows from the theorem below, that D is the most powerful statistic for testing $\mu = 0$ against $\mu = 2$.

Most Powerful Test when H_0 and H_1 are Simple

The following theorem, which is called the Neyman–Pearson Fundamental Lemma, yields a most powerful test statistic when both H_0 and H_1 are simple hypotheses.

Theorem 12.10.1. *Let H_0 and H_1 be simple hypotheses, and let $f_0(x)$ and $f_1(x)$ denote the probability of a typical outcome x under H_0 and under H_1, respectively. Then the statistic*

$$D(x) = f_1(x)/f_0(x) \tag{12.10.1}$$

is most powerful for testing H_0 against H_1.

PROOF. Let α be an achievable significance level for D, and let d_α be the value of D such that $P(D \geq d_\alpha | H_0) = \alpha$. The size α critical region C_α is the set of x-values for which $D(x) \geq d_\alpha$. Note that, by (12.10.1), we have

$$f_1(x) \geq d_\alpha f_0(x) \qquad \text{for } x \in C_\alpha; \tag{12.10.2}$$
$$f_1(x) < d_\alpha f_0(x) \qquad \text{for } x \in \bar{C}_\alpha. \tag{12.10.3}$$

Let C'_α be the size α critical region for any comparable test statistic D', and consider the partition of the sample space into four disjoint regions

$$S = (C_\alpha C'_\alpha) \cup (C_\alpha \bar{C}'_\alpha) \cup (\bar{C}_\alpha C'_\alpha) \cup (\bar{C}_\alpha \bar{C}'_\alpha).$$

We use p's to denote the probabilities of these regions under H_0, and q's to denote their probabilities under H_1 (see Table 12.10.1).

Table 12.10.1. Probabilities Under the Null Hypothesis and Under the Alternative Hypothesis for the Four Regions of the Sample Space Defined by Two Size α Critical Regions

H_0	C'_α	$\bar{C}'_\alpha$	Total	H_1	C'_α	$\bar{C}'_\alpha$	Total
C_α	p_{11}	p_{12}	α	C_α	q_{11}	q_{12}	K_α
$\bar{C}_\alpha$	p_{21}	p_{22}	$1-\alpha$	$\bar{C}_\alpha$	q_{21}	q_{22}	$1-K_\alpha$
Total	α	$1-\alpha$	1	Total	K'_α	$1-K'_\alpha$	1

Since C_α and C'_α are size α critical regions, we have

$$p_{11} + p_{12} = P(X \in C_\alpha | H_0) = \alpha = P(X \in C'_\alpha | H_0) = p_{11} + p_{21},$$

and hence $p_{12} = p_{21}$. The size α power is

$$K_\alpha = P(X \in C_\alpha | H_1) = q_{11} + q_{12} \qquad \text{for } D;$$
$$K'_\alpha = P(X \in C'_\alpha | H_1) = q_{11} + q_{21} \qquad \text{for } D'$$

and the difference in power is

$$K_\alpha - K'_\alpha = q_{12} - q_{21}.$$

Since $C_\alpha \bar{C}'_\alpha$ is a subset of C_α, (12.10.3) gives

$$q_{12} = \Sigma f_1(x) \geq d_\alpha \Sigma f_0(x) = d_\alpha p_{12}$$

where the sums are taken over $x \in C_\alpha \bar{C}'_\alpha$. Similarly, since $\bar{C}_\alpha C'_\alpha$ is a subset of $\bar{C}_\alpha$, (12.10.3) gives

$$q_{21} = \Sigma f_1(x) < d_\alpha \Sigma f_0(x) = d_\alpha p_{21}.$$

Now, since $p_{12} = p_{21}$, we have

$$K_\alpha - K'_\alpha = q_{12} - q_{21} > d_\alpha p_{12} - d_\alpha p_{21} = 0.$$

This result holds for all comparable statistics D' and achievable significance levels α, and hence the theorem follows. □

EXAMPLE 12.10.2. Let $X \sim N(\mu, 1)$, and consider a test of the simple null hypothesis H_0: $\mu = \mu_0$ versus the simple alternative hypothesis H_1: $\mu = \mu_1$. The theorem gives

$$D(x) = f_1(x)/f_0(x) = \exp\{-\tfrac{1}{2}(x-\mu_1)^2 + \tfrac{1}{2}(x-\mu_0)^2\}$$
$$= \exp\{x(\mu_1 - \mu_0) + \tfrac{1}{2}(\mu_0^2 - \mu_1^2)\}$$

as a most powerful statistic for testing H_0 against H_1.

If $\mu_1 > \mu_0$, large values of D correspond to large values of X. The size α critical region has the form $X \geq b_\alpha$ where b_α is chosen so that

$$P\{X \geq b_\alpha | H_0 \text{ is true}\} = \alpha.$$

Since $X \sim N(\mu_0, 1)$ under H_0, we find that $b_\alpha = \mu_0 + z_\alpha$ where z_α is the value exceeded with probability α in a standardized normal distribution. The size α power is

$$\begin{aligned} K_\alpha &= P\{X \geq b_\alpha | H_1\} = P\{X \geq b_\alpha | X \sim N(\mu_1, 1)\} \\ &= P\{Z \geq b_\alpha - \mu_1 | Z \text{ is } N(0, 1)\} \\ &= 1 - F(\mu_0 + z_\alpha - \mu_1) \end{aligned}$$

where F is the c.d.f. of $N(0, 1)$.

Similarly, if $\mu_1 < \mu_0$, the critical region has the form $X \leq a_\alpha$, and the size α power is found to be

$$K_\alpha = 1 - F(\mu_1 + z_\alpha - \mu_0).$$

In accepting only large values of X as evidence against $H: \mu = \mu_0$, we achieve maximum power against departures on the high side ($\mu_1 > \mu_0$), but we lose the ability to detect departures on the low side ($\mu_1 < \mu_0$). The cost of increased sensitivity to one particular type of departure is a decrease in sensitivity to other types. This point was discussed previously in Example 12.5.2.

The likelihood ratio statistic for testing $H: \mu = \mu_0$ is

$$D' = (X - \mu_0)^2$$

which ranks outcomes according to the magnitude of $|X - \mu_0|$. The LR statistic is not most powerful for testing $\mu = \mu_0$ against any particular alternative value μ_1, but it does have reasonably high power for departures in both directions.

Composite Alternative Hypothesis

Now consider a slightly more general problem in which H_0 is simple but H_1 is composite. Suppose that a typical outcome x has probability $f(x; \theta)$ where θ is a real-valued parameter. The null hypothesis is taken to be $H_0: \theta = \theta_0$. The alternative hypothesis has the form $H_1: \theta \in \Omega_1$ where Ω_1 is a set of possible parameter values; for instance $H_1: \theta \neq \theta_0$ and $H_1: \theta > \theta_0$ have this form.

Given any particular value $\theta_1 \in \Omega_1$, a most powerful statistic for testing $\theta = \theta_0$ versus $\theta = \theta_1$ is given by

$$D(x) = f(x; \theta_1)/f(x; \theta_0).$$

If this statistic defines the same ranking of outcomes (i.e. the same critical region) for all $\theta_1 \in \Omega_1$, then D is *uniformly most powerful* for testing H_0 versus H_1.

In Example 12.10.2, one obtains the same ranking of outcomes from smallest (most favorable) to largest (least favorable) whenever $\mu_1 > \mu_0$. Hence there exists a uniformly most powerful statistic for testing $H_0: \mu = \mu_0$ versus

$H_1: \mu > \mu_0$. Similarly, there exists a uniformly most powerful statistic for testing $H_0: \mu = \mu_0$ versus $H_1: \mu < \mu_0$. However there is no uniformly most powerful statistic for testing $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$ because a different ranking is obtained for $\mu_1 < \mu_0$ than for $\mu_1 > \mu_0$.

A similar result can be established for any distribution belonging to the exponential family (Example 15.1.6).

Discussion

1. We have seen that there generally will not exist a statistic which is uniformly most powerful for testing $H_0: \theta = \theta_0$ against a two-sided alternative $\theta \neq \theta_0$. In fact, uniformly most powerful tests will rarely exist except in simple textbook examples. To make further progress in defining a theoretically optimum test, additional restrictions must be placed on the types of test to be considered. The restrictions usually suggested seem arbitrary and unconvincing. The situation is even less satisfactory when both the null and alternative hypotheses are composite, and we shall not give details here.

2. Although power considerations will not identify an optimum test statistic except in very special cases, a comparison of power may still be helpful in choosing between two test statistics D and D'. Given a statistic D for testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$, one can determine the size α power as a function of θ_1,

$$K_\alpha(\theta_1) = P(\text{SL} \leq \alpha | \theta = \theta_1).$$

A graphical comparison of the power functions $K_\alpha(\theta)$ and $K'_\alpha(\theta)$ for selected values of α may suggest that one of the statistics is preferable.

3. Another use to which power has been put is the determination of sample size. Suppose that we intend to test $H: \theta = \theta_0$ using a test statistic D, and that we want to be 90% sure of obtaining a significance level of 5% or less if in fact $\theta = \theta_1$. Then the sample size n should be selected so that $K_{0.05}(\theta_1) = 0.9$. For another approach to experimental planning, see the discussion of expected information in Section 11.6.

4. Power comparisons are not likely to be very useful unless one can be quite specific with respect to the alternative hypothesis. In many applications of significance tests, one will have only a vague idea concerning the types of departure that may occur. One would like to avoid building elaborate models to explain these until the need for them has been demonstrated in a test of significance.

CHAPTER 13

Analysis of Normal Measurements

The normal distribution plays a central role in the modelling and statistical analysis of continuous measurements. Many types of measurements have distributions which are approximately normal, and the Central Limit Theorem helps to explain why this is so. Statistical methods for analyzing normally distributed measurements are relatively simple, and most of these methods give reasonable results under moderate departures from normality.

Section 1 discusses the basic assumptions and describes the models to be considered in later sections. All of these are examples of normal linear models, which will be discussed in greater generality in Chapter 14. Section 13.2 describes statistical methods for such models. These methods are applied to the one-sample and two-sample models in Sections 3 and 4, and to the straight line model in Sections 5 and 6. Section 7 discusses the analysis of paired measurements, such as measurements taken on the same subject before and after treatment.

13.1. Introduction

Suppose that n determinations $y_1, y_2, \ldots, y_n$ are made of the same quantity y under various different conditions. For instance, gasoline mileages achieved by a car over a fixed distance might be recorded for several driving speeds, weather conditions, etc. We wish to formulate a model which describes or explains the way in which y depends upon these conditions. Hopefully the model will help us to understand how the various factors affect mileage, and to estimate the magnitudes of their effects.

Any realistic model will have to take natural variability into account. If we

measure mileages repeatedly under conditions that are identical, or as close to identical as we can make them, we will not always get exactly the same result. There will be scatter, or variability, in observations made under identical conditions. We model this by assuming that the observations $y_1, y_2, \ldots, y_n$ are observed values of random variables $Y_1, Y_2, \ldots, Y_n$. The problem is then to determine how the probability distribution of Y_i depends upon the conditions under which this observation was made.

If the conditions are very different for Y_i than for Y_j, the probability distributions for Y_i and Y_j may be of completely different types. For instance, suppose that we are observing failure times of plastic gears at various temperatures. Gears fail due to melting at very high temperatures, whereas at low temperatures they become brittle and tend to fracture. There is no reason to suppose that the distributions of lifetimes will be similar at these two extremes.

In most studies we deal with relatively small changes in conditions. Then we expect the distributions of $Y_1, Y_2, \ldots, Y_n$ to be similar to one another. Thus we might reasonably assume that the Y_i's all have the same type of distribution, and that the effect of changing conditions is to alter the value of a parameter in this distribution. This is the sort of assumption we made in Section 10.5, where we were examining the dependence of the response rate on the dose of a drug. We assumed that all of the Y_i's were independent and binomially distributed, and that the only effect of changing the dose was to alter the response probability p.

The Basic Assumptions

In this chapter and the next one, we develop the model and analysis under the assumption that the Y_i's are independent and normally distributed with the same variance σ^2, so that

$$Y_i \sim N(\mu_i, \sigma^2) \qquad \text{for } i = 1, 2, \ldots, n. \tag{13.1.1}$$

Under these assumptions, the effect of changing conditions is to alter μ_i, the expected value of Y_i, but the shape and spread of the distribution are not affected.

The model can also be written in terms of n independent error variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, where

$$\varepsilon_i \equiv Y_i - \mu_i \sim N(0, \sigma^2).$$

We can then write Y_i as the sum of a "target value" μ_i and a "random error" or "noise" component ε_i:

$$Y_i \equiv \mu_i + \varepsilon_i \qquad \text{where } \varepsilon_i \sim N(0, \sigma^2). \tag{13.1.2}$$

If we take repeated measurements under the same conditions, then μ_i stays the same, but the random error ε_i varies from one repetition to the next.

The standard deviation σ measures the amount of random variability (scatter, noise) that one would expect to obtain in repeated measurements taken under the same conditions. Suppose that Y_i and Y_i' are independent measurements with the same expected value:

$$Y_i \sim N(\mu_i, \sigma^2);\ Y_i' \sim N(\mu_i, \sigma^2) \qquad \text{independent.}$$

Then by (6.6.7) we have

$$Y_i - Y_i' \sim N(0, 2\sigma^2).$$

After standardizing, we consult Table B2 to obtain

$$P\{|Y_i - Y_i'| \geq \sigma\} = P\left\{|Z| \geq \frac{1}{\sqrt{2}}\right\} = 0.48.$$

If the experiment is repeated under identical conditions, we expect about half the measurements to change by more than σ.

If $\sigma = 0$, then $\varepsilon_i = 0$ and $Y_i = \mu_i$ with probability one. In this case, repeating the experiment would produce exactly the same measurements $y_1, y_2, \ldots, y_n$. Most real experiments do involve some natural variability, and so $\sigma > 0$.

We are assuming that σ is the same for all conditions under which observations are being taken. This means that all n measurements $y_1, y_2, \ldots, y_n$ are made with the same precision, and therefore they should be treated equally in the analysis. If the Y_i's had unequal variances, it would be necessary to modify the analysis so that greater weight was given to the more precise observations. This can be done easily provided that all of the variance ratios $\text{var}(Y_i)/\text{var}(Y_j)$ are known.

Note that we are assuming the independence of $Y_1, Y_2, \ldots, Y_n$. Whether or not this is a reasonable assumption depends upon the way in which the observations are made. For instance, it is usually not appropriate to assume that repeat measurements on the same subject or specimen are independent (see Section 13.7).

The normality assumption is less critical than the assumptions of independence and equal variances. Most of the methods developed for normally distributed measurements will give reasonable results under moderate departures from normality. The normal distribution provides a good first approximation in many applications, and the Central Limit Theorem (Section 6.7) helps to explain why this is so.

If the original measurements are decidedly non-normal, it may still be possible to use the normal analysis, provided that a suitable nonlinear transformation is first applied to the y_i's. For instance, lifetime distributions usually have a long tail to the right, and are far from normal in shape. Also, there is usually more variability in the data under conditions which produce long lifetimes than under conditions which produce short lifetimes. It would be inappropriate to assume normality and equal variances in such cases. A common practice is to apply the normal analysis to the logarithms of the

observed lifetimes. The distribution of log-lifetimes is generally much closer to normal in shape, and the log transformation also helps to stabilize the variance.

Assumptions Concerning $\mu_1, \mu_2, \ldots, \mu_n$

The basic model (13.1.1) involves $n + 1$ parameters $\mu_1, \mu_2, \ldots, \mu_n$ and σ. It is not possible to estimate all $n + 1$ parameters with only n observations. Unless some of the parameter values are known, it is necessary to simplify the model by reducing the number of unknown parameters. We do this by expressing the means $\mu_1, \mu_2, \ldots, \mu_n$ as functions of q parameters, where $q < n$. The form of the assumptions concerning $\mu_1, \mu_2, \ldots, \mu_n$ will depend upon the conditions under which the observations were made.

In the simplest case, we have n measurements $y_1, y_2, \ldots, y_n$ which were all taken under the same conditions. The y_i's should differ from one another only because of random variation, and so we assume that

$$\mu_1 = \mu_2 = \cdots = \mu_n = \alpha.$$

In this case the μ_i's are written as functions of a single unknown parameter α, so $q = 1$. This is the one-sample model (see Section 13.3).

Alternatively, we might have two groups or samples of measurements — for instance, data for males and data for females. We might be willing to assume that observations belonging to the same sample differ from one another only because of random error. We could then take

$$\mu_i = \alpha \qquad \text{for measurements in group 1;}$$

$$\mu_i = \alpha + \beta \qquad \text{for measurements in group 2.}$$

The μ_i's are expressed as functions of two unknown parameters, so $q = 2$. This is the two-sample model which will be considered in Section 13.4.

In Sections 13.5 and 13.6 we shall consider the straight-line model

$$\mu_i = \alpha + \beta x_i \qquad \text{for } i = 1, 2, \ldots, n$$

where $x_1, x_2, \ldots, x_n$ are known constants. For instance, this model might be appropriate if the y_i's were blood pressure measurements for n subjects, and the x_i's were their ages. The μ_i's are functions of two unknown parameters α and β, so again we have $q = 2$.

Linear Models

Each of the models described above has the property that the μ_i's are expressed as *linear* functions of the q unknown parameters $\alpha, \beta, \ldots$. Models with this property are called linear models (see Chapter 14).

Because all of the models considered in this chapter are linear models, their analysis is very similar. The general form of the analysis will be described in the next section. Before proceeding, the reader may wish to review the material on the χ^2, t, and F distributions in Sections 6.9 and 6.10.

13.2. Statistical Methods

In this section, we describe statistical methods for normal linear models. These methods will be applied to the one-sample, two-sample, and straight line models in the following sections. See Chapter 14 for derivations and additional examples of normal linear models.

Under the basic model (13.1.1) or (13.1.2), the n measurements $y_1, y_2, \ldots, y_n$ are observed values of independent random variables $Y_1, Y_2, \ldots, Y_n$, where

$$Y_i \sim N(\mu_i, \sigma^2) \qquad \text{for } i = 1, 2, \ldots, n.$$

The joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$ is

$$\begin{aligned} f(y_1, y_2, \ldots, y_n) &= \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\Sigma(y_i - \mu_i)^2\right\}. \end{aligned}$$

If the measurement intervals are all small (see Section 9.4), the log likelihood function is

$$l = -n \log \sigma - \frac{1}{2\sigma^2}\Sigma(y_i - \mu_i)^2. \tag{13.2.1}$$

We also assume that the μ_i's can be written as linear functions of q unknown parameters $\alpha, \beta, \ldots$. The number of unknown parameters and the equations relating the μ_i's to $\alpha, \beta, \ldots$ will depend upon the situation (see Section 13.1).

To find the MLE's $\hat{\alpha}, \hat{\beta}, \ldots$, we substitute for the μ_i's in (13.2.1) and maximize. In order to maximize l by choice of $\alpha, \beta, \ldots$, we need to minimize the sum of squares

$$S = \Sigma(y_i - \mu_i)^2 = \Sigma\varepsilon_i^2 \tag{13.2.2}$$

where $\varepsilon_i = y_i - \mu_i$ is the error associated with the ith measurement. Because $\hat{\alpha}, \hat{\beta}, \ldots$ are chosen to minimize the sum of squares of the errors, they are also referred to as *least squares estimates.*

Note that S does not depend upon σ. Since $\hat{\alpha}, \hat{\beta}, \ldots$ may be found by minimizing S, they do not depend upon σ. The MLE's of $\alpha, \beta, \ldots$ are the same

whether or not the value of σ is known. For this reason, we can treat σ as known in obtaining $\hat{\alpha}, \hat{\beta}, \ldots$, and then adjust later in the analysis for the case in which σ is unknown.

Since the μ_i's are assumed to be linear functions of $\alpha, \beta, \ldots$, the sum of squares S is of the second degree in these parameters, and the derivatives of S are linear in $\alpha, \beta, \ldots$. Thus we can find $\hat{\alpha}, \hat{\beta}, \ldots$ by solving the simultaneous linear equations

$$\frac{\partial S}{\partial \alpha} = 0; \qquad \frac{\partial S}{\partial \beta} = 0; \ldots. \tag{13.2.3}$$

Iterative methods are not required with normal linear models.

The equations (13.2.3) will be linear in the measurements $y_1, y_2, \ldots, y_n$ as well as in the parameters $\alpha, \beta, \ldots$. As a result, each of the estimates $\hat{\alpha}, \hat{\beta}, \ldots$ will be a linear combination of the y_i's. We shall use this fact below in discussing significance tests and confidence intervals.

Fitted Values and Residuals

Each of the μ_i's is a linear function of the parameters $\alpha, \beta, \ldots$. Upon replacing these parameters by their estimates $\hat{\alpha}, \hat{\beta}, \ldots$, we obtain $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n$, the MLE's of $\mu_1, \mu_2, \ldots, \mu_n$. These are called the *fitted values.*

The difference between the ith measurement y_i and its estimated mean (fitted value) $\hat{\mu}_i$ is

$$\hat{\varepsilon}_i = y_i - \hat{\mu}_i.$$

This is called the ith *residual,* and it provides an estimate of $\varepsilon_i = y_i - \mu_i$, the random error associated with the ith measurement.

The n residuals $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n$ provide two types of information: information about the adequacy of the model, and information about the magnitude of σ.

The adequacy of the model is assessed by plotting the residuals and examining them for unusual patterns. See Section 14.5 for a discussion of residual plots.

Inferences about σ are based on the residual sum of squares, $\Sigma\hat{\varepsilon}_i^2$. It can be shown that, if the model is correct, then

$$\frac{1}{\sigma^2}\Sigma\hat{\varepsilon}_i^2 \sim \chi^2_{(n-q)} \tag{13.2.4}$$

where q is the number of unknown parameters $\alpha, \beta, \ldots$ in the model. Furthermore, $\Sigma\hat{\varepsilon}_i^2$ is distributed independently of $\hat{\alpha}, \hat{\beta}, \ldots$. See Section 14.6 for derivations of these results.

Maximizing the full log likelihood function (13.2.1) over σ and $\alpha, \beta, \ldots$ gives $\hat{\sigma}^2 = (1/n)\Sigma\hat{\varepsilon}_i^2$. This estimate is unsatisfactory when q is large. It does not allow for the fact that, since q parameters $\alpha, \beta, \ldots$ are estimated from the data, there are effectively only $n - q$ observations for estimating σ^2 (see Section 10.8).

The usual estimate of σ^2 is

$$s^2 = \frac{1}{n-q}\Sigma\hat{\varepsilon}_i^2 = \frac{\text{numerator of } \chi^2 \text{ quantity}}{\text{degrees of freedom}}. \tag{13.2.5}$$

We shall show at the end of this section that s^2 is the MLE of σ^2 based on the marginal distribution of $\Sigma\hat{\varepsilon}_i^2$.

Inferences for α, β, ...: σ^2 Known

We noted earlier that each of the estimates $\hat{\alpha}$, $\hat{\beta}$, ... is a linear combination of the y_i's. Thus we have

$$\hat{\alpha} = a_1 y_1 + a_2 y_2 + \cdots + a_n y_n$$

for some constants $a_1, a_2, \ldots, a_n$.

Since the y_i's are assumed to be observed values of independent normal variates, it follows by (6.6.7) that the sampling distribution of $\hat{\alpha}$ is normal. It can be shown that $E(\hat{\alpha}) = \alpha$, and (5.5.7) gives

$$\text{var}(\hat{\alpha}) = \Sigma a_i^2 \text{ var}(Y_i) = \sigma^2 c$$

where $c = \Sigma a_i^2$. It follows that

$$Z \equiv \frac{\hat{\alpha} - \alpha}{\sqrt{\sigma^2 c}} \sim N(0, 1). \tag{13.2.6}$$

If σ^2 is known, inferences about α are based on (13.2.6). To test $H: \alpha = \alpha_0$, we compute the observed value of Z when $\alpha = \alpha_0$, and then find

$$\text{SL} = P\{|Z| \geq |Z_{\text{obs}}|\} = P\{\chi^2_{(1)} \geq Z^2_{\text{obs}}\}.$$

It can be shown that the likelihood ratio statistic for testing $H: \alpha = \alpha_0$ is Z^2, and so the procedure just described is the likelihood ratio test.

To construct a 95% confidence interval for α when σ is known, we note from Table B1 or B2 that

$$P\{-1.96 \leq Z \leq 1.96\} = 0.95.$$

Substituting for Z and solving gives

$$\hat{\alpha} - 1.96\sqrt{\sigma^2 c} \leq \alpha \leq \hat{\alpha} + 1.96\sqrt{\sigma^2 c}$$

as the 95% confidence interval. This is also a 5% significance interval: it consists of all parameter values α_0 such that a likelihood ratio test of $H: \alpha = \alpha_0$ gives SL ≥ 0.05. It is also a likelihood or maximum likelihood interval for α.

Inferences for α, β, ...: σ^2 Unknown

Usually σ^2 is unknown, and is estimated by s^2 as defined in (13.2.5). By (13.2.4) we have

$$V \equiv \frac{(n-q)s^2}{\sigma^2} \equiv \frac{1}{\sigma^2}\Sigma\hat{\varepsilon}_i^2 \sim \chi^2_{(n-q)}.$$

Since $\Sigma\hat{\varepsilon}_i^2$ is distributed independently of $\hat{\alpha}$, it follows that V is distributed independently of Z in (13.2.6).

When σ^2 is unknown, we consider the quantity

$$T \equiv \frac{\hat{\alpha} - \alpha}{\sqrt{s^2 c}}$$

which we get by replacing σ^2 by s^2 in (13.2.6). Note that

$$T \equiv \frac{\hat{\alpha} - \alpha}{\sqrt{\sigma^2 c}} \div \sqrt{\frac{s^2}{\sigma^2}} \equiv Z \div \sqrt{\frac{V}{n-q}}$$

where $Z \sim N(0, 1)$ and $V \sim \chi^2_{(n-q)}$, independently of Z. It follows by (6.10.5) that T has Student's distribution with $n-q$ degrees of freedom:

$$T \equiv \frac{\hat{\alpha} - \alpha}{\sqrt{s^2 c}} \sim t_{(n-q)}. \tag{13.2.7}$$

Note that T has the same degrees of freedom as the variance estimate s^2.

To test $H: \alpha = \alpha_0$, we compute the observed value of T when $\alpha = \alpha_0$, and then use Table B3 to find

$$\text{SL} = P\{|t_{(n-q)}| \geq |T_{\text{obs}}|\}.$$

Alternatively, we have

$$\text{SL} = P\{t^2_{(n-q)} \geq T^2_{\text{obs}}\} = P\{F_{1,\,n-q} \geq T^2_{\text{obs}}\}$$

by (6.10.7), which can be evaluated from Table B5 for the F distribution. It can be shown that the likelihood ratio statistic for testing $H: \alpha = \alpha_0$ is an increasing function of T^2, and hence the procedure just described is the likelihood ratio test.

To construct a 95% confidence interval for α, we use Table B3 to find the value t such that

$$P\{-t \leq t_{(n-q)} \leq t\} = 0.95.$$

Substituting from (13.2.7) and solving gives

$$\hat{\alpha} - t\sqrt{s^2 c} \leq \alpha \leq \hat{\alpha} + t\sqrt{s^2 c}$$

as the 95% confidence interval. This is also a 5% significance interval and a maximum likelihood interval for α.

Inferences for σ

It can be argued that, when the parameters $\alpha, \beta, \ldots$ are unknown, the residual sum of squares $\Sigma\hat{\varepsilon}_i^2$ carries all of the information from the y_i's concerning σ. Inferences about σ will therefore be based on the marginal distribution of $\Sigma\hat{\varepsilon}_i^2$.

By (13.2.4) we have

$$V \equiv \frac{1}{\sigma^2}\Sigma\hat{\varepsilon}_i^2 \equiv \frac{(n-q)s^2}{\sigma^2} \sim \chi^2_{(n-q)}.$$

By (6.9.1), the p.d.f. of V is

$$f(v) = k_v v^{(\nu/2)-1} e^{-v/2} \qquad \text{for } v > 0,$$

where $\nu = n - q$ and k_ν is a constant. If we now change variables using (6.1.11), we find that the p.d.f. of $\Sigma\hat{\varepsilon}_i^2$ is

$$f(v)\cdot\left|\frac{dv}{d\Sigma\hat{\varepsilon}_i^2}\right| = k_\nu v^{(\nu/2)-1} e^{-v/2}\cdot\frac{1}{\sigma^2}$$

$$= k_\nu\left[\frac{\nu s^2}{\sigma^2}\right]^{(\nu/2)-1}\exp\left\{-\frac{\nu s^2}{2\sigma^2}\right\}\cdot\frac{1}{\sigma^2}$$

Based on this distribution, the log likelihood function of σ is

$$l(\sigma) = -\nu\log\sigma - \frac{\nu s^2}{2\sigma^2} \qquad \text{for } \sigma > 0.$$

Setting $l'(\sigma) = 0$ gives $\sigma = s$. Thus s is the MLE of σ, and s^2 is the MLE of σ^2, based on the marginal distribution of the residual sum of squares.

The log relative likelihood function of σ is

$$r(\sigma) = l(\sigma) - l(s)$$

$$= -\nu\log\sigma - \frac{\nu s^2}{2\sigma^2} + \nu\log s + \frac{\nu}{2}$$

$$= -\frac{\nu}{2}\left[\frac{s^2}{\sigma^2} - 1 - \log\frac{s^2}{\sigma^2}\right] \tag{13.2.8}$$

where $\nu = n - q$. We can plot $r(\sigma)$ to obtain likelihood intervals for σ, and we can use the fact that

$$D \equiv -2r(\sigma) \approx \chi^2_{(1)} \tag{13.2.9}$$

to test hypotheses or to obtain approximate confidence intervals for σ.

Note that $r(\sigma)$ has the same form as $r(\theta)$ in Example 11.3.3. The results given in Table 11.3.3 for $n = 1, 2, 3, \ldots$ are applicable to the present situation with $\nu = 2, 4, 6, \ldots$. The χ^2 approximation (13.2.9) is accurate enough for most practical purposes even when ν is as small as 2.

To obtain an approximate 95% confidence interval for σ, we note from Table B4 that

$$P\{\chi^2_{(1)} \le 3.841\} = 0.95.$$

We then solve the inequality

$$-2r(\sigma) \le 3.841,$$

either by plotting $r(\sigma)$ or by using Newton's method as in Section 9.8. The interval obtained is also a 14.7% likelihood interval and an approximate 5% significance interval for σ.

Alternately, a 95% confidence interval for σ can be obtained directly from (13.2.4). From Table B4, we find values a, b such that

$$P\{\chi^2_{(n-q)} \le a\} = P\{\chi^2_{(n-q)} \ge b\} = 0.025.$$

We then have

$$P\left\{a \le \frac{(n-q)s^2}{\sigma^2} \le b\right\} = 0.95,$$

and therefore the interval

$$\frac{(n-q)s^2}{b} \le \sigma^2 \le \frac{(n-q)s^2}{a}$$

has coverage probability 0.95.

The second construction involves less arithmetic than the first, but it does not produce a likelihood interval. The interval will include some values of σ^2 at the high end which are less likely than values excluded at the lower end (see Problem 11.4.10 and Example 13.3.2). For this reason, the first construction based on the likelihood ratio statistic is preferable.

Problems for Section 13.2

1. Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent $N(\alpha, \sigma^2)$.
   (a) Show that, if σ is known, the likelihood ratio statistic for testing a value of α is $D \equiv Z^2$, where Z is defined in (13.2.6) and $c = 1/n$.
   (b) Show that, if σ is unknown, the likelihood ratio statistic for testing a value of α is given by
   $$D \equiv n\log\left[1 + \frac{1}{n-1}T^2\right]$$
   where T is defined in (13.2.7) and $c = 1/n$.
2. Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent $N(\alpha x_i, \sigma^2)$ variates where $x_1, x_2, \ldots, x_n$ are known constants. Show that the results of the preceding problem still hold, but with $c = 1/\Sigma x_i^2$.
   Hint: You will need to show that
   $$\Sigma(y_i - \alpha x_i)^2 = \Sigma(y_i - \hat{\alpha}x_i)^2 + (\hat{\alpha} - \alpha)^2\Sigma x_i^2.$$

3. Let $Y_1, Y_2, \ldots, Y_n$ be independent variates, with

$$Y_i \sim N(\mu_i, \sigma^2/w_i) \qquad \text{for } i = 1, 2, \ldots, n,$$

where $w_1, w_2, \ldots, w_n$ are known positive constants. The μ_i's are assumed to be linear functions of q unknown parameters $\alpha, \beta, \ldots$.

(a) Derive the log likelihood function, and show that $\hat{\alpha}, \hat{\beta}, \ldots$ are the parameter values which minimize the weighted sum of squares $\Sigma w_i(y_i - \mu_i)^2$.
(b) Show that $\hat{\alpha}, \hat{\beta}, \ldots$ are linear combinations of the y_i's.

Note: This type of model might be used when the measurements $y_1, y_2, \ldots, y_n$ are not made with equal precision. Observations made with high precision are given large weights w_i, while less precise observations are given smaller weights. The estimates $\hat{\alpha}, \hat{\beta}, \ldots$ are called weighted least squares estimates. Because of (b), the statistical methods described in this section can be extended easily to the weighted least squares case. It can be shown that $\Sigma w_i \hat{\varepsilon}_i^2/\sigma^2 \sim \chi^2_{(n-q)}$, and so the appropriate variance estimate is

$$s^2 = \frac{1}{n-q}\Sigma w_i \hat{\varepsilon}_i^2.$$

13.3. The One-Sample Model

As in Sections 13.1 and 13.2, we consider n measurements $y_1, y_2, \ldots, y_n$. These are modelled as observed values of n independent random variables $Y_1, Y_2, \ldots, Y_n$, where

$$Y_i \sim N(\mu_i, \sigma^2) \qquad \text{for } i = 1, 2, \ldots, n.$$

We also assume that the μ_i's are linear functions of q unknown parameters $\alpha, \beta, \ldots$.

In this section we consider the simplest case, in which all n measurements were taken under the same conditions. Then the y_i's should differ from one another only because of random variation, and their expected values $\mu_1, \mu_2, \ldots, \mu_n$ should all be equal. Thus we consider the one-sample model

$$\mu_1 = \mu_2 = \cdots = \mu_n = \alpha. \tag{13.3.1}$$

The μ_i's are expressed as functions of a single unknown parameter α, and so $q = 1$.

Substituting for the μ_i's in (13.2.2) gives

$$S = \Sigma(y_i - \mu_i)^2 = \Sigma(y_i - \alpha)^2;$$

$$\frac{\partial S}{\partial \alpha} = -2\Sigma(y_i - \alpha) = -2[\Sigma y_i - n\alpha].$$

Note that the derivative of S is linear in both α and the observations $y_1, y_2, \ldots, y_n$.

Setting the derivative equal to zero gives

$$\hat{\alpha} = \frac{1}{n}\Sigma y_i = \bar{y}.$$

The MLE of α is the sample mean $\bar{y}$. Note that $\hat{\alpha}$ is a linear combination of the y_i's:

$$\hat{\alpha} = a_1 y_1 + a_2 y_2 + \cdots + a_n y_n$$

where $a_1 = a_2 = \cdots = a_n = \frac{1}{n}$. The sampling distribution of $\hat{\alpha}$ is $N(\alpha, \sigma^2 c)$, where

$$c = a_1^2 + a_2^2 + \cdots + a_n^2 = n\left(\frac{1}{n}\right)^2 = \frac{1}{n}$$

(see (6.6.8)).

Upon replacing α by $\hat{\alpha}$ in (13.3.1), we obtain the fitted values

$$\hat{\mu}_1 = \hat{\mu}_2 = \cdots = \hat{\mu}_n = \hat{\alpha} = \bar{y}.$$

Thus the ith residual is

$$\hat{\varepsilon}_i = y_i - \hat{\mu}_i = y_i - \bar{y},$$

and the residual sum of squares is

$$\Sigma \hat{\varepsilon}_i^2 = \Sigma(y_i - \bar{y})^2.$$

This is called the *corrected sum of squares* of the y_i's, and is often denoted by S_{yy}.

The one-sample model (13.3.1) replaces n parameters $\mu_1, \mu_2, \ldots, \mu_n$ by a single parameter α, thus reducing the number of unknown parameters by $n-1$. There are $n-1$ degrees of freedom for estimating σ^2. By (13.2.5), the variance estimate is

$$s^2 = \frac{1}{n-1}\Sigma(y_i - \bar{y})^2, \tag{13.3.2}$$

and this is called the *sample variance.*

To calculate s^2 by computer, one evaluates the n residuals $y_i - \bar{y}$, sums their squares, and then divides by $n-1$. For hand calculation, the following identities are useful:

$$S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - n\bar{y}^2 = \Sigma y_i^2 - \frac{1}{n}(\Sigma y_i)^2. \tag{13.3.3}$$

The latter two formulas must be used cautiously, because they are highly susceptible to roundoff errors.

Having obtained $\hat{\alpha}$, c, and s^2, one can use the methods described in Section 13.2 to make inferences about α and σ.

EXAMPLE 13.3.1. A standard drug produces blood pressure increases which are normally distributed with mean $\mu = 22$ and standard deviation $\sigma = 9.2$. A new drug is also expected to produce normally distributed increases, but with possibly different values of μ and σ. The new drug was given to ten individuals, and it produced the following blood pressure increases:

18 27 23 15 18 15 18 20 17 8.

Is there evidence that $\mu \neq 22$? that $\sigma \neq 9.2$?

SOLUTION. We assume that the measurements $y_1, y_2, \ldots, y_{10}$ are observed values of independent $N(\mu_i, \sigma^2)$ random variables, and that

$$\mu_1 = \mu_2 = \cdots = \mu_{10} = \alpha.$$

We find that

$$n = 10; \Sigma y_i = 179; \Sigma y_i^2 = 3433; \hat{\alpha} = \bar{y} = 17.9$$

$$S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{1}{n}(\Sigma y_i)^2 = 228.9$$

$$s^2 = \frac{1}{n-1}\Sigma(y_i - \bar{y})^2 = 25.4$$

$$\text{var}(\hat{\alpha}) = c\sigma^2 \qquad \text{where } c = \frac{1}{n} = 0.1.$$

The variance estimate has $n - 1 = 9$ degrees of freedom.

When σ is unknown, inferences about α are based on

$$T \equiv \frac{\hat{\alpha} - \alpha}{\sqrt{s^2 c}} \sim t_{(9)}.$$

Note that T has the same degrees of freedom as s^2. To test H: $\alpha = 22$, we calculate

$$T_{\text{obs}} = \frac{\hat{\alpha} - 22}{\sqrt{s^2 c}} = -2.57;$$

$$\text{SL} = P\{|t_{(9)}| \geq 2.57\} \approx 0.03$$

from Table B3. There is evidence that the mean increase for the new drug is different from 22.

By (13.2.8), the likelihood ratio statistic for testing H: $\sigma = \sigma_0$ is

$$D \equiv -2r(\sigma_0) \equiv \nu\left[\frac{s^2}{\sigma_0^2} - 1 - \log\frac{s^2}{\sigma_0^2}\right],$$

where here $\nu = n - 1 = 9$. Taking $\sigma_0 = 9.2$ and $s^2 = 25.4$, we find that $D_{\text{obs}} = 4.53$. Since $D \approx \chi^2_{(1)}$ if H is true, Table B4 gives

$$\text{SL} \approx P\{\chi^2_{(1)} \geq 4.53\} \approx 0.03.$$

There is evidence that $\sigma \neq 9.2$.

The new drug produces a lower mean increase than the standard drug. It is also less variable in its effect. This is an advantage, because the effect of the new drug on an individual can be more accurately predicted. □

EXAMPLE 13.3.2. Eight plastic gears were tested at 21°C until they failed. The times to failure (in millions of cycles) were as follows:

2.37 2.01 2.47 2.20 1.87 2.32 2.00 2.86

A common assumption in such situations is that the logarithms of the failure times are normally distributed. The natural logarithms of the eight times listed above are as follows:

0.863 0.698 0.904 0.788 0.626 0.842 0.693 1.051

Assuming these to be independent observations from $N(\mu, \sigma^2)$, find 95% confidence intervals for μ and σ.

SOLUTION. The 8 log failure times $y_1, y_2, \ldots, y_8$ are assumed to be observed values of independent $N(\mu_i, \sigma^2)$ variates with

$$\mu_1 = \mu_2 = \cdots = \mu_8 = \mu.$$

This is the one-sample model (13.3.1), except that μ rather than α is used to denote the common mean. Proceeding as in the Example 13.3.1, we obtain

$$n = 8; \hat{\mu} = \bar{y} = 0.8081; s^2 = 0.01876;$$

$$\text{var}(\hat{\mu}) = \sigma^2 c \qquad \text{where } c = \frac{1}{n} = \frac{1}{8}.$$

The variance estimate has $n - 1 = 7$ degrees of freedom.

Inferences concerning μ are based on

$$T \equiv \frac{\hat{\mu} - \mu}{\sqrt{s^2 c}} \sim t_{(7)}.$$

From Table B3, we find that

$$P\{-2.365 \leq t_{(7)} \leq 2.365\} = 0.95.$$

Hence the 95% confidence interval for μ is

$$\mu \in \hat{\mu} \pm 2.365\sqrt{s^2 c} = 0.8081 \pm 0.1145;$$

that is, $0.6936 \leq \mu \leq 0.9226$. If we tested H: $\mu = \mu_0$ for any parameter value μ_0 in this interval, we would obtain a significance level of 5% or more. Also, parameter values inside the interval have higher maximum relative likelihood than values outside.

To obtain an approximate 95% confidence interval for σ, we use the fact that

$$D \equiv -2r(\sigma) \approx \chi^2_{(1)}.$$

(a) Assuming that the counts are independent $N(\mu, \sigma^2)$, find estimates of μ and σ^2. Hence estimate the probability of a negative count under this model.
(b) A more reasonable assumption in this example is that the (natural) logarithms of the counts are independent normal. Repeat (a) under this assumption.

8. It is sometimes convenient to relocate and rescale measurements $y_1, y_2, \ldots, y_n$ before analysis. The new measurements are then given by $z_i = (y_i - a)/b$ where a, b are known constants. Suppose that the y_i's are modelled as independent $N(\mu, \sigma^2)$. How should the z_i's be modelled? How can estimates and confidence intervals for μ and σ^2 be obtained from an analysis of the z_i's?

13.4. The Two-Sample Model

Consider $n = n_1 + n_2$ independent measurements in two samples. The first sample consists of n_1 measurements $y_{11}, y_{12}, \ldots, y_{1n_1}$ recorded under one set of conditions, and the second sample consists of n_2 measurements $y_{21}, y_{22}, \ldots, y_{2n_2}$ recorded under a different set of conditions. For instance, the y_{1j}'s might be blood pressure increases for n_1 subjects who received drug A, while the y_{2j}'s are increases for n_2 different subjects who received drug B.

As in the preceding sections, we model the y_{ij}'s as observed values of independent normal random variables Y_{ij} with the same variance σ^2:

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2) \qquad \text{for } j = 1, 2, \ldots, n_i;\ i = 1, 2.$$

The two-sample model is as follows:

$$\left.\begin{aligned} \mu_{11} = \mu_{12} = \cdots = \mu_{1n_1} = \mu_1; \\ \mu_{21} = \mu_{22} = \cdots = \mu_{2n_2} = \mu_2. \end{aligned}\right\} \tag{13.4.1}$$

This model states that observations within a sample differ from one another only because of random variation. There are $q = 2$ unknown parameters, μ_1 and μ_2.

Another way to write the two-sample model is

$$\left.\begin{aligned} \mu_{11} = \mu_{12} = \cdots = \mu_{1n_1} = \alpha \\ \mu_{21} = \mu_{22} = \cdots = \mu_{2n_2} = \alpha + \beta \end{aligned}\right\} \tag{13.4.2}$$

where $\alpha = \mu_1$ and $\beta = \mu_2 - \mu_1$. An advantage of this parametrization is that the difference in the means $\mu_2 - \mu_1$, which is usually the quantity of primary interest, is explicitly represented by the parameter β.

Upon substituting (13.4.1) into (13.2.2), we find that the error sum of squares is

$$\begin{aligned} S &= \Sigma\Sigma(y_{ij} - \mu_{ij})^2 \\ &= \sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2. \end{aligned}$$

Setting the derivatives of S equal to zero and solving gives

$$\hat{\mu}_1 = \bar{y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j};$$

$$\hat{\mu}_2 = \bar{y}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j}.$$

It now follows that

$$\hat{\alpha} = \hat{\mu}_1 = \bar{y}_1; \hat{\beta} = \hat{\mu}_2 - \hat{\mu}_1 = \bar{y}_2 - \bar{y}_1.$$

The fitted values and residuals are given by

$$\hat{\mu}_{ij} = \hat{\mu}_i = \bar{y}_i;$$

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{\mu}_{ij} = y_{ij} - \bar{y}_i.$$

The residual sum of squares is

$$\begin{aligned} \Sigma\Sigma\hat{\varepsilon}_{ij}^2 &= \Sigma\Sigma(y_{ij} - \bar{y}_i)^2 \\ &= \Sigma(y_{1j} - \bar{y}_1)^2 + \Sigma(y_{2j} - \bar{y}_2)^2 \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 \end{aligned}$$

where s_1^2 and s_2^2 are the two sample variances:

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2;$$

$$s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2.$$

Since there are $q = 2$ unknown parameters μ_1 and μ_2 (or α and β), there are $n - 2$ degrees of freedom for variance estimation, and (13.2.5) gives

$$s^2 = \frac{1}{n-2} \Sigma\Sigma\hat{\varepsilon}_{ij}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

The combined or pooled estimate s^2 based on both samples is a weighted average of the two sample variances s_1^2 and s_2^2, with weights equal to their degrees of freedom $n_1 - 1$ and $n_2 - 1$.

Inferences for $\beta = \mu_2 - \mu_1$

For inferences about β, we follow the procedures described in Section 13.2. First we find the sampling distribution of $\hat{\beta} = \bar{y}_2 - \bar{y}_1$. Note that

$$\bar{Y}_1 \sim N\left(\mu_1, \frac{1}{n_1}\sigma^2\right); \bar{Y}_2 \sim N\left(\mu_2, \frac{1}{n_2}\sigma^2\right).$$

Since $P\{\chi^2_{(1)} \le 3.841\} = 0.95$, the required interval is found by solving the inequality

$$-2r(\sigma) \le 3.841.$$

Since $\nu = n - 1 = 7$ and $s^2 = 0.01876$, (13.2.8) gives

$$-2r(\sigma) = 7\left[\frac{0.01876}{\sigma^2} - 1 - \log\frac{0.01876}{\sigma^2}\right].$$

By plotting this function or using Newton's method, we find the interval to be $0.088 \le \sigma \le 0.258$. This is also a 14.7% likelihood interval for σ, and an approximate 5% significance interval for σ.

Alternatively, we can construct a 95% confidence interval for σ by using the fact that

$$\frac{(n-1)s^2}{\sigma^2} \equiv \frac{1}{\sigma^2}\Sigma\hat{\varepsilon}_i^2 \sim \chi^2_{(n-1)}.$$

Here $n - 1 = 7$, and Table B4 gives

$$P\{\chi^2_{(7)} \ge 16.01\} = P\{\chi^2_{(7)} \le 1.690\} = 0.025.$$

It follows that

$$P\left\{1.690 \le \frac{7s^2}{\sigma^2} \le 16.01\right\} = 0.95,$$

and therefore the interval

$$\frac{7s^2}{16.01} \le \sigma^2 \le \frac{7s^2}{1.690}$$

has coverage probability 0.95. Upon substituting for s^2 and taking square roots, we obtain $0.091 \le \sigma \le 0.279$ as the 95% confidence interval for σ.

Note that the second construction does not produce a likelihood interval. The interval $0.091 \le \sigma \le 0.279$ includes values at the high end which are less likely than some values excluded at the lower end. The first construction does produce a likelihood interval, and it is therefore preferable. □

Problems for Section 13.3

1. The following are the initial velocities in meters per second of seven projectiles fired from the same gun:

   451 447 454 450 454 449 452

   Assuming that velocity is normally distributed, obtain a 90% confidence interval for the mean velocity μ.

2. The relationship between parental age and the incidence of mongolism was discussed in Section 7.5. It is known that, in a certain population, the mean age of mothers at normal births is 31.25. The average age of the mothers at 50 births of mongolian children was 37.25 years, with sample variance $s^2 = 49.35$. Are these observations consistent with the hypothesis $\mu = 31.25$?

3.† Under a special diet, twelve rats made the following weight gains (in grams) from birth to age three months:

55.3	54.8	65.9	60.7	59.4	62.0
62.1	58.7	64.5	62.3	67.6	61.2

Assuming that weight gains are independent $N(\mu, \sigma^2)$, obtain 95% confidence intervals for μ and for σ^2. Use two methods to find the confidence interval for σ^2, and compare the results.

4.† A manufacturer wishes to determine the mean breaking strength μ of string "to within a pound", which we interpret as requiring that the 95% confidence interval for μ should have length at most 2 pounds. If measurements are independent $N(\mu, \sigma^2)$, and if ten preliminary measurements gave $\Sigma(x_i - \bar{x})^2 = 80$, how many additional measurements would you advise the manufacturer to make?

5. Sixteen packages are randomly selected from the production of a detergent packaging machine. Their weights (in grams) were as follows:

287	293	295	295	297	298	299	300
300	302	302	303	306	307	308	311

It may be assumed that weights are independent $N(\mu, \sigma^2)$.

(a) Determine 95% confidence intervals for μ and σ.
(b) Assuming that μ and σ are equal to their estimates, find an interval which contains the weight of a new randomly chosen package with probability 0.95.

6. Ten steel ingots chosen at random from a large shipment gave the following hardness measures:

71.7	71.1	68.0	69.6	69.1
69.4	68.8	70.4	69.3	68.2

If the manufacturing process is under control, the hardness measures should be independent $N(\mu, \sigma^2)$ with $\sigma^2 = 1.2$.

(a) Are the ten observations consistent with the hypothesis $\sigma^2 = 1.2$?
(b) Assuming that $\sigma^2 = 1.2$, find a 90% confidence interval for μ.
(c) Find a 90% confidence interval for μ if σ^2 is unknown and must be estimated from the data.

7.† The following are trypanosome counts (in thousands) in cattle seven days after infection:

17.0	2.1	1.7	44.2	5.1	2.9	3.5	19.6
28.0	7.0	17.1	0.7	34.5	13.0	1.5	5.2
9.0	5.9	3.9	11.5	14.5	16.2	33.3	12.2

Furthermore, $\bar{Y}_1$ and $\bar{Y}_2$ are independent because $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are independent of $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$. It follows by (6.6.7) that

$$\bar{Y}_2 - \bar{Y}_1 \sim N\left(\mu_2 - \mu_1, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

Hence the sampling distribution of $\hat{\beta}$ is $N(\beta, \sigma^2 c)$, where

$$c = \frac{1}{n_1} + \frac{1}{n_2}.$$

We now standardize and replace σ^2 by s^2 to get

$$T \equiv \frac{\hat{\beta} - \beta}{\sqrt{s^2 c}} \sim t_{(n-2)}.$$

Here s^2 is the pooled variance estimate with $n-2$ degrees of freedom, and T has the same degrees of freedom as s^2. We can now test an hypothesis $\beta = \beta_0$ or find confidence intervals for β as in Section 13.2.

EXAMPLE 13.4.1. Cuckoos lay their eggs in the nests of other birds. Table 13.4.1 gives the lengths in millimeters of $n = 24$ cuckoos' eggs found in nests of reed-warblers and wrens. The data are from a paper by O.H. Latter in *Biometrika* (1902). The table also shows the two sample means and sample variances.

The average length of cuckoos' eggs is 22.20 in the first sample, and only 21.12 in the second sample. It appears as though the lengths of cuckoos' eggs may depend upon the locations in which they are found. On the other hand, since only 24 eggs were measured, it may be that the observed difference between $\bar{y}_1$ and $\bar{y}_2$ is due merely to random variation. We wish to determine whether the observed difference is too great to be attributed to random variation.

We model the 24 measurements as observed values of independent normal variates with the same variance. We assume that the expected length is μ_1 for

Table 13.4.1. Lengths in Millimeters of 24 Cuckoos' Eggs

Sample 1			Sample 2				
Eggs from reed-warblers' nests			Eggs from wrens' nests				
21.2	21.6	21.9	19.8	20.0	20.3	20.8	20.9
22.0	22.0	22.2	20.9	21.0	21.0	21.0	21.2
22.8	22.9	23.2	21.5	22.0	22.0	22.1	22.3
$n_1 = 9$; $\bar{y}_1 = 22.20$			$n_2 = 15$; $\bar{y} = 21.12$				
$s_1^2 = 0.4225$ (8 d.f.)			$s_2^2 = 0.5689$ (14 d.f.)				

the first sample, and μ_2 for the second sample. This is the two-sample model. We wish to know whether $H: \mu_1 = \mu_2$, or equivalently $H: \beta = 0$, is consistent with the data.

Inferences about β are based on

$$T \equiv \frac{\hat{\beta} - \beta}{\sqrt{s^2 c}} \sim t_{(n-2)}$$

where $\hat{\beta} = \bar{y}_2 - \bar{y}_1 = -1.08$, $c = \frac{1}{9} + \frac{1}{15}$, $n = 24$, and s^2 is the pooled variance estimate:

$$s^2 = \frac{8s_1^2 + 14s_2^2}{8 + 14} = 0.5156 \text{ (22 d.f.)}.$$

To test $H: \beta = 0$, we compute

$$T_{\text{obs}} = \frac{\hat{\beta} - 0}{\sqrt{s^2 c}} = -3.57;$$

$$\text{SL} = P\{|t_{(22)}| \geq 3.57\} \approx 0.002$$

from Table B3. If μ_1 and μ_2 were equal, a difference as large as that observed would very rarely occur, and so there is strong evidence that $\mu_1 \neq \mu_2$.

EXAMPLE 13.4.2. The log-lifetimes of 8 plastic gears tested at 21°C were analyzed in Example 13.3.2. The following are the log-lifetimes of 4 additional gears tested at 30°C:

$$0.364 \quad 0.695 \quad 0.558 \quad 0.359.$$

It may be assumed that log-lifetimes are independent and normally distributed with the same variance σ^2, but with expected value μ_1 at 21°C and μ_2 at 30°C. Find 95% confidence intervals for μ_1, $\mu_2 - \mu_1$, and σ^2.

SOLUTION. The sample means and variances are as follows:

Sample 1 (21°C)

$$n_1 = 8 \qquad \bar{y}_1 = 0.8081 \qquad s_1^2 = 0.01876 \text{ (7 d.f.)}.$$

Sample 2 (30°C)

$$n_2 = 4 \qquad \bar{y}_2 = 0.4940 \qquad s_2^2 = 0.02654 \text{ (3 d.f.)}.$$

The pooled variance estimate is

$$s^2 = \frac{7s_1^2 + 3s_2^2}{7 + 3} = 0.02109 \text{ (10 d.f.)}.$$

The sampling distribution of $\hat{\mu}_1 = \bar{y}_1$ is $N(\mu_1, \sigma^2 c)$, where $c = 1/n_1$. Inferences about μ_1 are based on

$$T \equiv \frac{\hat{\mu}_1 - \mu_1}{\sqrt{s^2 c}} \sim t_{(n-2)}$$

where s^2 is the pooled variance estimate with $n-2=10$ degrees of freedom. From Table B3 we find that

$$P\{-2.228 \le t_{(10)} \le 2.228\} = 0.95$$

and hence the 95% confidence interval is

$$\mu_1 \in \hat{\mu}_1 \pm 2.228\sqrt{s^2 c} = 0.8081 \pm 0.1144.$$

This differs from the construction given in Example 13.3.2 because now we are using *both* samples to estimate σ^2. Thus s^2 is different, and there are more degrees of freedom for T.

Next we consider $\beta = \mu_2 - \mu_1$. The estimate $\hat{\beta} = \bar{y}_2 - \bar{y}_1$ has sampling distribution $N(\beta, \sigma^2 c)$, where now $c = \frac{1}{8} + \frac{1}{4}$. Inferences about β are based on

$$T \equiv \frac{\hat{\beta} - \beta}{\sqrt{s^2 c}} \sim t_{(10)},$$

and the 95% confidence interval for β is

$$\beta \in \hat{\beta} \pm 2.228\sqrt{s^2 c} = -0.3140 \pm 0.1981;$$

that is, $0.1159 \le \mu_1 - \mu_2 \le 0.5121$.

Two methods of constructing confidence intervals for σ were described at the end of Section 13.2. We shall use the first method, which is based on (13.2.9). Here we have $s^2 = 0.02109$ with $\nu = 10$ degrees of freedom, so

$$D = -2r(\sigma) = 10\left[\frac{0.02109}{\sigma^2} - 1 - \log\frac{0.02109}{\sigma^2}\right].$$

We solve the inequality $-2r(\sigma) \le 3.841$ to obtain

$$0.0991 \le \sigma \le 0.2425$$

as the (approximate) 95% confidence interval for σ. □

More Than Two Samples

A similar model and analysis can be developed when there are more than two samples. Suppose that there are $n = n_1 + n_2 + \cdots + n_k$ measurements in k samples. We assume that all n measurements are independent normal with variance σ^2, and that the n_i measurements in the ith sample have expected value $\mu_i (i = 1, 2, \ldots, k)$. Let $\bar{y}_i$ and s_i^2 denote the sample mean and variance for the ith sample. Then $\hat{\mu}_i = \bar{y}_i$, and the pooled variance estimate is

$$s^2 = \frac{1}{n-k}\Sigma\Sigma(y_{ij} - \bar{y}_i)^2 = \frac{(n_1-1)s_1^2 + \cdots + (n_k-1)s_k^2}{(n_1-1) + \cdots + (n_k-1)}.$$

There are k unknown parameters $\mu_1, \mu_2, \ldots, \mu_k$, and therefore $n-k$ degrees of freedom for variance estimation.

The comments in Section 9.2 concerning the pooling of estimates are relevant here. It is usually not a good idea to combine estimates $s_1^2, s_2^2, \ldots, s_k^2$ which are significantly different from one another. A formal test for homogeneity can be carried out as in Section 12.3. For details, see the problems at the end of this section.

EXAMPLE 13.4.3. Table 13.4.2 shows data from a study on a pulse-jet pavement breaker using nozzles of three different diameters. The measurement is the penetration (in millimeters) of a concrete slab produced by a single discharge. The table also gives the sample mean $\bar{y}_i$ and sample variance s_i^2 for each sample. Since $n_1 = n_2 = n_3 = 9$, each of the sample variances has 8 degrees of freedom.

Let y_{ij} denote the jth observation for the ith nozzle type, where $i = 1, 2, 3$ and $j = 1, 2, \ldots, 9$. We assume that the y_{ij}'s are observed values of independent $N(\mu_{ij}, \sigma^2)$ variates, and that measurements made with the same type of nozzle have the same expected value:

$$\mu_{i1} = \mu_{i2} = \cdots = \mu_{i9} = \mu_i \qquad \text{for } i = 1, 2, 3.$$

This is the three-sample model with $n_1 = n_2 = n_3 = 9$. The variance estimate for this model is

$$s^2 = \frac{8s_1^2 + 8s_2^2 + 8s_3^2}{8+8+8} = \frac{3360}{24} = 140,$$

with $n-3 = 24$ degrees of freedom.

Suppose that we are interested in $\mu_3 - \mu_2$, which is the difference in expected penetration for large and medium nozzles. This is estimated by $\bar{y}_3 - \bar{y}_2$, which has sampling distribution $N(\mu_3 - \mu_2, \sigma^2 c)$ where $c = \frac{1}{9} + \frac{1}{9}$. Inferences about $\mu_3 - \mu_2$ are based on

$$T \equiv \frac{(\bar{y}_3 - \bar{y}_2) - (\mu_3 - \mu_2)}{\sqrt{s^2 c}} \sim t_{(24)}.$$

In particular, the 95% confidence interval is

$$\mu_3 - \mu_2 \in \bar{y}_3 - \bar{y}_2 \pm 2.064\sqrt{s^2 c} = 1.16 \pm 11.51.$$

Since this interval contains zero, a test of $H: \mu_3 - \mu_2 = 0$ would give a large

Table 13.4.2. Penetration of Concrete for 27 Discharges of a Jet Pulse Machine with Three Nozzle Sizes

Nozzle	Penetration (in millimeters)									$\bar{y}_i$	s_i^2
Small	67	47.5	46	62.5	49	53.5	42	55.5	39	51.33	85.00
Medium	88	60	72	73.5	62	72.5	73.5	44	54.5	66.67	167.38
Large	83	53	87	71	78	51.5	68	58	61	67.83	167.63

significance level. There is no evidence of a difference in expected penetration for medium and large nozzles.

PROBLEMS FOR SECTION 13.4.

1. The following are measurements of ultimate tensile strength (UTS) for twelve specimens of insulating foam of two different densities.

High density	98.5	105.5	111.6	114.5	126.5	127.1
Low density	79.7	84.5	85.2	98.0	105.2	113.6

Assuming normality and equality of variances, obtain 95% confidence intervals for the common variance and for the difference in mean UTS at the two densities.

2.†Twenty-seven measurements of yield were made on two industrial processes, with the following results:

Process 1:	$n_1 = 11$	$\bar{y}_1 = 6.23$	$s_1^2 = 3.79$
Process 2:	$n_2 = 16$	$\bar{y}_2 = 12.74$	$s_2^2 = 4.17$

Assuming that the yields are normally distributed with the same variance, find 95% confidence intervals for the means μ_1 and μ_2, and for the difference in means $\mu_1 - \mu_2$.

3. An experiment to determine the effect of a drug on the blood glucose concentration of diabetic rats gave the following results:

Control rats	2.05	1.82	2.00	1.94	2.12	
Treated rats	1.71	1.37	2.04	1.50	1.69	1.83

Test the hypothesis that the treatment has no effect on mean blood glucose concentration. State the assumptions upon which this test depends.

4. An experiment to discover the movement of an antibiotic in a certain variety of broad bean plants was carried out by treating 10 cut shoots and 10 rooted plants for 18 hours with a solution of the antibiotic. Assay results giving the concentration of the antibiotic per gram of plant weight are given below.

Cut shoots	55	65	61	48	57	58	60	68	52	63
Rooted plants	53	48	50	39	43	44	46	56	35	51

Assuming that concentrations are independent normal with constant variance, obtain a 99% confidence interval for the difference in mean concentration for the two types of plant.

5. *Testing equality of variances.* Let $s_1^2, s_2^2, \ldots, s_k^2$ be independent variance estimates, where

$$\nu_i s_i^2/\sigma_i^2 \sim \chi^2_{(\nu_i)} \qquad \text{for } i = 1, 2, \ldots, k.$$

(a) Find the joint log likelihood function of $\sigma_1, \sigma_2, \ldots, \sigma_k$, and show that it is maximized for $\sigma_i^2 = s_i^2$ $(i = 1, 2, \ldots, k)$.

(b) Show that, if $\sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma$, then the MLE of σ^2 is given by

$$s^2 = (\Sigma \nu_i s_i^2)/(\Sigma \nu_i).$$

(c) Show that the likelihood ratio statistic for testing $H: \sigma_1 = \sigma_2 = \cdots = \sigma_k$ is given by

$$D = \Sigma \nu_i \log(s^2/s_i^2).$$

Note: The distribution of D is approximately $\chi^2_{(k-1)}$ if H is true.

6.†Fourteen welded girders were cyclically stressed at 1900 pounds per square inch, and the numbers of cycles to failure were observed. The sample mean and variance of the log failure times were $\bar{y} = 14.564$ and $s^2 = 0.0914$. Similar tests on four additional girders with repaired welds gave $\bar{y} = 14.291$ and $s^2 = 0.0422$. Log failure times are assumed to be independent and normally distributed.

(a) Test the hypothesis that the variance of log failure time is the same for repaired welds as for normal welds.
(b) Assuming equal variances, obtain a 90% confidence interval for the difference in mean log failure time.

7. A common final examination was written by 182 honors mathematics students, of whom 61 were in the co-operative program. The results were as follows:

Co-op students	$n_1 = 61$	$\bar{y}_1 = 68.30$	$s_1 = 10.83$
Others	$n_2 = 121$	$\bar{y}_2 = 65.93$	$s_2 = 15.36$

(a) Assuming that the examination marks are normally distributed, determine whether the variances are significantly different for the two groups.
(b) Estimate the proportion of students obtaining a mark of 90 or more, and the proportion obtaining 50 or more, for each group.

8. Readings produced by a set of scales are independent and normally distributed about the true weight with constant variance. Six weighings of each of two objects gave the following results:

Object 1:	4.83	4.98	4.91	4.96	5.05	4.93
Object 2:	3.02	2.95	2.83	3.00	3.02	3.09

(a) Test the hypothesis that the variance in the readings is the same for the two objects.
(b) Assuming a common variance, obtain a 90% confidence interval for the difference in weights of the two objects.

9.†The following are the distances traveled (in miles) by 15 rockets used to test 3 different fuels.

Fuel 1	16.2	17.3	17.0	16.6	17.4
Fuel 2	18.6	18.6	19.0	19.5	20.0
Fuel 3	19.7	19.4	20.0	19.2	18.9

(a) Test the hypothesis that the variability in distance traveled is the same for the three fuels.

(b) Assuming equal variances, obtain a 95% confidence interval for the common variance.
(c) Find a 95% confidence interval for the difference in mean distance traveled for fuels 2 and 3.
(d) State the assumptions upon which the analysis in (a), (b), and (c) depends.

10. Let $s_1^2, s_2^2, \ldots, s_k^2$ be independent variate estimates as in problem 5 above. Consider the hypothesis

$$H: \sigma_i^2 = \sigma^2/w_i \qquad \text{for } i = 1, 2, \ldots, n,$$

where $w_1, w_2, \ldots, w_k$ are known positive constants and σ^2 is unknown.

(a) Show that, under H, the MLE of σ^2 is

$$s^2 = (\Sigma v_i w_i s_i^2)/(\Sigma v_i).$$

(b) Derive the likelihood ratio statistic for testing H.

11. Consider the two-sample model (13.4.1) with σ known. Show that the likelihood ratio statistic for testing $H: \mu_1 = \mu_2$ is $D \equiv Z^2$, where

$$Z \equiv (\bar{Y}_1 - \bar{Y}_2)/\sqrt{c\sigma^2}; \qquad c = \frac{1}{n_1} + \frac{1}{n_2}.$$

Hence show that, if H is true, D has a χ^2 distribution with one degree of freedom.

13.5. The Straight Line Model

Consider n measurements $y_1, y_2, \ldots, y_n$, but now suppose that each measurement y_i has associated with it a value x_i of another variable, and that the x-values can be used to explain or predict the corresponding y-values. We call x the explanatory variable or independent variable, and y the response variable or dependent variable.

For instance, $y_1, y_2, \ldots, y_n$ could be blood pressure increases for n subjects who received doses $x_1, x_2, \ldots, x_n$ of a drug. Or $y_1, y_2, \ldots, y_n$ might be gasoline mileages achieved by a car in n tests at driving speeds $x_1, x_2, \ldots, x_n$. Or $y_1, y_2, \ldots, y_n$ might be log lifetimes of n plastic gears tested at temperatures $x_1, x_2, \ldots, x_n$. In each case, knowledge of the x-value will help to predict or explain the value of y.

The x_i's will be treated as known constants in the analysis, and the y_i's will be modelled as observed values of random variables $Y_1, Y_2, \ldots, Y_n$. We shall assume that the Y_i's are independent and normally distributed with the same variance σ^2, so that

$$Y_i \sim N(\mu_i, \sigma^2) \qquad \text{for } i = 1, 2, \ldots, n.$$

The means $\mu_1, \mu_2, \ldots, \mu_n$ will then be modelled as functions of the explanatory variable x.

In many applications it is reasonable to assume that the dependence of

$E(Y)$ on x is linear. Even if the relationship is nonlinear, a straight line will often give a satisfactory approximation over a restricted range of x-values. It is advisable to plot the data in order to see whether a straight line model will be satisfactory (see Example 13.5.3).

Under the straight-line model, we have

$$\mu_i = \alpha + \beta x_i \qquad \text{for } i = 1, 2, \ldots, n. \tag{13.5.1}$$

There are $q = 2$ unknown parameters, the intercept α and the slope β.

For historical reasons, the straight line model is also called a *simple linear regression model.* The origin of the term "regression" is explained in Section 7.5.

Estimation of α and β

Upon substituting (13.5.1) into (13.2.2), we find that the error sum of squares is

$$S = \Sigma(y_i - \mu_i)^2 = \Sigma(y_i - \alpha - \beta x_i)^2.$$

The derivatives of S are

$$\frac{\partial S}{\partial \alpha} = -2\Sigma(y_i - \alpha - \beta x_i); \qquad \frac{\partial S}{\partial \beta} = -2\Sigma x_i(y_i - \alpha - \beta x_i).$$

Putting $\partial S/\partial \alpha = 0$ gives

$$\Sigma y_i - n\hat{\alpha} - \hat{\beta}\Sigma x_i = 0.$$

We divide by n and solve for $\hat{\alpha}$ to obtain

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \tag{13.5.2}$$

We now put $\partial S/\partial \beta$ equal to zero and substitute for $\hat{\alpha}$ to obtain

$$\begin{aligned} 0 &= \Sigma x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) \\ &= \Sigma x_i(y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i) \\ &= \Sigma x_i(y_i - \bar{y}) - \hat{\beta}\Sigma x_i(x_i - \bar{x}). \end{aligned}$$

It follows that

$$\hat{\beta} = \frac{\Sigma x_i(y_i - \bar{y})}{\Sigma x_i(x_i - \bar{x})} = \frac{S_{xy}}{S_{xx}}. \tag{13.5.3}$$

The numerator in (13.5.3) is called the corrected sum of products, and it can be rewritten in several forms:

$$S_{xy} = \Sigma(y_i - \bar{y})x_i = \Sigma(x_i - \bar{x})y_i = \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

$$= \Sigma x_i y_i - n\bar{x}\bar{y} = \Sigma x_i y_i - \frac{1}{n}(\Sigma x_i)(\Sigma y_i).$$

The denominator is the corrected sum of squares of the x_i's:

$$S_{xx} = \Sigma(x_i - \bar{x})x_i = \Sigma(x_i - \bar{x})^2$$
$$= \Sigma x_i^2 - n\bar{x}^2 = \Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2.$$

Variance Estimation

The fitted values and residuals are given by

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta}x_i = \bar{y} + \hat{\beta}(x_i - \bar{x});$$
$$\hat{\varepsilon}_i = y_i - \hat{\mu}_i = (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}).$$

The residual sum of squares is

$$\Sigma\hat{\varepsilon}_i^2 = \Sigma[(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})]^2$$
$$= \Sigma(y_i - \bar{y})^2 - 2\hat{\beta}\Sigma(x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}^2\Sigma(x_i - \bar{x})^2$$
$$= S_{yy} - 2\hat{\beta}S_{xy} + \hat{\beta}^2 S_{xx}.$$

Since $\hat{\beta} = S_{xy}/S_{xx}$, it follows that

$$\Sigma\hat{\varepsilon}_i^2 = S_{yy} - \hat{\beta}S_{xy}. \tag{13.5.4}$$

Since there are $q = 2$ unknown parameters α and β, there are $n - 2$ degrees of freedom for variance estimation, and (13.2.5) gives

$$s^2 = \frac{1}{n-2}\Sigma\hat{\varepsilon}_i^2. \tag{13.5.5}$$

Formula (13.5.4) is useful for hand calculation, but it is susceptible to large roundoff errors. For calculation by computer, it is better to evaluate the residuals $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n$, square them, and sum to get $\Sigma\hat{\varepsilon}_i^2$.

EXAMPLE 13.5.1. The following table gives the age (x) and systolic blood pressure (y) for each of $n = 12$ women:

x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

The data are plotted in Figure 13.5.1. The graph shows a roughly linear increase in blood pressure with age. The amount of scatter about the line does not show any systematic change with x, and so the assumption of constant variance σ^2 is reasonable.

We assume that the y's are observed values of independent $N(\mu_i, \sigma^2)$ variates and that the straight line model (13.5.1) holds. From the data we obtain

$$\Sigma x_i = 628 \qquad \Sigma y_i = 1684$$
$$\Sigma x_i^2 = 34416 \qquad \Sigma y_i^2 = 238822 \qquad \Sigma x_i y_i = 89894.$$

From these we find the sample means and corrected sums:

$$\bar{x} = 52.33 \qquad \bar{y} = 140.33$$
$$S_{xx} = 1550.67 \qquad S_{yy} = 2500.67 \qquad S_{xy} = 1764.67.$$

Now we have

$$\hat{\beta} = S_{xy}/S_{xx} = 1.138; \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 80.78.$$

The fitted line $y = 80.78 + 1.138x$ is shown in Figure 13.5.1.

By (13.5.4), the residual sum of squares is

$$\Sigma\hat{\varepsilon}_i^2 = S_{yy} - \hat{\beta}S_{xy} = 492.47.$$

The estimate of the variance about the line is

$$s^2 = \frac{1}{n-2}\Sigma\hat{\varepsilon}_i^2 = 49.247$$

with $n - 2 = 10$ degrees of freedom.

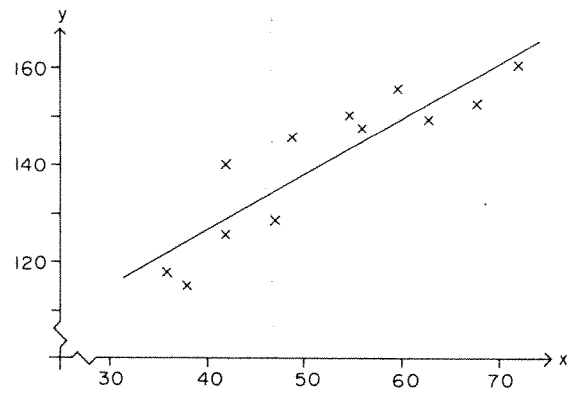

Figure 13.5.1. Scatterplot of blood pressure (y) versus age (x).

EXAMPLE 13.5.2. In Examples 13.3.2 and 13.4.2 we considered data from endurance tests of plastic gears at 21°C and 30°C. These data came from an experiment in which $n = 40$ gears were tested at nine different temperatures.

Examination of the 40 lifetimes revealed that the lifetime distribution has a long tail to the right, and that there is more variability in the lifetimes at the lower temperatures. It would not be appropriate to assume that the lifetimes were normally distributed with constant variance. Instead, we analyze log lifetimes as in Examples 13.3.2 and 13.4.2.

The natural logarithms of the observed lifetimes (in millions of cycles) are given in Table 13.5.1, and are plotted against operating temperature in Figure 13.5.2. Note that the amount of scatter in the log lifetimes is about the same at all temperatures, and the dependence of mean log lifetime on temperature is roughly linear.

There are 40 pairs (x_i, y_i), where x_i is the operating temperature and y_i is the log lifetime. Note that there are repeated x-values:

$$x_1 = x_2 = x_3 = x_4 = -16;$$

$$x_5 = x_6 = x_7 = x_8 = 0;$$

and so on. We assume that the y_i's are observed values of independent $N(\mu_i, \sigma^2)$ variates where $\mu_i = \alpha + \beta x_i$.

To find the parameter estimates, we first compute

$$\Sigma x_i = 4(-16) + 4(0) + 4(10) + 8(21) + \cdots = 1096$$

$$\Sigma x_i^2 = 4(-16)^2 + 4(0)^2 + 4(10)^2 + 8(21)^2 + \cdots = 53816$$

$$\Sigma y_i = 24.996 \quad \Sigma y_i^2 = 37.506862 \quad \Sigma x_i y_i = -15.781$$

We can now compute means, corrected sums of squares, and estimates as in the preceding example. The fitted line is

$$y = 1.432 - 0.02946x$$

and the residual sum of squares is 1.24663 with 38 degrees of freedom, giving the variance estimate $s^2 = 0.03281$.

When there are repeated x-values, it is possible to test the goodness of fit of the straight-line model (see Section 14.4). In this case, the test indicates a poor fit. The poor fit can also be seen in Figure 13.5.2, since at 5 of the 9 temperatures, all of the observed log lifetimes lie on the same side of the fitted line. See Section 14.4 for further discussion of this example, and for a possible explanation of the poor fit.

Table 13.5.1. Log Lifetimes of Plastic Gears at Nine Operating Temperatures

Temperature x(°C)	Number tested	y = Natural logarithm of lifetime (in millions of cycles)			
−16	4	1.690	1.779	1.692	1.857
0	4	1.643	1.584	1.585	1.462
10	4	1.153	0.991	1.204	1.029
21	8	0.863	0.698	0.904	0.788
		0.626	0.842	0.693	1.051
30	4	0.364	0.695	0.558	0.359
37	4	0.412	0.425	0.574	0.649
47	4	0.116	0.501	0.296	0.099
57	4	−0.355	−0.269	−0.354	−0.459
67	4	−0.736	−0.343	−0.965	−0.705

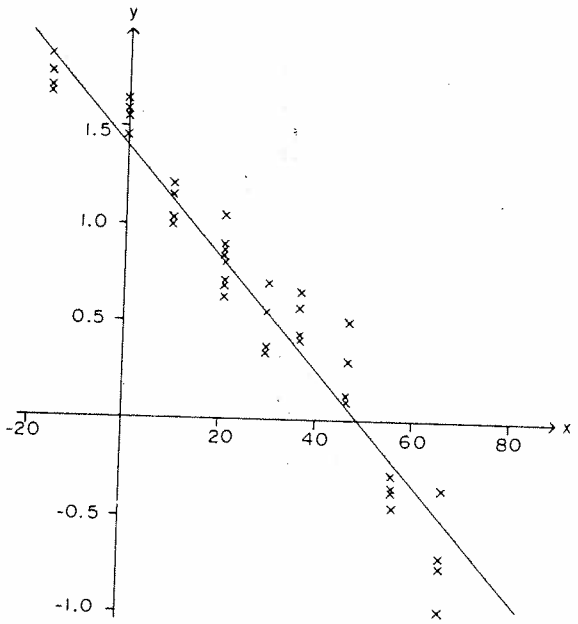

Figure 13.5.2. Scatterplot of log lifetimes (y) versus temperature (x).

Plot the Data!

It is possible to compute $\hat{\alpha}$, $\hat{\beta}$, and s^2 for any set of n pairs (x_i, y_i). Nothing in the arithmetic tells us whether fitting a straight line model is a sensible thing to do. It is important to plot the data and check that the straight-line model gives a reasonable fit. The graph may reveal difficulties with the model, or special features of the data which affect the interpretation. This point is illustrated in the following example, which was given by F.J. Anscombe, Graphs in Statistical Analysis, *The American Statistician* 27 (1973), pages 17–21.

EXAMPLE 13.5.3. Four data sets each consisting of 11 pairs (x_i, y_i) are given in Table 13.5.2. All four sets give approximately the same numerical results

$$\hat{\alpha} = 3, \hat{\beta} = 0.5, s^2 = 1.528.$$

However, as Figure 13.5.3 shows, the appropriate conclusions will be qualitatively different in each case.

Table 13.5.2.

Set 1		Set 2		Set 3		Set 4	
x	y	x	y	x	y	x	y
4	4.26	4	3.10	4	5.39	8	6.58
5	5.68	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.42	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.74	8	6.89
14	9.96	14	8.10	14	8.84	19	12.50

The points of the first data set appear to be scattered randomly about the fitted line $Y = 3 + 0.5x$. The straight line model gives a satisfactory fit to the data, and there are no peculiarities which need to be pointed out.

For the second data set, the dependence of Y on x is clearly not linear. The straight line model is inappropriate, and instead a quadratic polynomial model

$$\mu_i = \alpha + \beta x_i + \gamma x_i^2 \qquad \text{for } i = 1, 2, \ldots, n$$

could be tried.

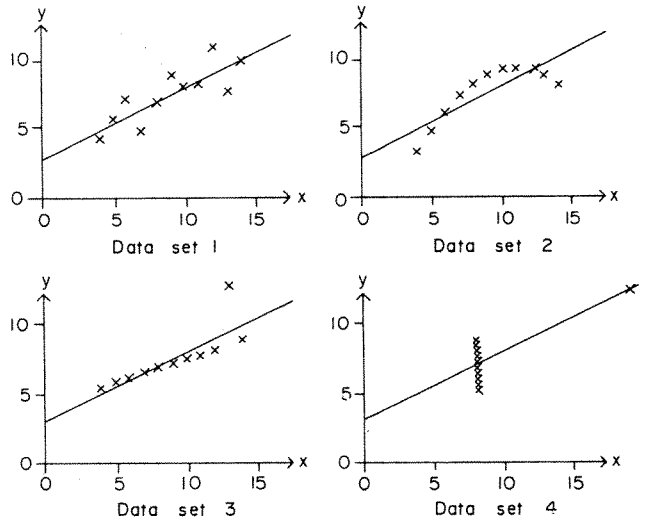

Figure 13.5.3. Scatterplots of the four data sets in Table 13.5.2.

With data set #3, there is an outlying point at $x = 13$. It causes the fitted line to be shifted upwards, so that it does not properly fit the remaining ten points either. If we remove this point and recalculate, the fitted line is

$$Y = 4 + 0.346x$$

which gives a close fit to the remaining ten points and a much smaller variance estimate. Both the outlier and the revised analysis should be reported.

The fourth data set shows good agreement with the straight line model, but the estimate of the slope depends entirely upon a single observation. If this observation were found to be in error and deleted, the slope could not be estimated. Furthermore, without measurements at additional values of x, there is no way of determining whether the actual dependence of y on x is even close to being linear. The fact that the analysis depends so heavily on a single observation should be reported along with the numerical results.

In summary, although all four data sets yield the same numerical results for the straight line model, different conclusions are appropriate in the four cases. Examination of a graph is an indispensable part of the statistical analysis.

Some additional graphical methods for examining the adequacy of the model will be described in Section 14.5.

Problems for Section 13.5

1. Theory suggests that a linear relationship exists between the shearing strength of steel bolts and their diameters. The following table gives the diameter x and strength y for 9 bolts of a particular type.

x	1/8	1/4	3/8	1/2	5/8	3/4	7/8	1	3/2
y	47	72	97	126	165	186	233	257	311

   (a) Fit a straight line to the data, and compute the variance estimate.
   (b) Plot the data and the fitted line. Note that one of the observations is seriously out of line with the others.
   (c) Recalculate the fitted line and variance estimate with the outlying observation omitted, and plot the new line on the graph in (b). Briefly describe the effect of this observation on the analysis.

2.† The following are the breaking strengths of six bolts at each of five different diameters.

Diameter	0.1	0.2	0.3	0.4	0.5
	1.62	1.71	1.86	2.14	2.45
	1.73	1.78	1.86	2.07	2.42
Breaking	1.70	1.79	1.90	2.11	2.33
strength	1.66	1.86	1.95	2.18	2.36
	1.74	1.70	1.96	2.17	2.38
	1.72	1.84	2.00	2.07	2.31

(a) Fit a straight line to the data and compute the variance estimate.
(b) Plot the data and the fitted line. Does the dependence of breaking strength on diameter appear to be linear? How should the model be modified?

3. The analysis in the preceding example assumed that the variance in breaking strength was the same at all five diameters. To check this assumption, compute the sample variance for the six measurements at diameter 0.1. Repeat for each of the other diameters to obtain five sample variances, each with five degrees of freedom. Now carry out a likelihood ratio test of the hypothesis that the variance is the same at all five diameters. (See Problem 13.4.5.)

4. The following table gives x, the water content of snow on April 1, and y, the water yield from April to July (in inches), for the Snake River watershed in Wyoming for 17 years (1919–35).

x	y	x	y	x	y
10.5	23.1	16.7	32.8	18.2	31.8
17.0	32.0	16.3	30.4	10.5	24.0
23.1	39.5	12.4	24.2	24.9	52.5
22.8	37.9	14.1	30.5	12.9	25.1
8.8	12.4	17.4	35.1	14.9	31.5
10.5	21.1	16.1	27.6		

(a) Fit a straight line to these data, and calculate the variance estimate. Plot the data and the fitted line. Can you spot any difficulties?
(b) Suppose that the observations with the smallest and largest x-values are dropped from the analysis. Without redoing the calculations, explain what effects this will have on the estimates of the intercept, slope, and variance.

5. Archeologists use both tree ring dating and carbon dating in estimating the age of artifacts. In one study of Indian ruins, the estimated ages (in years) by tree ring dating (R) and carbon dating (C) were as follows:

R	C	R	C	R	C
710	795	212	222	415	432
717	764	822	765	272	352
350	320	612	543	204	187
323	360	647	642	206	192
500	612	513	533	824	764
620	642	722	724	641	701
832	786	724	745	527	529
669	690	400	409	569	582
917	878	396	456	693	646
423	436	812	652	471	360

Taking R as the dependent variable (y) and C as the independent variable (x), fit a straight line to the data. Repeat with the roles of R and C interchanged. Plot the data and both fitted lines. Why are the two lines different?

Note: It is not clear that either of the above analyses is appropriate, because there is no natural choice for the dependent and independent variables in this example. In fact, both R and C could be considered to be dependent on actual age.

6.† The following measurements of atmospheric pressure (AP) and the boiling point of water (BP) were taken at various altitudes in the Alps and Scotland. Theory suggests that the boiling point of water should change linearly with changes in the (natural) logarithm of the atmospheric pressure.

BP	AP	BP	AP	BP	AP
194.5	20.79	200.9	23.89	209.5	28.49
194.3	20.79	201.1	23.99	208.6	27.76
197.9	22.40	201.4	24.02	210.7	29.04
198.4	22.67	201.3	24.01	211.9	29.88
199.4	23.15	203.6	25.14	212.2	30.06
199.9	23.35	204.6	26.57		

(a) Fit a straight line model $E(\text{BP}) = \alpha + \beta \log(\text{AP})$ to the data and compute the variance estimate. Use the fitted line to estimate the boiling point of water for atmospheric pressures 20, 25, and 30.
(b) How would the results in (a) be affected if one used logarithms to the base 10 rather than natural logarithms?
(c) Plot the data and the fitted line. Can you spot any difficulties? If so, how would you suggest that the analysis should be modified?

13.6. The Straight Line Model (Continued)

In Section 13.5 we derived estimates of the parameters α, β, and σ^2 in the straight line model. In this section we apply the methods described in Section 13.2 to obtain significance tests and confidence intervals.

We are considering n observed pairs (x_i, y_i) for $i = 1, 2, \ldots, n$. The x_i's are treated as known constants, and the y_i's are modelled as observed values of independent $N(\mu_i, \sigma^2)$ variates, where $\mu_i = \alpha + \beta x_i$.

Throughout this section, we use s^2 to denote the appropriate variance estimate based on the straight line model, as defined in (13.5.5). This estimate has $n - 2$ degrees of freedom. Confidence intervals for σ can be obtained by either of the methods described at the end of Section 13.2.

Inferences about the Slope β

By (13.5.3), the MLE of β is

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\Sigma(x_i - \bar{x})y_i}{S_{xx}} = \Sigma a_i y_i$$

where the a_i's are constants:

$$a_i = (x_i - \bar{x})/S_{xx} \qquad \text{for } i = 1, 2, \ldots, n.$$

The sampling distribution of $\hat{\beta}$ is $N(\beta, \sigma^2 c)$, where

$$c = \Sigma a_i^2 = \Sigma(x_i - \bar{x})^2/S_{xx}^2 = 1/S_{xx}.$$

Inferences about β are based on

$$T \equiv \frac{\hat{\beta} - \beta}{\sqrt{s^2 c}} \sim t_{(n-2)} \tag{13.6.1}$$

where s^2 is the variance estimate for the straight line model.

The 95% confidence interval for β is

$$\beta \in \hat{\beta} \pm t\sqrt{s^2 c} = \hat{\beta} \pm ts\sqrt{1/S_{xx}}$$

where t is the value (from Table B3) such that

$$P\{-t \le t_{(n-2)} \le t\} = 0.95.$$

Note that we will be able to determine β precisely (i.e. the confidence interval will be narrow) if S_{xx} is large. Thus, if we are planning an experiment to obtain information about β, we should select $x_1, x_2, \ldots, x_n$ so that $S_{xx} = \Sigma(x_i - \bar{x})^2$ is large. To maximize the information, we would need to make half of the x_i's as large as possible, and the other half as small as possible. However if we did this, we would be unable to check the assumption that the dependence of $E(Y)$ on x is linear. As a result, one would usually compromise by taking observations over the whole range of x-values, but with more observations at the extremes than in the middle of the range. It would then be possible to check the fit of the model, and also to make fairly precise statements about β.

Inferences about $E(Y)$

Given a particular value for x, the expected value of Y is $\mu = \alpha + \beta x$, with MLE

$$\hat{\mu} = \hat{\alpha} + \hat{\beta}x.$$

Since $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, it follows that

$$\begin{aligned}\hat{\mu} &= \bar{y} + \hat{\beta}(x - \bar{x}) \\ &= \frac{1}{n}\Sigma y_i + (x - \bar{x})\Sigma a_i y_i \\ &= \Sigma\left[\frac{1}{n} + (x - \bar{x})a_i\right] y_i\end{aligned}$$

where the a_i's are as defined above. Hence the sampling distribution of $\hat{\mu}$ is $N(\mu, \sigma^2 c')$, where

$$\begin{aligned}c' &= \Sigma\left[\frac{1}{n} + (x - \bar{x})a_i\right]^2 \\ &= \frac{1}{n} + \frac{2(x - \bar{x})}{n}\Sigma a_i + (x - \bar{x})^2 \Sigma a_i^2.\end{aligned}$$

But since $\bar{x} = \frac{1}{n}\Sigma x_i$, it follows that

$$\Sigma(x_i - \bar{x}) = \Sigma x_i - n\bar{x} = 0,$$

and hence that $\Sigma a_i = 0$. Since $\Sigma a_i^2 = 1/S_{xx}$, we have

$$c' = \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}.$$

Inferences about μ are now based on

$$T' \equiv \frac{\hat{\mu} - \mu}{\sqrt{s^2 c'}} \sim t_{(n-2)}. \tag{13.6.2}$$

The 95% confidence interval for $\mu = \alpha + \beta x$ is

$$\mu \in \hat{\mu} \pm t\sqrt{s^2 c'} = \hat{\mu} \pm ts\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right]^{1/2}$$

where t is the value such that $P\{-t \le t_{(n-2)} \le t\} = 0.95$. This interval is narrowest when $x = \bar{x}$, and its width increases as $|x - \bar{x}|$ increases. We can estimate $\alpha + \beta x$ the most precisely when x is close to $\bar{x}$, the mean of the x-values used in fitting the line.

Inferences for the Intercept α

The intercept α is the expected value of Y when $x = 0$. Upon substituting $x = 0$ in the preceding two paragraphs, we find that $\hat{\alpha}$ has sampling distribution $N(\alpha, \sigma^2 c'')$, where

$$c'' = \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}.$$

Inferences about α are thus based on

$$T'' \equiv \frac{\hat{\alpha} - \alpha}{\sqrt{s^2 c''}} \sim t_{(n-2)}.$$

EXAMPLE 13.6.1. In Example 13.5.1, a straight line model was fitted to $n = 12$ blood pressure measurements for women of various ages. The following

results were obtained:

$$\hat{\alpha} = 80.78 \qquad \hat{\beta} = 1.138 \qquad s^2 = 49.247 \text{ (10 d.f.)}$$
$$\bar{x} = 52.33 \qquad S_{xx} = 1550.67.$$

We shall use these results to obtain 99% confidence intervals for β, $\alpha + 50\beta$, and $\alpha + 70\beta$.

Inferences about β are based on (13.6.1) with $c = 1/S_{xx}$. Table B3 gives

$$P\{-3.169 \le t_{(10)} \le 3.169\} = 0.99,$$

and hence the 99% confidence interval for β is

$$\beta \in \hat{\beta} \pm 3.169\sqrt{s^2 c} = 1.138 \pm 0.565.$$

This is also a 1% significance interval. If we test H: $\beta = \beta_0$ for any parameter value β_0 outside this interval, we will obtain SL < 0.01. In particular, the hypothesis $\beta = 0$ is very strongly contradicted by the data.

According to the model, the mean blood pressure of women aged 50 is $\mu = \alpha + 50\beta$. Inferences about μ are based on (13.6.2) with

$$c' = \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} = \frac{1}{12} + \frac{(50 - 52.33)^2}{1550.67} = 0.0868.$$

The 99% confidence interval for $\alpha + 50\beta$ is

$$\hat{\alpha} + 50\hat{\beta} \pm 3.169\sqrt{s^2 c'} = 137.68 \pm 6.55.$$

Similarly, the mean blood pressure of women aged 70 is $\mu' = \alpha + 70\beta$. Inferences about μ' are based on (13.6.2) with

$$c' = \frac{1}{12} + \frac{(70 - 52.33)^2}{1550.67} = 0.2847.$$

The 99% confidence interval for $\alpha + 70\beta$ is

$$\hat{\alpha} + 70\hat{\beta} \pm 3.169\sqrt{s^2 c'} = 160.44 \pm 11.87.$$

As we noted earlier, the width of the confidence interval for $\alpha + \beta x$ increases as $|x - \bar{x}|$ increases. We can estimate $\alpha + \beta x$ with the greatest precision when x is close to $\bar{x} = 52.33$, which is the average of the x-values used in fitting the model.

These confidence intervals are computed under the assumption that the straight line model is correct, and they may be quite misleading if the actual dependence of $E(Y)$ on x is nonlinear. Even if the model seems to fit the data very well, there is no guarantee that it will apply over a wider range of x-values. It is always dangerous to extrapolate beyond the range of x-values in the sample. For instance, in this example it would be unwise to use the straight line model to estimate the mean blood pressure of women aged 30. We have no observations near $x = 30$, and we cannot be sure that the same straight line model will hold in this region.

Problems for Section 13.6

1. In Problem 13.5.1(a), test the hypothesis that the line goes through the origin, and obtain a 95% confidence interval for the mean strength of bolts with diameter $x = 0.9$. Repeat using the revised analysis of Problem 13.5.1(c), and compare the results.

2.† Expected energy use in heating a house decreases as the amount of insulation in the attic increases. To a first approximation, the expected energy per degree day is a linear function of x, a rating of the attic insulation. Small values of x indicate a well-insulated attic, and large values of x indicate poor insulation. The following table gives the observed fuel consumption y and insulation rating x for 8 houses of similar construction:

Insulation rating x	1.4	1.1	0.9	0.7	0.5	0.4	0.3	0.2
Energy use y	1.56	1.30	1.34	1.12	1.08	1.09	1.05	1.21

   (a) Fit a straight line to the data, and calculate the variance estimate. Plot the data and the fitted line. Does the straight line model give a reasonable fit to the data?
   (b) Find a 95% confidence interval for the slope of the line.
   (c) Find a 95% confidence interval for the mean energy use of such houses with insulation rating $x = 0.4$.

3. A study was carried out to investigate evaporation loss from packages of food in storage. Nine similar packages were stored for various periods of time, and the weight losses were recorded.

Days in storage	2	7	9	12	18	23	30	35	40
Weight loss	15	25	40	65	105	105	136	175	180

   (a) Fit a straight line to the data and calculate the variance estimate.
   (b) Plot the data and the fitted line, and comment on the adequacy of the straight line model.
   (c) Find a 95% confidence interval for the mean weight loss in packages stored for two weeks.

4. Suppose that both of the following models are fitted by least squares to n observed points (x_i, y_i).

$$\text{Model 1:} \quad \mu_i = \alpha + \beta x_i;$$
$$\text{Model 2:} \quad \mu_i = \gamma + \delta(x_i - \bar{x}).$$

   Show that $\hat{\gamma} = \bar{y}$, $\hat{\delta} = \hat{\beta}$, and that both models give the same fitted values $\hat{\mu}_i$ and residuals $\hat{\varepsilon}_i$.

5. Suppose that $x_i = 0$ for n_1 observations, and $x_i = 1$ for the other n_2 observations $(n_1 + n_2 = n)$. Show that, in this case, (13.5.2) and (13.5.3) give the appropriate estimates $\hat{\alpha}$, $\hat{\beta}$ for the two-sample model (13.4.2).

6. *Fitting a Straight Line Through the Origin.* Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent $N(\mu_i, \sigma^2)$. Consider the model $\mu_i = \beta x_i$ where $x_1, x_2, \ldots, x_n$ are known constants and β is an unknown parameter.

(a) Show that $\hat{\beta} = \Sigma x_i y_i / \Sigma x_i^2$, and that the sampling distribution of $\hat{\beta}$ is $N(\beta, \sigma^2/\Sigma x_i^2)$.
(b) Show that the variance estimate is

$$s^2 = \frac{1}{n-1}[\Sigma y_i^2 - \hat{\beta}\Sigma x_i y_i].$$

7.† A new procedure is being investigated for measuring calcium content. The following table gives the actual calcium content x for each of ten samples, and the measurement y given by the new procedure.

x	4.0	8.0	12.5	16.0	20.0	25.0	31.0	36.0	40.0	40.0
y	3.7	7.8	12.1	15.6	19.8	24.5	31.1	35.5	39.4	39.5

(a) Fit a straight line to the data and calculate the residual sum of squares. Plot the data and the fitted line.
(b) Test the hypothesis that the slope of the line is 1.
(c) Test the hypothesis that the line passes through the origin.
(d) Fit a straight line through the origin to these data, and test the hypothesis that the slope is 1. Why is the result different from that obtained in (b)?

13.7. Analysis of Paired Measurements

In Section 12.8 we considered an example in which drugs A and B were administered to the same n subjects in order to see which was more likely to produce nausea. Because the drugs were given to the same subjects, we could not assume that observations for drug A were independent of observations for drug B.

In Section 12.8 we were concerned only with the absence or presence of an effect. However, similar problems can arise when we are measuring the size of an effect.

Suppose that an experiment yields n pairs of measurements (A_i, B_i) for $i = 1, 2, \ldots, n$. Because of the way the data were collected, we expect A_i and B_i to be more alike than A_i and B_j for $i \neq j$. For instance, B_i and A_i might be measurements of blood pressure on the same subject before and after the administration of a drug, or measurements of gasoline mileage for the same car before and after a tuneup.

In such situations, it would be incorrect to treat $A_1, A_2, \ldots, A_n$ and $B_1, B_2, \ldots, B_n$ as two independent samples. The analysis must take the pairing of the data into account. We shall describe two ways in which this can be done.

Analysis of Differences

The simplest approach is to replace each pair (A_i, B_i) by a suitable summary statistic such as the difference

$$y_i = A_i - B_i.$$

Only the y_i's are used in the analysis, and it is usually reasonable to assume that they are independent. No assumptions are required about the distributions of the individual measurements A_i and B_i.

The details of the analysis will depend upon what is assumed about the distributions of the y_i's. If the y_i's are assumed to be observed values of $N(\mu_i, \sigma^2)$ variates, then the methods developed in this chapter will apply. We model the μ_i's as linear functions of unknown parameters $\beta_1, \beta_2, \ldots, \beta_q$ as called for by the situation, and choose parameter estimates to minimize

$$S = \Sigma(y_i - \mu_i)^2. \tag{13.7.1}$$

There will be $n - q$ degrees of freedom for variance estimation.

EXAMPLE 13.7.1. The following table gives the average number of manhours per month lost due to accidents in eight factories of similar size over a period of one year before and after the introduction of an industrial safety program.

Factory i	1	2	3	4	5	6	7	8
After A_i	28.7	62.2	28.9	0.0	93.5	49.6	86.3	40.2
Before B_i	48.5	79.2	25.3	19.7	130.9	57.6	88.8	62.1
Difference y_i	−19.8	−17.0	3.6	−19.7	−37.4	−8.0	−2.5	−21.9

There is a natural pairing of the data by factory. Factories with the best safety records before the safety program tend to have the best records after the safety program as well. The analysis of the data must take this pairing into account. One way of doing this is to analyze only the differences $y_i = A_i - B_i$.

We assume that the y_i's are observed values of independent $N(\mu_i, \sigma^2)$ variates, and that $\mu_1 = \mu_2 = \cdots = \mu_n = \alpha$, say. The $n = 8$ differences are thus analyzed as a one-sample problem (see Section 13.3). We have $\hat{\alpha} = \bar{y} = -15.34$, and the variance estimate is

$$s^2 = \frac{1}{n-1}\Sigma(y_i - \bar{y})^2 = 164.11$$

with $n - 1 = 7$ degrees of freedom.

The sampling distribution of $\hat{\alpha}$ is $N(\alpha, \sigma^2 c)$ where $c = \frac{1}{n}$, and inferences concerning α are based on

$$T \equiv \frac{\hat{\alpha} - \alpha}{\sqrt{s^2 c}} \sim t_{(n-1)}.$$

Since $P\{-2.365 \leq t_{(7)} \leq 2.365\}$, the 95% confidence interval for α is

$$\alpha \in \hat{\alpha} \pm 2.365\sqrt{s^2/n} = -15.34 \pm 10.71.$$

The 95% confidence interval for the mean decrease in lost manhours per month is (4.63, 26.05). Since zero does not belong to this interval, there is evidence of a real decrease in the mean number of lost manhours.

Incorrect Analysis

Suppose that we ignore the pairing of the data, and analyze the original $2n$ measurements as a two-sample problem (see Section 13.4). We find that

$$\bar{A} = 48.68;\ \bar{B} = 64.01;\ s_A^2 = 977;\ s_B^2 = 1295$$

and the pooled variance estimate is

$$s^2 = \frac{7 \times 977 + 7 \times 1295}{7+7} = 1136$$

with 14 degrees of freedom.

Let $\alpha = \mu_A - \mu_B$ be the decrease in the mean number of lost manhours. The MLE of α is $\hat{\alpha} = \bar{A} - \bar{B}$, with sampling distribution $N(\alpha, \sigma^2 c)$, where $c = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$. Inferences about α are now based on

$$T' \equiv \frac{\hat{\alpha} - \alpha}{\sqrt{s^2 c}} \sim t_{(14)}$$

where s^2 is the pooled variance estimate. The 95% confidence interval for α based on the two-sample model is

$$\alpha \in \hat{\alpha} \pm 2.145\sqrt{s^2/4} = -15.34 \pm 36.18.$$

The interval is now much wider than before, and the decrease in mean number of lost manhours is no longer statistically significant. The variance estimate is inflated by the large differences among factories, and so the effect of the safety program does not show up.

Models with Block Effects

Instead of analyzing differences, we could retain all $2n$ measurements (A_i, B_i) in the analysis and build a model which allows for the fact that A_i and B_i are likely to be similar. For instance, we could assume that the $2n$ measurements are independent and normally distributed with the same variance, and take

$$E\{B_i\} = \pi_i;\ E\{A_i\} = \pi_i + \mu_i \qquad \text{for } i = 1, 2, \ldots, n.$$

The parameters $\pi_1, \pi_2, \ldots, \pi_n$ are the pair or block effects, and $\mu_1, \mu_2, \ldots, \mu_n$ represent the effects of the treatment in the n pairs. Measurements A_i, B_i from the same pair share the parameter π_i, whereas measurements A_i, B_j from different pairs have different parameters π_i, π_j.

The μ_i's can now be modelled as linear functions of unknown parameters $\beta_1, \beta_2, \ldots, \beta_q$ as appropriate for the situation. Parameter estimates are found by minimizing

$$S = \Sigma(B_i - \pi_i)^2 + \Sigma(A_i - \pi_i - \mu_i)^2.$$

There are $n + q$ parameters to estimate from the data, and so there will be $2n - (n+q) = n - q$ degrees of freedom for variance estimation.

This analysis gives essentially the same results as the analysis of differences. To see this, we note that the solution of $\partial S/\partial \pi_i = 0$ is

$$\pi_i = \tfrac{1}{2}(A_i + B_i - \mu_i).$$

Upon substituting for π_i and simplifying, we find that the minimum of the error sum of squares over the π_i's is

$$S_{\min} = \tfrac{1}{2}\Sigma(y_i - \mu_i)^2$$

where $y_i = A_i - B_i$. This is one-half of (13.7.1). As a result, the $\hat{\beta}_j$'s and $\hat{\mu}_i$'s will be the same as in the analysis of differences. The variance estimates will differ by a factor of 2 because now $\sigma^2 = \text{var}(A_i) = \text{var}(B_i)$ is the variance of a single measurement, whereas previously we used σ^2 to denote the variance of the difference $A_i - B_i$.

If A_i and B_i are independent normal, then $A_i - B_i$ has a normal distribution as we assumed in the analysis of differences. However it is possible to have $A_i - B_i$ normally distributed without having either the A_i's or the B_i's normal. The revised model with block effects involves more stringent assumptions than are necessary for the analysis of differences.

An advantage of models with block effects is that they can be constructed for situations in which the measurements occur in blocks of three or more rather than pairs, and it is not necessary that all blocks be of the same size. The analysis of differences does not easily generalize to these situations.

Design of Experiments

Often it is advantageous to design the experiment so that it will produce paired measurements. For instance, suppose that treatments A and B are to be compared using $2n$ subjects, n for each treatment. Before treatments are assigned, subjects are grouped or blocked into n pairs so that the members of a pair are as similar as possible with respect to potentially important factors like age, weight, prognosis, etc. The treatments are then assigned randomly within each pair, so that for each pair we obtain two measurements A_i, B_i. Using the analysis of differences, or a model with block effects, one can eliminate differences between pairs from the analysis and obtain a more precise comparison of the treatments.

Problems for Section 13.7

1. The following table gives the results of a series of measurements of the corrosion of coated and uncoated underground pipes:

Soil type:	1	2	3	4	5	6
Coated:	15.6	21.0	22.6	56.8	13.2	20.9
Uncoated:	10.9	46.7	25.7	69.7	36.7	20.4

Soil type:	7	8	9	10	11	12
Coated:	8.6	31.2	25.4	8.5	11.2	35.8
Uncoated:	29.4	10.2	71.6	42.8	23.9	49.2

Obtain a 99% confidence interval for the mean difference in the amounts of corrosion for the two types of pipe.

2. Two analysts carried out simultaneous measurements of the percentage of ammonia in a plant gas on nine successive days to find the extent of the bias, if any, between their results. Their measurements were:

Day	1	2	3	4	5	6	7	8	9
Analyst A	4	37	35	43	34	36	48	33	33
Analyst B	18	37	38	36	47	48	57	28	42

Obtain a 95% confidence interval for the mean difference in their measurements. On what assumptions does your analysis depend?

3.† Six automobiles of different models were used to compare two brands of tires. Each car was fitted with tires of brand A and driven over a difficult course until one of its tires could no longer be used. Tires of brand B were then fitted to the same cars, and the procedure was repeated. The following are the observed mileages to tire failure in thousands of miles:

Car	1	2	3	4	5	6
Brand A	18	23	16	27	19	17
Brand B	15	22	16	21	15	16

(a) Test whether these data are consistent with the hypothesis that the mean lifetimes for the two brands are equal.
(b) What factor, other than difference in tire quality, might account for the lower mileage achieved with brand B? Suggest an improvement in the design of the experiment which would have helped to eliminate this source of bias.

4. Two methods of treating sewage were compared. Each day for eight days, two similar batches of sewage were selected. One batch was randomly chosen to receive treatment A, and the other received treatment B. The following table shows the coliform density per ml for the sixteen batches after treatment.

Day	1	2	3	4	5	6	7	8
A	16.44	22.00	18.17	20.09	11.02	20.09	24.53	13.46
B	24.53	22.20	29.96	33.12	14.88	18.16	33.12	16.44

(a) Assuming that differences in coliform density are normally distributed, test the hypothesis that the treatments are equally effective.
(b) A more reasonable assumption in this case is that the logarithmic differences $\log A_i - \log B_i$ are independent $N(\mu, \sigma^2)$. Repeat the test in (a) under this assumption.

5. A study was carried out to investigate the effect of trap color on the catch of whiteflies. Two similar traps, one yellow and one green, were hung side by side on each of 8 plants in a greenhouse. The following table shows the weight of whiteflies caught in each trap.

Plant	1	2	3	4	5	6	7	8
Yellow trap	20.5	42.7	19.4	100.7	23.9	45.4	99.1	125.9
Green trap	20.0	38.5	15.5	103.6	18.0	47.9	96.4	126.0

(a) Set up a normal model appropriate for examining the difference in effectiveness of the two trap colors.
(b) Test the hypothesis that yellow and green traps are equally effective in catching whiteflies, and state your conclusions carefully.
(c) Discuss briefly the advantages of conducting the study in the manner described rather than by hanging the 16 traps on 16 different plants.

6. Twenty pigs were grouped into ten pairs in such a way that the two pigs in a pair had nearly equal weight. One pig was randomly chosen from each pair to receive diet X, and the other received diet Y. The following are the observed weight gains per day:

Pair	1	2	3	4	5	6	7	8	9	10
Diet X	21	21	19	16	26	19	18	29	22	19
Diet Y	30	25	25	16	29	18	18	19	24	22

Give a 95% confidence interval for the difference in mean weight gain under the two diets, and state the assumptions on which your analysis is based.

7. An experiment was carried out to determine how the defect rate y in a highway surface depends on the amount x of asphalt cement used in the paving material. Seven samples with known asphalt content were prepared. Each sample was split in two, and two separate tests were made to determine the defect rate.

Asphalt content	50	75	100	125	200	250	275
Defect rate	195	172	164	175	145	115	108
	197	175	163	177	147	115	109

(a) Using all 14 observations (x, y), fit a straight line to the data. Plot the data and the fitted line. Obtain a 95% confidence interval for the mean defect rate when $x = 100$ and show it on the graph.
(b) The graph in (a) suggests that it is not appropriate to model the two measurements on the same sample as independent. Instead, it is better to replace the two measurements by their average. Redo the analysis in (a) using the seven observed pairs $(x, \bar{y}_x)$, and compare the results.

8.† A new technique for determining the fraction x of a given gas in a mixture of gases was investigated. Eleven gas mixtures with known x were prepared, and each of them was divided into three portions. For each portion, the quantity y of the gas which dissolved in a liquid was recorded.

x = content	y = amount dissolving			x = content	y = amount dissolving		
0.080	2.67	2.68	2.75	0.131	4.46	4.40	4.43
0.082	2.73	2.69	2.62	0.139	4.78	4.80	4.86
0.091	2.88	3.02	3.04	0.164	5.77	5.85	5.82
0.095	3.17	3.28	3.18	0.189	6.56	6.65	6.49
0.096	3.27	3.28	3.08	0.231	7.88	7.97	7.76
0.106	3.51	3.68	3.58				

(a) Using all 33 observed pairs (x, y), fit a straight line model to the data. Find a 95% confidence interval for the expected amount dissolving in mixtures with $x = 0.1$.

(b) The analysis in (a) assumes that the 3 measurements taken at each value of x are independent replicates. This is a questionable assumption because these measurements were obtained by dividing one gas mixture into three portions rather than by preparing three different mixtures with the same x. One way around this difficulty is to replace the three repeat observations at each x by their average $\bar{y}_x$. Repeat (a) using the 11 observed pairs $(x, \bar{y}_x)$, and compare the results.

Review Problems for Chapter 13

1. Two experiments were carried out to determine μ, the mean increase in blood pressure due to a certain drug. Six different subjects were used, three in each experiment, and the following increases were observed:

Experiment 1:	4.5	5.6	4.9
Experiment 2:	−1.2	9.8	21.4

Indicate, with reasons, which experiment produces stronger evidence that the drug does have an effect on blood pressures. Which experiment points to the greater effect?

2.† Fourteen men were used in an experiment to determine which of two drugs produces a greater increase in blood pressure. Drug 1 was given to seven of the men chosen at random, and drug 2 was given to the remaining seven. The observed increases in blood pressure are:

Drug 1:	0.7	−0.2	3.4	3.7	0.8	0.0	2.0
Drug 2:	1.9	1.1	4.4	5.5	1.6	4.6	3.4

(a) Are these data consistent with the hypothesis of equal variances in blood pressure for the two drugs?

(b) Assuming the variances to be equal, obtain a 95% confidence interval for the difference in mean blood pressure increase $\mu_2 - \mu_1$, and for the common variance σ^2.

(c) It is possible that the increase in blood pressure with both drugs may depend upon the initial blood pressure of the subject. How should the design of the experiment and the analysis be modified to allow for this possibility?

3. The following are yields (in pounds) of 16 tomato plants grown on 8 separate uniform plots of land. One plant in each plot was treated with fertilizer A and the other with fertilizer B.

Plot	1	2	3	4	5	6	7	8
Fertilizer A	4.0	5.7	4.0	6.9	5.5	4.6	6.5	8.4
Fertilizer B	4.8	5.5	4.4	4.8	5.9	4.2	4.4	6.3

Test the hypothesis that the two fertilizers are equally effective, and state the assumptions upon which the test is based.

4. A study was carried out to investigate the dependence of fuel oil consumption on the mean atmospheric temperature. The following are the results observed on ten winter days.

Temperature	−3	−2	−10	+1	−5	−6	−15	−4	−9	−2
Consumption	150	141	238	132	186	168	218	163	210	169

(a) Fit a straight line to the data and calculate the variance estimate. Plot the fitted line and the data, and comment on any difficulties.

(b) Obtain 90% confidence intervals for the intercept, and for the mean fuel consumption on days when the mean temperature is −5.

5. An experiment was performed to compare two different methods of measuring the phosphate content of material. Ten samples were chosen so that the material within a sample was relatively homogeneous. Each sample was then divided in half, one half being analysed by method A and the other half by method B.

Sample	1	2	3	4	5	6	7	8	9	10
Method A	55.6	62.4	48.9	45.5	75.4	89.6	38.4	96.8	92.5	98.7
Method B	58.4	66.3	51.2	46.1	74.3	92.5	40.2	97.3	94.8	99.0

Find a 95% confidence interval for the mean difference in phosphate content as measured by the two methods, and state the assumptions upon which your analysis depends.

6.† In a progeny trial, the clean fleece weights of 9 ewe lambs from each of four sires were as follows:

Sire 1:	2.74	3.50	3.22	2.98	2.97	3.47	3.47	3.68	4.22
Sire 2:	3.88	3.36	4.29	4.08	3.90	4.71	4.25	3.41	3.84
Sire 3:	3.28	3.92	3.66	3.47	2.94	3.26	3.57	2.62	3.76
Sire 4:	3.52	3.54	4.13	3.29	3.26	3.04	3.77	2.88	2.90

(a) Test the hypothesis that the variance in fleece weight is the same for all four sires.

(b) Assuming the variances to be equal, obtain a 95% confidence interval for the common variance σ^2.

CHAPTER 14

Normal Linear Models

In Chapter 13 we considered some simple models for normally distributed measurements. The basic assumptions for these models were discussed in Section 13.1, and some statistical methods were described in Section 13.2.

All of the models considered in Chapter 13 are special cases of the normal linear model, which is the subject of this chapter. Section 14.1 describes matrix notation for linear models and gives several examples. Section 2 considers the estimation of parameters in linear models, and likelihood ratio tests are derived in Section 3. Section 4 gives some further discussion of the statistical methods described in Section 13.2. Section 5 describes some graphical procedures for checking the adequacy of the model. The distributions of the residual sum of squares and the additional sum of squares due to a linear hypothesis are derived in Section 6.

14.1. Matrix Notation

As in Chapter 13, we consider n measurements $y_1, y_2, \ldots, y_n$ of the same quantity taken under various different conditions. We assume that the y_i's are observed values of independent random variables $Y_1, Y_2, \ldots, Y_n$, where

$$Y_i \sim N(\mu_i, \sigma^2) \qquad \text{for } i = 1, 2, \ldots, n. \tag{14.1.1}$$

See Section 13.1 for discussion of these basic assumptions.

The basic model (14.1.1) involves $n+1$ unknown parameters $\mu_1, \mu_2, \ldots, \mu_n$ and σ, but we have only n observations. Before we can estimate σ, we must reduce the number of unknown parameters. We do this by writing the μ_i's as functions of q unknown parameters $\beta_1, \beta_2, \ldots, \beta_q$ where $q < n$. We then have effectively $n - q$ observations available for estimating σ, and we say that there are $n - q$ degrees of freedom for variance estimation.

The model is called *linear* if each of the μ_i's may be written as a linear function of the unknown parameters $\beta_1, \beta_2, \ldots, \beta_q$. In a linear model, all of the partial derivatives $\partial\mu_i/\partial\beta_j$ are known constants. The one-sample, two-sample, and straight line models are examples of linear models with $q = 1, 2, 2$, respectively.

A linear model can be described by a set of n linear equations:

$$\mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iq}\beta_q \qquad \text{for } 1 \leq i \leq n. \tag{14.1.2}$$

This is also called a *multiple regression model*. The β_j's are unknown parameters, and $x_{i1}, x_{i2}, \ldots, x_{iq}$ are known constants which describe the conditions under which the ith observation is made. The x_{ij}'s may be values of quantitative variables such as temperature or age, or values of $0-1$ indicator variables, or a mixture of these.

The n linear equations (14.1.2) can be represented by a single matrix equation

$$\mu = X\beta \tag{14.1.3}$$

where μ is $n \times 1$, β is $q \times 1$, and X is $n \times q$:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_n \end{bmatrix}; \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_q \end{bmatrix}; \qquad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix}$$

X has one row for each observation y_i, and one column for each unknown parameter β_j. To obtain X, we just write out the n equations (14.1.2) one below the other and detach the coefficients of the β_j's.

We shall assume that the q columns of X are linearly independent. If they were not, it would be possible to rewrite the n equations using only $q-1$ of the unknown parameters $\beta_1, \beta_2, \ldots, \beta_q$.

The remainder of this section describes a few of the many situations covered by linear models. In particular, all of the models considered in Chapter 13 are linear models. Results derived for linear models in the following sections are applicable to all of these situations and many others as well.

Straight Line Model (Section 13.5)

The n equations defining the straight line model are

$$\mu_1 = \beta_1 + \beta_2 x_1 = 1 \cdot \beta_1 + x_1 \cdot \beta_2$$
$$\mu_2 = \beta_1 + \beta_2 x_2 = 1 \cdot \beta_1 + x_2 \cdot \beta_2$$
$$\cdots\cdots\cdots\cdots$$
$$\mu_n = \beta_1 + \beta_2 x_n = 1 \cdot \beta_1 + x_n \cdot \beta_2$$

These can be written in the form $\mu = X\beta$ where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_n \end{bmatrix}; \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}; \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix}.$$

Polynomial Model

As a generalization of the straight line model, one might consider a second-degree polynomial model

$$\mu_1 = \beta_1 + x_1\beta_2 + x_1^2\beta_3$$
$$\mu_2 = \beta_1 + x_2\beta_2 + x_2^2\beta_3$$
$$\cdots\cdots\cdots\cdots$$
$$\mu_n = \beta_1 + x_n\beta_2 + x_n^2\beta_3.$$

Here we have $\mu = X\beta$ where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}; \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}; \qquad X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdots & \cdots & \cdots \\ 1 & x_n & x_n^2 \end{bmatrix}.$$

Similarly, cubic and higher degree polynomials can be written in the form $\mu = X\beta$ for a suitable choice of X, and are examples of linear models. The components of X may be any known constants, such as 0, 1, x_i, x_i^2, $\log x_i$, $\sin x_i$, and so on. The model is still a linear model if the μ_i's are *linear functions of the unknown parameters* $\beta_1, \beta_2, \ldots, \beta_q$.

One-Sample Problem (Section 13.3)

In the one-sample problem, we assume that the n means $\mu_1, \mu_2, \ldots, \mu_n$ are all equal to the same unknown value β_1, say. Thus the n equations are

$$\mu_1 = 1 \cdot \beta_1; \qquad \mu_2 = 1 \cdot \beta_1; \qquad \ldots; \qquad \mu_n = 1 \cdot \beta_1,$$

and we have $\mu = X\beta$ where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}; \qquad \beta = [\beta_1]; \qquad X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

In this case X is an $n \times 1$ matrix whose components are all equal to 1.

Two-Sample Problem (Section 13.4)

Here we assume that m of the means are equal to β_1, say, and the other $n - m$ are equal to β_2:

$$\mu_1 = \mu_2 = \cdots = \mu_m = \beta_1,$$
$$\mu_{m+1} = \mu_{m+2} = \cdots = \mu_n = \beta_2.$$

We can write this in the form $\mu = X\beta$ where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_n \end{bmatrix}; \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}; \qquad X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \cdots & \cdots \\ 1 & 0 \\ \hline 0 & 1 \\ 0 & 1 \\ \cdots & \cdots \\ 0 & 1 \end{bmatrix} \begin{matrix} \} \, m \times 2 \\ \\ \\ \\ \\ \} \, (n-m) \times 2 \\ \\ \\ \end{matrix}$$

Weighing Experiment (see Example 10.1.1)

Suppose that three objects with unknown weights β_1, β_2, and β_3 are weighed on a set of scales in all possible combinations, giving 7 independent measurements $Y_1, Y_2, \ldots, Y_7$. We assume that the Y_i's are independent $N(\mu_i, \sigma^2)$, where

$$\mu_1 = \beta_1; \qquad \mu_2 = \beta_2; \qquad \mu_3 = \beta_3;$$
$$\mu_4 = \beta_1 + \beta_2; \qquad \mu_5 = \beta_1 + \beta_3; \qquad \mu_6 = \beta_2 + \beta_3;$$
$$\mu_7 = \beta_1 + \beta_2 + \beta_3.$$

This model has the form $\mu = X\beta$ where X is a 7×3 matrix with 0, 1 entries. Its transpose is

$$X^t = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Parallel Line Model

Suppose that for the first m observations we wish to assume a straight line model $\mu_i = \beta_1 + \beta_3 x_i$, and that for the remaining $n - m$ observations we wish to assume another straight line model $\mu_i = \beta_2 + \beta_3 x_i$ which has the same slope but a different intercept. This model can be written $\mu = X\beta$ where the transpose of X is

$$X^t = \begin{bmatrix} 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_m & x_{m+1} & x_{m+2} & \dots & x_n \end{bmatrix}.$$

Problems for Section 14.1

1.† Consider six measurements $Y_1, Y_2, \dots, Y_6$ with expected values $\mu_1, \mu_2, \dots, \mu_6$. Four possible linear models are described below. In each case define a matrix X with linearly independent columns and a parameter vector β such that the model can be written in the form $\mu = X\beta$.

(a) $\mu_1 = \mu_2 = \mu_3$ and $\mu_4 = \mu_5 = \mu_6$
(b) $\mu_i = \beta_1 + i\beta_2$ for $i \le 3$ and $\mu_i = \beta_1 + i\beta_3$ for $i \ge 4$
(c) $\mu_1 = \mu_2$ and $\mu_3 + \mu_4 = \mu_5 + \mu_6$
(d) $\mu_1 = \mu_2 = \mu_3$

2. Consider n measurements $Y_1, Y_2, \dots, Y_n$ with expected values $\mu_1, \mu_2, \dots, \mu_n$. A straight line model $\mu_i = \beta_1 + \beta_2 x_i$ is to be assumed for the first m measurements, and a different straight line model $\mu_i = \beta_3 + \beta_4 x_i$ is to be assumed for the remaining $n - m$ measurements.

(a) Define a matrix X such that the model can be written in the form $\mu = X\beta$.
(b) Show that, by adding two columns of X in (a), one can obtain X for two straight lines with equal slopes or with equal intercepts.

3. A standard treatment A and two new treatments B, C are to be compared in an experiment using 3 mice from each of four different litters. Twelve measurements Y_i are to be taken according to the following scheme:

Measurement No.	1	2	3	4	5	6	7	8	9	10	11	12
Litter No.	1	2	3	4	1	2	3	4	1	2	3	4
Treatment	A	A	A	A	B	B	B	B	C	C	C	C

It is assumed that

$$E\{Y_i\} = \alpha_i; \qquad E\{Y_{i+4}\} = \alpha_i + \gamma_1; \qquad E\{Y_{i+8}\} = \alpha_i + \gamma_2$$

for $i = 1, 2, 3, 4$. Here γ_1 and γ_2 represent the effects of treatments B and C relative to the standard treatment A. Define a matrix X and parameter vector β such that the model can be written in the form $\mu = X\beta$.

4. A standard treatment A and three new treatments B, C, D are to be compared in an experiment using 3 mice from each of four litters. Twelve measurements Y_i are to be taken according to the following scheme.

Measurement No.	1	2	3	4	5	6	7	8	9	10	11	12
Litter No.	1	2	3	4	1	2	3	4	1	2	3	4
Treatment	A	A	A	B	B	B	C	C	C	D	D	D

Set up a linear model similar to that in the preceding problem, and write it in matrix notation.

14.2. Parameter Estimates

The linear models considered in Chapter 13 were simple enough so that we could obtain an algebraic formula for each of the parameter estimates. With more complicated models, the estimates are usually determined numerically by computer using matrix arithmetic.

As in Section 14.1, we suppose that $Y_1, Y_2, \dots, Y_n$ are independent $N(\mu_i, \sigma^2)$, and consider the linear model

$$\mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iq}\beta_q \qquad \text{for } i = 1, 2, \dots, n.$$

This can be written in matrix notation as $\mu = X\beta$ where X is $n \times q$. We assume that the q columns of X are linearly independent (see Section 14.1).

We noted in Section 13.2 that, under these assumptions, the log-likelihood function is

$$l = -n \log \sigma - \frac{1}{2\sigma^2} S$$

where S is the error sum of squares,

$$S = \Sigma \varepsilon_i^2 = \Sigma (y_i - \mu_i)^2.$$

The MLE's $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_q$ are chosen to minimize S. The fitted values and residuals are defined by

$$\hat{\mu}_i = x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + \cdots + x_{iq}\hat{\beta}_q;$$
$$\hat{\varepsilon}_i = y_i - \hat{\mu}_i.$$

To put this in matrix notation, we let $y = (y_i)$, $\hat{\mu} = (\hat{\mu}_i)$, and $\hat{\varepsilon} = (\hat{\varepsilon}_i)$ be $n \times 1$ vectors, and let $\hat{\beta} = (\hat{\beta}_j)$ be the $q \times 1$ vector of parameter estimates. Then we have

$$\hat{\mu} = X\hat{\beta}; \qquad \hat{\varepsilon} = y - \hat{\mu}. \tag{14.2.1}$$

Derivation of $\hat{\beta}$

The derivative of S with respect to β_j is

$$\frac{\partial S}{\partial \beta_j} = 2\Sigma \varepsilon_i \frac{\partial \varepsilon_i}{\partial \beta_j} = -2\Sigma \varepsilon_i \frac{\partial \mu_i}{\partial \beta_j} = -2\Sigma \varepsilon_i x_{ij}.$$

Thus $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q$ satisfy the q simultaneous equations

$$\sum_i \hat{\varepsilon}_i x_{ij} = 0 \qquad \text{for } j = 1, 2, \ldots, q.$$

These q equations are equivalent to the matrix equation

$$X^t \hat{\varepsilon} = 0. \tag{14.2.2}$$

Substituting $\hat{\varepsilon} = y - X\hat{\beta}$ gives

$$X^t(y - X\hat{\beta}) = 0.$$

It follows that

$$X^t X\hat{\beta} = X^t y. \tag{14.2.3}$$

This is a set of q linear equations in $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q$.

Since X is $n \times q$, the product $X^t X$ is $q \times q$. It can be shown that, since X has linearly independent columns, the product $X^t X$ is nonsingular, and its inverse $(X^t X)^{-1}$ exists. Multiplying (14.2.3) by $(X^t X)^{-1}$ gives

$$\hat{\beta} = (X^t X)^{-1} X^t y = X^L y \tag{14.2.4}$$

where $X^L = (X^t X)^{-1} X^t$.

The matrix X^L is $q \times n$, and it has the property that

$$X^L X = (X^t X)^{-1} X^t X = I_q$$

where I_q is the $q \times q$ identity matrix. Thus X^L is a *left inverse* of X. Note that

$$XX^L = X(X^t X)^{-1} X^t$$

which is $n \times n$ and will *not* equal I_n unless X is a square matrix ($q = n$).

Computation

Calculations for linear models are usually done by computer. The main labor is in finding the $q \times n$ matrix $X^L = (X^t X)^{-1} X^t$. From this we can easily get $\hat{\beta} = X^L y$, $\hat{\mu} = X\hat{\beta}$, and $\hat{\varepsilon} = Y - \hat{\mu}$. Squaring and summing the n components of $\hat{\varepsilon}$ gives the residual sum of squares $\Sigma \hat{\varepsilon}_i^2$. The variance estimate is then

$$s^2 = \frac{1}{n-q} \Sigma \hat{\varepsilon}_i^2$$

with $n - q$ degrees of freedom. Various other quantities useful for assessing the adequacy of the model or for testing hypotheses about the β_j's can also be obtained from X^L (see Sections 14.4 and 14.5).

To analyze one of the examples from Chapter 13 in this way, we must define the appropriate matrix X as indicated in Section 14.1. For instance, in Example 13.5.1 a straight line model $\mu_i = \alpha + \beta x_i$ is to be fitted to $n = 12$ observations. We take y to be the 12×1 vector of observed blood pressure measurements 147, 125, ..., 155, and X to be 12×2 matrix with transpose

$$X^t = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_{12} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 56 & 42 & \ldots & 60 \end{bmatrix}.$$

Then $X^L y = (X^t X)^{-1} X^t y$ is a 2×1 vector whose components are $\hat{\alpha}$ and $\hat{\beta}$.

The computer language APL is particularly convenient for linear models because it has a built-in operator ⌹ for handling the necessary calculations. Having defined an $n \times q$ matrix X and a list of n y-values, one enters ⌹X to obtain X^L, or Y ⌹ X to obtain $\hat{\beta} = X^L y$. Alternatively, statistical software packages such as SAS, SPSS, BMDP, and GLIM may be used for fitting linear models.

EXAMPLE 14.2.1. Data were collected to investigate how the amount of fuel oil required to heat a home depends upon the outdoor air temperature and wind velocity. Table 14.2.1 contains the result for $n = 10$ winter days.

We expect fuel consumption to increase as the wind velocity v increases, and to decrease as the temperature increases. As a first approximation, we assume that these changes are linear, and that the effect of wind velocity is the same at all temperatures. Thus the y_i's are assumed to be observed values of

Table 14.2.1. Fuel Consumption (y), Temperature (t), and Wind Velocity (v) on Each of Ten Winter Days

Day	y	t	v
1	14.96	−3.0	15.3
2	14.10	−1.8	16.4
3	23.76	−10.0	41.2
4	13.20	0.7	9.7
5	18.60	−5.1	19.3
6	16.79	−6.3	11.4
7	21.83	−15.5	5.9
8	16.25	−4.2	24.3
9	20.98	−8.8	14.7
10	16.88	−2.3	16.1

independent $N(\mu_i, \sigma^2)$ variates, where

$$\mu_i = \beta_1 + \beta_2 t_i + \beta_3 v_i \qquad \text{for } i = 1, 2, \ldots, 10.$$

Here β_2 is the effect on mean fuel consumption of a unit increase in temperature assuming that the wind velocity is held fixed, and β_3 is the effect on mean consumption of a unit increase in wind velocity with the temperature fixed. The "general constant term" β_1 represents the mean fuel consumption when $t = v = 0$.

The model can be written as $\mu = X\beta$ where X is the 10×3 matrix shown in Figure 14.2.1. The left inverse

$$X^L = (X^t X)^{-1} X^t$$

was obtained by computer using the APL operator ⌹. Its transpose has the same shape as X and is shown rounded to four decimal places in Figure 14.2.1. The vector of parameter estimates $\hat{\beta} = X^L y$ is found next, and the fitted model is

$$y = 11.934 - 0.6285t + 0.1298v.$$

Next we obtain the vector of fitted values $\hat{\mu} = X\hat{\beta}$, and the vector of residuals $\hat{\varepsilon} = y - \hat{\mu}$. We can then find the residual sum of squares, $\Sigma\hat{\varepsilon}_i^2 = 10.533$. The variance estimate is $s^2 = \frac{1}{7}\Sigma\hat{\varepsilon}_i^2 = 1.505$ with $10 - 3 = 7$ degrees of freedom.

$$X = \begin{bmatrix} 1 & -3.0 & 15.3 \\ 1 & -1.8 & 16.4 \\ 1 & -10.0 & 41.2 \\ 1 & 0.7 & 9.7 \\ 1 & -5.1 & 19.3 \\ 1 & -6.3 & 11.4 \\ 1 & -15.5 & 5.9 \\ 1 & -4.2 & 24.3 \\ 1 & -8.8 & 14.7 \\ 1 & -2.3 & 16.1 \end{bmatrix}; \quad (X^L)^t = \begin{bmatrix} 0.2072 & 0.0127 & -0.0020 \\ 0.2159 & 0.0189 & -0.0006 \\ -0.4710 & -0.0177 & 0.0270 \\ 0.4085 & 0.0302 & -0.0080 \\ 0.0770 & 0.0030 & 0.0023 \\ 0.2003 & -0.0044 & -0.0072 \\ 0.0769 & -0.0511 & -0.0152 \\ 0.0024 & 0.0083 & 0.0083 \\ 0.0737 & -0.0162 & -0.0037 \\ 0.2092 & 0.0163 & -0.0010 \end{bmatrix}$$

$$y = \begin{bmatrix} 14.96 \\ 14.10 \\ 23.76 \\ 13.20 \\ 18.60 \\ 16.79 \\ 21.83 \\ 16.25 \\ 20.98 \\ 16.88 \end{bmatrix}; \quad \hat{\beta} = \begin{bmatrix} 11.9339 \\ -0.6285 \\ 0.1298 \end{bmatrix}; \quad \hat{\mu} = \begin{bmatrix} 15.81 \\ 15.19 \\ 23.57 \\ 12.75 \\ 17.64 \\ 17.37 \\ 22.44 \\ 17.73 \\ 19.37 \\ 15.47 \end{bmatrix}; \quad \hat{\varepsilon} = \begin{bmatrix} -0.85 \\ -1.09 \\ 0.19 \\ 0.45 \\ 0.96 \\ -0.58 \\ -0.61 \\ -1.48 \\ 1.61 \\ 1.41 \end{bmatrix}.$$

Figure 14.2.1. Calculations for the fuel consumption example.

Problems for Section 14.2

1. Show that

$$\Sigma\hat{\varepsilon}_i^2 = \Sigma y_i^2 - \hat{\beta}^t(X^t y).$$

This formula is useful when calculations are to be done by hand, but it is susceptible to roundoff errors.

2.† Set up the straight line model of Example 13.5.1 in matrix notation. Calculate $(X^t X)^{-1}$ and $X^t y$, and hence obtain the parameter estimates. Use the formula in Problem 1 to obtain the residual sum of squares.

3. Set up the 3-sample model of Example 13.4.3 in matrix notation. Calculate $(X^t X)^{-1}$ and $X^t y$, and hence obtain the parameter estimates. Use the formula in Problem 1 to find the residual sum of squares.

4. The following measurements are from the weighing experiment described in Section 14.1.

Objects weighed	1	2	3	1&2	1&3	2&3	1&2&3
Measurement	12.5	23.9	28.1	31.5	41.9	49.5	61.7

(a) Evaluate $X^t X$ and $X^t y$, and show that

$$(X^t X)^{-1} = \frac{1}{8}\begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

Hence show that the estimated weights are

$$\hat{\beta}_1 = 11.875 \qquad \hat{\beta}_2 = 21.375 \qquad \hat{\beta}_3 = 28.675$$

(b) Use the formula in Problem 1 to evaluate $\Sigma\hat{\varepsilon}_i^2$, and show that $s^2 = 3.08375$ with 4 degrees of freedom.

5.† The yield Y of a chemical process was measured at each of nine different temperatures t_i with the following results:

t_i	15	16	17	18	19	20	21	22	23
y_i	90	91.9	90.7	87.9	86.4	82.5	80.0	76.0	70.0

Consider the 2nd degree polynomial model

$$\mu_i = \beta_1 + (t_i - 19)\beta_2 + (t_i - 19)^2\beta_3 \qquad \text{for } i = 1, 2, \ldots, 9.$$

(a) Write this model in the form $\mu = X\beta$. Calculate the parameter estimates and the residual sum of squares.

Note:
$$\begin{bmatrix} 9 & 0 & 60 \\ 0 & 60 & 0 \\ 60 & 0 & 708 \end{bmatrix}^{-1} = \begin{bmatrix} 0.255411 & 0 & -0.021645 \\ 0 & 0.016667 & 0 \\ -0.021645 & 0 & 0.003247 \end{bmatrix}$$

(b) Use the fitted model to estimate
(i) the expected yield μ when $t = 17.5$;
(ii) the temperature t for which the expected yield is greatest.

(c) Plot the data and the fitted model. Does the model appear to give a good fit to the data?

6. Consider the linear model $\mu = X\beta$, and suppose that the columns of X are mutually orthogonal:

$$x_{1j}x_{1k} + x_{2j}x_{2k} + \cdots + x_{nj}x_{nk} = 0 \qquad \text{for } j \neq k.$$

Let c_j denote the sum of squares of elements in the jth column:

$$c_j = x_{1j}^2 + x_{2j}^2 + \cdots + x_{nj}^2.$$

Prove the following results:

$$\hat{\beta}_j = \frac{1}{c_j}\sum_{i=1}^{n} x_{ij}y_i; \qquad \operatorname{var}(\hat{\beta}_j) = \sigma^2/c_j.$$

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} y_i^2 - \sum_{j=1}^{q} \hat{\beta}_j^2 c_j.$$

7. Suppose that a linear model contains a general constant (intercept) term β_1, so that

$$\mu_i = 1 \cdot \beta_1 + x_{i2}\beta_2 + \cdots + x_{iq}\beta_q \qquad \text{for } i = 1, 2, \ldots, n.$$

Show that, in this case, the residuals $\hat{\varepsilon}_i$ must sum to zero.

14.3. Testing Hypotheses in Linear Models

In this section we shall derive the likelihood ratio statistic for testing an hypothesis about the parameters $\beta_1, \beta_2, \ldots, \beta_q$ in a linear model.

As in the preceding section we suppose that $Y_1, Y_2, \ldots, Y_n$ are independent $N(\mu_i, \sigma^2)$, and that

$$\mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iq}\beta_q \qquad \text{for } i = 1, 2, \ldots, n.$$

In matrix notation this is $\mu = X\beta$ where X is an $n \times q$ matrix with linearly independent columns. The model involves $q+1$ unknown parameters $\beta_1, \beta_2, \ldots, \beta_q$ and σ. By (13.2.1) the log likelihood function is

$$l(\beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} S(\beta)$$

where $S(\beta)$ is the error sum of squares:

$$S(\beta) = \Sigma\varepsilon_i^2 = \Sigma(y_i - \mu_i)^2.$$

Note that

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} S(\beta)$$

and so the MLE of σ for a given value of β is

$$\hat{\sigma}^2(\beta) = \frac{1}{n} S(\beta).$$

Let $\hat{\beta}$ be the MLE of β under the model $\mu = X\beta$, and let $\hat{\varepsilon} = Y - X\hat{\beta}$. Then

$$\hat{\sigma}^2 = \frac{1}{n} S(\hat{\beta}) = \frac{1}{n}\Sigma\hat{\varepsilon}_i^2,$$

and the maximum of the log likelihood is

$$l(\hat{\beta}, \hat{\sigma}) = -n \log \hat{\sigma} - \frac{1}{2\hat{\sigma}^2} S(\hat{\beta}) = -\frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2}.$$

Now let H be an hypothesis which expresses the q parameters $\beta_1, \beta_2, \ldots, \beta_q$ as functions of p new parameters $\gamma_1, \gamma_2, \ldots, \gamma_p$, where $p < q$. If the γ_j's are functionally independent, there are $q - p$ degrees of freedom for testing H (see Section 12.3).

Assuming H to be true, we can find the MLE's $\tilde{\gamma}_1, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_p$, and use these to compute $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_q$ and $\tilde{\varepsilon} = Y - X\tilde{\beta}$. The new MLE of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n} S(\tilde{\beta}) = \frac{1}{n}\Sigma\tilde{\varepsilon}_i^2,$$

and the maximum log likelihood under H is

$$l(\tilde{\beta}, \tilde{\sigma}) = -n \log \tilde{\sigma} - \frac{1}{2\tilde{\sigma}^2} S(\tilde{\beta}) = -\frac{n}{2} \log \tilde{\sigma}^2 - \frac{n}{2}.$$

The likelihood ratio statistic for testing H is twice the difference between the two maximum log likelihoods:

$$D = 2[l(\hat{\beta}, \hat{\sigma}) - l(\tilde{\beta}, \tilde{\sigma})] = n \log(\tilde{\sigma}^2/\hat{\sigma}^2).$$

It follows that

$$D = n \log(\Sigma\tilde{\varepsilon}_i^2/\Sigma\hat{\varepsilon}_i^2) = n \log\left[1 + \frac{Q}{\Sigma\hat{\varepsilon}_i^2}\right] \tag{14.3.1}$$

where Q is the increase in the residual sum of squares due to the hypothesis:

$$Q = \Sigma\tilde{\varepsilon}_i^2 - \Sigma\hat{\varepsilon}_i^2. \tag{14.3.2}$$

Q is called the *additional sum of squares due to* H. Since $D \geq 0$, it follows that $Q \geq 0$.

There are $q - p$ degrees of freedom for testing H, and so (12.3.2) gives

$$\text{SL} \approx P\{\chi^2_{(q-p)} \geq D_{\text{obs}}\}.$$

This approximation will be accurate whenever n is much larger than q.

In what follows, we shall consider the special case of a linear hypothesis H. We shall see that, when H is linear, the significance level can be computed exactly.

Testing Linear Hypotheses

Let H be an hypothesis which expresses $\beta_1, \beta_2, \ldots, \beta_q$ as functions of p new parameters $\gamma_1, \gamma_2, \ldots, \gamma_p$, where $p < q$. H is called a linear hypothesis if the β_i's can be written as linear functions of the γ_j's:

$$\beta_i = b_{i1}\gamma_1 + b_{i2}\gamma_2 + \cdots + b_{ip}\gamma_p \qquad \text{for } i = 1, 2, \ldots, q.$$

In matrix notation, the linear hypothesis is

$$H\colon \beta = b\gamma$$

where b is a $q \times p$ matrix of constants. We can assume that the columns of b are linearly independent, since otherwise it would be possible to rewrite the β_i's as functions of only $p-1$ of the γ_j's.

If H is true, then

$$\mu = X\beta = Xb\gamma = W\gamma$$

where $W = Xb$ is $n \times p$. Under H we have another linear model $\mu = W\gamma$, and so the MLE $\tilde{\gamma}$ can be found as in Section 14.2.

We shall show in Section 14.6 that, if H is a linear hypothesis, then Q is distributed independently of the smaller residual sum of squares $\Sigma\hat{\varepsilon}_i^2$, and

$$Q/\sigma^2 \sim \chi^2_{(q-p)} \qquad \text{if } H \text{ is true;}$$
$$\Sigma\hat{\varepsilon}_i^2/\sigma^2 \sim \chi^2_{(n-q)}.$$

If H is true, the quantity $Q/\Sigma\hat{\varepsilon}_i^2$ in (14.3.1) is distributed as a ratio of independent χ^2 random variables.

To obtain a quantity whose distribution is tabulated, we divide each χ^2 variate by its degrees of freedom before taking their ratio. Thus we consider the F-statistic:

$$F = \frac{(Q/\sigma^2) \div (q-p)}{(\Sigma\hat{\varepsilon}_i^2/\sigma^2) \div (n-q)} = \frac{Q \div (q-p)}{s^2}.$$

It follows by (6.10.1) that, if H is true, then F has a variance ratio (F) distribution with $q-p$ numerator and $n-q$ denominator degrees of freedom:

$$F = \frac{Q \div (q-p)}{s^2} \sim F_{q-p,\,n-q}. \tag{14.3.3}$$

The numerator of F is the variance estimate based on the additional sum of squares. The denominator is the variance estimate $s^2 = \Sigma\hat{\varepsilon}_i^2/(n-q)$ for the model $\mu = X\beta$.

Note that the likelihood ratio statistic D is an increasing function of F:

$$D = n \log[1 + Q/\Sigma\hat{\varepsilon}_i^2] = n \log\left[1 + \frac{q-p}{n-q}F\right].$$

Large values of D correspond to large values of F, and so

$$\begin{aligned} \text{SL} &= P\{D \geq D_{\text{obs}} | H \text{ is true}\} \\ &= P\{F \geq F_{\text{obs}} | H \text{ is true}\} \\ &= P\{F_{q-p,\,n-q} \geq F_{\text{obs}}\} \end{aligned} \tag{14.3.4}$$

which can be evaluated using Table B5.

EXAMPLE 14.3.1. Consider the fuel consumption data of Example 14.2.1. We fitted the linear model $\mu_i = \beta_1 + \beta_2 t_i + \beta_3 v_i$ and obtained $\Sigma\hat{\varepsilon}_i^2 = 10.533$ and $s^2 = 1.505$ with $n - q = 10 - 3 = 7$ degrees of freedom.

Consider the hypothesis H: $\beta_3 = 0$, which states that wind velocity v has no effect on the mean fuel consumption. The hypothesized model is $\mu_i = \beta_1 + \beta_2 t_i$, which is another linear model with $q = 2$ unknown β_j's. We can fit this model to the data as in Section 13.5, or we can omit the last column of X and repeat the calculation of Example 14.2.1. In either case, we find that the residual sum of squares is $\Sigma\tilde{\varepsilon}_i^2 = 24.944$ with $n - p = 8$ degrees of freedom. The additional sum of squares due to H is then

$$Q = \Sigma\tilde{\varepsilon}_i^2 - \Sigma\hat{\varepsilon}_i^2 = 24.944 - 10.533 = 14.411$$

with $q - p = 1$ degree of freedom.

To test H: $\beta_3 = 0$ we compute the observed value of the F-statistic:

$$F_{\text{obs}} = \frac{Q \div 1}{s^2} = \frac{14.411}{1.505} = 9.58.$$

The significance level is then

$$\text{SL} = P\{F_{1,\,7} \geq 9.58\} \approx 0.02$$

from Table B5. Thus there is evidence that wind velocity does have an effect on mean fuel consumption, and the wind velocity term should be kept in the linear model.

When $q - p = 1$ as in the present example, it is also possible to test H by the method described in Section 13.2. In the next section we will show that these two methods of testing H will always give the same significance level.

EXAMPLE 14.3.2. Table 13.4.2 shows 27 concrete penetration measurements from a study on a pulse-jet pavement breaker using nozzles of three sizes. These data were analyzed using a 3-sample model in Example 13.4.3. We assumed that the y_{ij}'s were observed values of independent $N(\mu_{ij}, \sigma^2)$ variates, and that

$$\mu_{i1} = \mu_{i2} = \cdots = \mu_{i9} = \mu_i \qquad \text{for } i = 1, 2, 3.$$

The variance estimate was found to be $s^2 = 140$ with $n - q = 24$ degrees of freedom, and so the residual sum of squares for this model is

$$\Sigma\Sigma\hat{\varepsilon}_{ij}^2 = 24s^2 = 3360.$$

Now consider the hypothesis

$$H: \mu_1 = \mu_2 = \mu_3$$

which states that the mean penetration is the same for all three nozzle sizes. There are $q - p = 2$ degrees of freedom for testing H. Under H, the 27 observations form a single sample, and the residual sum of squares is

$$\Sigma\Sigma\tilde{\varepsilon}_{ij}^2 = \Sigma\Sigma(y_{ij} - \bar{y})^2 = 4886.17$$

with 26 d.f., where $\bar{y}$ is the grand mean of all 27 observations.

The additional sum of squares due to H is

$$Q = \Sigma\Sigma\tilde{\varepsilon}_{ij}^2 - \Sigma\Sigma\hat{\varepsilon}_{ij}^2 = 1526.17$$

with 2 degrees of freedom. The observed value of the F-statistic is

$$F_{\text{obs}} = \frac{Q \div 2}{s^2} = \frac{1526.17 \div 2}{140} = 5.45.$$

Now Table B5 gives

$$\text{SL} = P\{F_{2,24} \geq 5.45\} \approx 0.01.$$

There is strong evidence against the hypothesis of equal means. The small nozzle diameter gives a significantly lower mean penetration than the medium or large diameter.

It would have been possible to design a test specifically for the purpose of determining whether mean penetration is significantly less for small diameter nozzles. For instance, the test statistic $D = \frac{1}{2}(\bar{Y}_2 + \bar{Y}_3) - \bar{Y}_1$ could be used. It would not be valid to do this unless this particular type of departure had been anticipated prior to examination of the data. See the discussion on tests suggested by the data in Section 12.1.

EXAMPLE 14.3.3. Table 13.5.1 gives the log lifetimes of 40 plastic gears tested at 9 different temperatures $x_1, x_2, \ldots, x_9$. A straight line model was fitted to these data in Example 13.5.2. It is also possible to ignore the values of $x_1, x_2, \ldots, x_9$ and analyze the data as nine independent samples. Using the residual sums of squares for these two models, we can test the adequacy of the straight line model.

Let y_{ij} be the log lifetime for the jth gear tested at the ith temperature. We assume that the y_{ij}'s are independent $N(\mu_{ij}, \sigma^2)$, and that

$$\mu_{i1} = \mu_{i2} = \cdots = \mu_{in_i} = \mu_i \qquad \text{for } i = 1, 2, \ldots, 9.$$

This is the 9-sample model. From Section 13.4, the MLE of μ_i is $\bar{y}_i$, the sample mean for the ith sample. The residual sum of squares is

$$\Sigma\Sigma\hat{\varepsilon}_{ij}^2 = \Sigma\Sigma(y_{ij} - \bar{y}_i)^2 = \Sigma(n_i - 1)s_i^2$$

where $s_1^2, s_2^2, \ldots, s_9^2$ are the nine sample variances. Using the results from Table 14.3.1 we get

$$\Sigma\Sigma\hat{\varepsilon}_{ij}^2 = 0.64046$$

Table 14.3.1. Sample Means and Variances for the Plastic Gear Data

Sample No. (i)	Temp. x_i	Sample size n_i	Sample mean $\bar{y}_i$	Sample variance s_i^2	df $n_i - 1$
1	−16	4	1.755	0.0064	3
2	0	4	1.569	0.0058	3
3	10	4	1.094	0.0101	3
4	21	8	0.808	0.0188	7
5	30	4	0.494	0.0265	3
6	37	4	0.515	0.0134	3
7	47	4	0.253	0.0353	3
8	57	4	−0.359	0.0060	3
9	67	4	−0.687	0.0661	3

with $\Sigma(n_i - 1) = n - 9 = 31$ degrees of freedom. The (pooled) variance estimate for the 9-sample model is

$$s^2 = \frac{1}{31}\Sigma\Sigma\hat{\varepsilon}_{ij}^2 = 0.02066.$$

Now consider the hypothesis

$$H: \mu_i = \alpha + \beta x_i \qquad \text{for } i = 1, 2, \ldots, 9.$$

This is a linear hypothesis which reduces the number of unknown parameters by 7, so there are 7 degrees of freedom for testing H. Under H we have a straight line model, and from Example 13.5.2 the residual sum of squares is

$$\Sigma\Sigma\tilde{\varepsilon}_{ij}^2 = 1.24663$$

with $n - 2 = 38$ degrees of freedom. The additional sum of squares due to H is

$$Q = \Sigma\Sigma\tilde{\varepsilon}_{ij}^2 - \Sigma\Sigma\hat{\varepsilon}_{ij}^2 = 0.60617$$

with 7 degrees of freedom.

To test H, we compute

$$F_{\text{obs}} = \frac{Q \div 7}{s^2} = \frac{0.60617 \div 7}{0.02066} = 4.19.$$

Now Table B5 gives

$$\text{SL} = P\{F_{7,31} \geq 4.19\} < 0.01.$$

The test gives very strong evidence against $H: \mu_i = \alpha + \beta x_i$.

The reason for this result is apparent from Figure 13.5.2, where at several temperatures all of the observations lie on the same side of the fitted line. There is no simple pattern to the departures from the line, and it is not obvious how the straight line model could be altered to give a satisfactory fit.

The most likely explanation of the small significance level is that, owing to

the way the experiment was performed, the variance estimate s^2 in the denominator of F_{obs} is too small. A considerable amount of time and effort was required to reset the test machine from one temperature to another. To save time, the experimenter sometimes ran two or more tests at the same temperature without resetting the test machine. Repeat measurements obtained without resetting will likely show less scatter than would be obtained if the machine were reset each time. These repeats do not reflect all possible sources of variability in the experiment, and consequently the variance estimate we obtained is likely to be too small. We do not have a valid estimate of σ^2, and so interpretation of the results is not clearcut.

The experiment should have been run in four complete replications. In the first replication, one gear would be tested at each temperature, with the order of testing being decided at random. This procedure would then be repeated three more times, with a different random order each time. The four measurements at the same temperature would then be genuine independent replicates, and it would be possible to obtain an estimate of σ^2 which takes into account all sources of variability in the experiment.

Problems for Section 14.3

1.† Measurements of breaking strength for six bolts at each of five diameters are given in Problem 13.5.2. Three different models are fitted to these data. The residual sum of squares is found to be 0.074317 for a 5-sample model, 0.14066 for a straight line model, and 0.07436 for a second degree polynomial model.

   (a) Assuming the 5-sample model to be correct, test the hypotheses H_1: $\mu_i = \beta_1 + \beta_2 d_i$ and H_2: $\mu_i = \beta_1 + \beta_2 d_i + \beta_3 d_i^2$.
   (b) Assuming the second degree polynomial model to be correct, test the hypothesis that $\beta_3 = 0$.

2. Show that the additional sum of squares due to H: $\mu_1 = \mu_2 = \cdots = \mu_k$ in the k-sample model is given by

$$Q = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2,$$

   where n_i and $\bar{y}_i$ are the sample size and mean for the ith sample, and $\bar{y}$ is the grand mean. Use this formula to check the value given for Q in Example 14.3.2.

3. Consider $n = n_1 + n_2 + \cdots + n_k$ pairs of measurements (x_i, y_{ij}) for $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$. The y_{ij}'s are observed values of independent $N(\mu_{ij}, \sigma^2)$ random variables, where

$$\mu_{i1} = \mu_{i2} = \cdots = \mu_{in_i} = \mu_i \qquad \text{for } i = 1, 2, \ldots, k.$$

   Show that the additional sum of squares due to H: $\mu_i = \alpha + \beta x_i$ for $i = 1, 2, \ldots, k$ is given by

$$Q = \sum_{i=1}^{k} n_i(\bar{y}_i - \hat{\alpha} - \hat{\beta}x_i)^2,$$

   where $\bar{y}_i$ is the mean of the n_i observations at $x = x_i$. Use this formula to check the value for Q in Example 14.3.3.

4. In Problem 13.4.9, test the hypothesis that the mean distance traveled is the same for all fuels, and state the assumptions upon which this test is based.

5.† Several chemical analyses of samples of a product were performed on each of four successive days, and the following table gives the percentage impurity found in each analysis.

Day 1:	2.6	2.6	2.9	2.0	2.1	2.1	2.0
Day 2:	3.1	2.9	3.1	2.5			
Day 3:	2.6	2.2	2.5	2.2	1.2	1.2	1.8
Day 4:	2.5	2.4	3.0	1.5	1.7		

   (a) Assuming equal variances, test whether there is a difference in the mean percentage impurity over the four days.
   (b) Check the equal-variance assumption (see Problem 13.4.5).

6. Three laboratories each carried out five independent determinations of the nicotine content of a brand of cigarettes. Their findings, in milligrams per cigarette, were as follows:

Laboratory A:	16.3	15.6	15.5	16.7	16.2
Laboratory B:	13.5	17.4	16.9	18.2	15.6
Laboratory C:	14.1	13.2	14.3	12.9	12.8

   Are there real differences among the results produced by the three laboratories?

7. Measurements of the ultimate tensile strength (UTS) were made for specimens of insulating foam of five different densities.

Density (x)	Ultimate tensile strength (y)							
4.155	82.8	95.5	97.5	102.8	105.6	107.8	115.7	118.9
3.555	79.7	84.5	85.2	98.0	105.2	113.6		
3.55	71.0	98.2	104.9	106.9	109.6	117.8		
3.23	67.1	77.0	80.3	81.8	83.0	84.1		
4.25	98.5	105.5	111.6	114.5	126.5	127.1		

   Calculate the residual sum of squares for a five-sample model, and for a straight line model. Hence test the hypothesis that the dependence of mean strength on density is linear.

8.† The following table gives measurements of systolic blood pressure for 20 men of various ages:

Age (years)	Blood pressure (mm Hg)				
30	108	110	106		
40	125	120	118	119	
50	132	137	134		
60	148	151	146	147	144
70	162	156	164	158	159

Calculate the residual sums of squares for a five-sample model and for a straight line model. Hence test the hypothesis that the dependence of mean blood pressure on age is linear.

9. Problem 13.7.8 presents eleven sets of three measurements of the amount of a gas dissolving in a liquid.
   (a) Calculate the residual sums of squares for an 11-sample model and a straight line model. Test the hypothesis that the mean amount dissolving is a linear function of the gas content in the mixture.
   (b) Explain why one might expect to obtain a small significance level in (a) even if the straight line model is correct.

10. A procedure sometimes used to check the adequacy of a linear model is to complicate the model by adding extra terms to it, refit, and then test whether the new terms are significantly different from zero. For instance, in Example 14.2.1, the residual sum of squares for the model $\mu_i = \beta_1 + \beta_2 t_i + \beta_3 v_i$ was found to be 10.533. If the more complicated model

$$\mu_i = \beta_1 + \beta_2 t_i + \beta_3 v_i + \beta_4 t_i^2 + \beta_5 v_i^2 + \beta_6 t_i v_i$$

is fitted to the data, the residual sum of squares decreases to 4.442. Using these results, test the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$. (A small significance level would indicate possible difficulties with the simpler model.)

11. Consider two linear models $\mu = X\beta$ and $\mu = W\gamma$ where X and W are $n \times q$ matrices with linearly independent columns. Suppose that $W = Xb$ where b is a $q \times q$ nonsingular matrix. (This means that the columns of W are linear combinations of the columns of X.)
   (a) Show that $\hat{\beta} = b\hat{\gamma}$.
   (b) Show that both models give the same fitted values $\hat{\mu}$ and residuals $\hat{\varepsilon}$.

14.4. More on Tests and Confidence Intervals

In Section 13.2 we described procedures for making inferences about a single parameter β_i in a linear model. These methods can also be used to make inferences about a single linear combination $\theta = b_1\beta_1 + b_2\beta_2 + \cdots + b_q\beta_q$. In this section we give some further discussion of these methods. We shall show that the significance tests described in Section 13.2 are equivalent to likelihood ratio tests, and that the confidence intervals obtained are in fact maximum likelihood intervals.

As in the preceding sections we assume that $Y_1, Y_2, \ldots, Y_n$ are independent $N(\mu_i, \sigma^2)$, and that

$$\mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iq}\beta_q \qquad \text{for } i = 1, 2, \ldots, n.$$

In matrix notation this is $\mu = X\beta$ where X is $n \times q$ with linearly independent

columns. From Section 14.2, the MLE of β is

$$\hat{\beta} = X^L Y$$

where $X^L = (X^t X)^{-1} X^t$ is a $q \times n$ matrix of constants.

Inferences about β_i

The MLE of β_i is a linear combination of the Y_i's:

$$\hat{\beta}_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

where $a_1, a_2, \ldots, a_n$ are the elements in the ith row of X^L. It follows by (6.6.7) that $\hat{\beta}_i \sim N(\Sigma a_j \mu_j, \sigma^2 \Sigma a_j^2)$.

Note that $\Sigma a_j \mu_j$ is the product of the ith row of X^L with the vector μ. This is the ith element of the matrix product $X^L \mu$. But since $\mu = X\beta$ and $X^L X = I$, it follows that

$$X^L \mu = X^L X \beta = I\beta = \beta.$$

Hence the ith element of $X^L \mu$ is β_i, and it follows that

$$E(\hat{\beta}_i) = \Sigma a_j \mu_j = \beta_i.$$

Similarly, Σa_j^2 is the product of the ith row of X^L with itself. This is the (i, i) element of

$$V = X^L (X^L)^t = (X^t X)^{-1}.$$

It follows that $\operatorname{var}(\hat{\beta}_i) = \sigma^2 v_{ii}$, where v_{ii} is the ith diagonal element of V. Similar arguments can be used to show that $\operatorname{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 v_{ij}$.

According to the procedure described in Section 13.2, we start with the sampling distribution $\hat{\beta}_i \sim N(\beta_i, \sigma^2 v_{ii})$. We then standardize and replace σ^2 by s^2, the variance estimate for the model $\mu = X\beta$, to obtain

$$T \equiv \frac{\hat{\beta}_i - \beta_i}{\sqrt{s^2 v_{ii}}} \sim t_{(n-q)}. \tag{14.4.1}$$

We can now test an hypothesis concerning β_i or set up confidence intervals for β_i as in Section 13.2.

Inferences for $\theta = \Sigma b_j \beta_j$

More generally, suppose that we are interested in a linear function of the β_j's,

$$\theta = b_1\beta_1 + b_2\beta_2 + \cdots + b_q\beta_q = b^t\beta$$

where the b_j's are constants and b is $q \times 1$. The MLE of θ is

$$\hat{\theta} = b_1\hat{\beta}_1 + b_2\hat{\beta}_2 + \cdots + b_q\hat{\beta}_q = b^t\hat{\beta}.$$

Since $\hat{\beta} = X^L Y$, it follows that

$$\hat{\theta} = b^t X^L Y = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

where a_i is the ith component of the $1 \times q$ vector $b^t X^L$.

Since $\hat{\theta}$ is a linear combination of the Y_i's, its sampling distribution is normal. Since $E(\hat{\beta}_i) = \beta_i$ we have

$$E(\hat{\theta}) = b_1\beta_1 + b_2\beta_2 + \cdots + b_q\beta_q = \theta.$$

The variance of $\hat{\theta}$ is $\sigma^2 c$, where

$$c = \Sigma a_i^2 = b^t X^L (b^t X^L)^t = b^t X^L (X^L)^t b = b^t V b.$$

Thus we have $\hat{\theta} \sim N(\theta, \sigma^2 c)$, where $c = b^t V b$.

Now, proceeding as in Section 13.2, we standardize and replace σ^2 by s^2 to get

$$T \equiv \frac{\hat{\theta} - \theta}{\sqrt{s^2 c}} \sim t_{(n-q)} \tag{14.4.2}$$

where $c = b^t V b$. Inferences about θ are based on this result.

Note that (14.4.1) is the special case of (14.4.2) in which $b_i = 1$, and $b_j = 0$ for $j \neq i$.

Connection with Likelihood Ratio Tests

We showed in Section 14.3 that the likelihood ratio statistic for testing a linear hypothesis H is

$$D = n \log\left[1 + \frac{Q}{\Sigma \hat{\varepsilon}_i^2}\right] = n \log\left[1 + \frac{q-p}{n-q} F\right]$$

where Q and F are defined in (14.3.2) and (14.3.3).

At the end of this section we shall show that the additional sum of squares due to the hypothesis

$$H: b_1\beta_1 + b_2\beta_2 + \cdots + b_q\beta_q = \theta$$

is given by

$$Q = (\hat{\theta} - \theta)^2 / c \qquad \text{where } c = b^t V b. \tag{14.4.3}$$

There is one degree of freedom for testing H, and so $q - p = 1$. By (14.4.3), (14.3.3), and (14.4.2) we have

$$F = \frac{Q \div 1}{s^2} = \frac{(\hat{\theta} - \theta)^2}{s^2 c} = T^2.$$

The likelihood ratio statistic for testing H is

$$D = n \log\left\{1 + \frac{1}{n-q} T^2\right\},$$

which is an increasing function of T^2. It follows that

$$P\{D \geq D_{\text{obs}}\} = P\{T^2 \geq T^2_{\text{obs}}\} = P\{|T| \geq |T_{\text{obs}}|\},$$

and so the significance tests described in Section 13.2 are equivalent to likelihood ratio tests.

It also follows from these results that the maximum log relative likelihood function of θ is

$$r_{\max}(\theta) = -\tfrac{1}{2}D = -\frac{n}{2} \log\left\{1 + \frac{1}{n-q} T^2\right\}.$$

To construct a confidence interval for θ, we take $-t \leq T \leq t$ where t is the appropriate value from Table B3. The parameter values belonging to the confidence interval are those for which

$$r_{\max}(\theta) \geq -\frac{n}{2} \log\left\{1 + \frac{1}{n-q} t^2\right\}.$$

Hence the confidence interval is a maximum likelihood interval for θ.

EXAMPLE 14.4.1. In Example 14.2.1 we fitted the model

$$\mu_i = \beta_1 + \beta_2 t_i + \beta_3 v_i, \qquad i = 1, 2, \ldots, 10$$

to the fuel consumption data of Table 14.2.1. The matrix $(X^L)^t$ is given in Figure 14.2.1, and from this we can find

$$V = X^L (X^L)^t = \begin{bmatrix} 0.57939 & 0.02503 & -0.01942 \\ 0.02503 & 0.00497 & 0.00017 \\ -0.01942 & 0.00017 & 0.00117 \end{bmatrix}.$$

Parameter β_3 measures the effect of wind velocity on expected fuel consumption. Inferences about β_3 are based on

$$T \equiv \frac{\hat{\beta}_3 - \beta_3}{\sqrt{s^2 v_{33}}} \sim t_{(7)}$$

where $s^2 = 1.505$ (7 d.f.), $\hat{\beta}_3 = 0.1298$, and $v_{33} = 0.00117$. To test H: $\beta_3 = 0$, we compute

$$T_{\text{obs}} = \frac{\hat{\beta}_3 - 0}{\sqrt{s^2 v_{33}}} = 3.09;$$

$$\text{SL} = P\{|t_{(7)}| \geq 3.09\} \approx 0.02$$

from Table B3.

A different procedure was used to test H: $\beta_3 = 0$ in Example 14.3.1. There we refitted the model with $\beta_3 = 0$ and calculated the additional sum of

squares Q. We then found

$$F_{\text{obs}} = \frac{Q \div 1}{s^2} = 9.58;$$

$$\text{SL} = P\{F_{1,7} \geq 9.58\}.$$

Since $F_{\text{obs}} = T_{\text{obs}}^2$, and since $t_{(7)}^2$ is distributed as $F_{1,7}$ by (6.10.7), it follows that

$$P\{F_{1,7} \geq F_{\text{obs}}\} = P\{t_{(7)}^2 \geq T_{\text{obs}}^2\} = P\{|t_{(7)}| \geq |T_{\text{obs}}|\}.$$

Both of these procedures are equivalent to the likelihood ratio test, and therefore they will always give the same significance level.

To complete the example, we shall find a 95% confidence interval for the mean fuel consumption on days when the temperature is -5 and the wind velocity is 20. According to the model, this is

$$\theta = \beta_1 - 5\beta_2 + 20\beta_3 = b^t\beta$$

where $b^t = (1 \ -5 \ 20)$. The MLE of θ is

$$\hat{\theta} = \hat{\beta}_1 - 5\hat{\beta}_2 + 20\hat{\beta}_3 = 17.67.$$

The variance of $\hat{\theta}$ is $\sigma^2 c$ where

$$\begin{aligned} c &= b^t V b \\ &= \begin{bmatrix} 1 & -5 & 20 \end{bmatrix} \begin{bmatrix} 0.57939 & 0.02503 & -0.01942 \\ 0.02503 & 0.00497 & 0.00017 \\ -0.01942 & 0.00017 & 0.00117 \end{bmatrix} \begin{bmatrix} 1 \\ -5 \\ 20 \end{bmatrix} \\ &= 0.11025. \end{aligned}$$

Now, by (14.4.2), the 95% confidence interval for θ is

$$\theta \in \hat{\theta} \pm 2.365\sqrt{s^2 c} = 17.67 \pm 0.96.$$

We know that this is also a maximum likelihood interval, so each value of θ belonging to the interval has a higher maximum relative likelihood than any value outside the interval.

PROOF OF (14.4.3). Let $\tilde{\beta}$ denote the MLE of β under the hypothesis

$$H: b_1\beta_1 + b_2\beta_2 + \cdots + b_q\beta_q = \theta,$$

and let $\tilde{\varepsilon} = Y - X\tilde{\beta}$ be the vector of residuals. Then $\tilde{\beta}$ is the value of β which minimizes

$$S = \Sigma(y_i - \mu_i)^2 = \Sigma(y_i - x_{i1}\beta_1 - \cdots - x_{iq}\beta_q)^2$$

subject to the restriction $\Sigma b_j\beta_j = \theta$.

To find $\tilde{\beta}$, we use the method of Lagrange. We define a new function of $q+1$ variables,

$$g(\beta_1, \beta_2, \ldots, \beta_q, \lambda) = S(\beta_1, \beta_2, \ldots, \beta_q) + 2\lambda(\Sigma b_j\beta_j - \theta).$$

The extra variable λ is called a Lagrange multiplier. We now minimize g over the $q+1$ variables $\beta_1, \beta_2, \ldots, \beta_q$, and λ.

The derivatives of g are

$$\frac{\partial g}{\partial \lambda} = 2(\Sigma b_j\beta_j - \theta);$$

$$\frac{\partial g}{\partial \beta_j} = -2\Sigma x_{ij}(y_i - x_{i1}\beta_1 - \cdots - x_{iq}\beta_q) + 2\lambda b_j.$$

Upon setting these derivatives equal to zero, we find that $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_q$ and $\tilde{\lambda}$ satisfy the $q+1$ equations

$$\Sigma b_j\tilde{\beta}_j = \theta;$$

$$\Sigma x_{ij}\tilde{\varepsilon}_i = \tilde{\lambda}b_j \qquad \text{for } j = 1, 2, \ldots, q.$$

Note that the restriction $\Sigma b_j\beta_j = \theta$ is satisfied at the minimum. Also, if $(\tilde{\beta}, \tilde{\lambda})$ minimizes g, then $\tilde{\beta}$ minimizes S.

In matrix notation, the $q+1$ equations are

$$X^t\tilde{\varepsilon} = \tilde{\lambda}b; \qquad b^t\tilde{\beta} = \theta$$

where b is $q \times 1$. Substituting $\tilde{\varepsilon} = Y - X\tilde{\beta}$ gives

$$X^t(y - X\tilde{\beta}) = \tilde{\lambda}b$$

and now multiplying by $V = (X^tX)^{-1}$ gives

$$\tilde{\beta} = (X^tX)^{-1}X^ty - \tilde{\lambda}(X^tX)^{-1}b = \hat{\beta} - \tilde{\lambda}Vb.$$

Since $b^t\tilde{\beta} = \theta$, it follows that

$$\theta = b^t\hat{\beta} - \tilde{\lambda}b^tVb = \hat{\theta} - \tilde{\lambda}c$$

where $\hat{\theta} = b^t\hat{\beta}$ and $c = b^tVb$, and therefore

$$\tilde{\lambda} = (\hat{\theta} - \theta)/c.$$

Also, since $\tilde{\beta} = \hat{\beta} - \tilde{\lambda}Vb$, we have

$$\tilde{\varepsilon} = y - X\tilde{\beta} = y - X\hat{\beta} + \tilde{\lambda}XVb = \hat{\varepsilon} + \tilde{\lambda}XVb.$$

The residual sum of squares under H is

$$\begin{aligned} \Sigma\tilde{\varepsilon}_i^2 = \tilde{\varepsilon}^t\tilde{\varepsilon} &= (\hat{\varepsilon} + \tilde{\lambda}XVb)^t(\hat{\varepsilon} + \tilde{\lambda}XVb) \\ &= \hat{\varepsilon}^t\hat{\varepsilon} + \tilde{\lambda}^2 b^tV^tX^tXVb + \text{cross-product terms.} \end{aligned}$$

The cross-product terms are zero because $X^t\hat{\varepsilon} = 0$ by (14.2.2). Since $V = (X^tX)^{-1}$ and $V^t = V$, it follows that

$$\Sigma\tilde{\varepsilon}_i^2 = \Sigma\hat{\varepsilon}_i^2 + \tilde{\lambda}^2 b^tVb = \Sigma\hat{\varepsilon}_i^2 + \tilde{\lambda}^2 c.$$

Hence the additional sum of squares due to H is

$$Q = \Sigma\tilde{\varepsilon}_i^2 - \Sigma\hat{\varepsilon}_i^2 = \tilde{\lambda}^2 c = (\hat{\theta} - \theta)^2/c$$

which is (14.4.3). □

PROBLEMS FOR SECTION 14.4

1.† Consider the data from the weighing experiment in Problem 14.2.4.

(a) Find a 95% confidence interval for β_3.
(b) Test the hypothesis $\beta_2 = 2\beta_1$ in two ways:
  (i) use (14.4.2);
  (ii) use the hypothesis to simplify the model, then refit and use the additional sum of squares method.
(c) Assuming the simplified model in b (ii), recalculate the 95% confidence interval for β_3.

2. Consider the chemical yield data in Problem 14.2.5.

(a) Find a 95% confidence interval for the expected yield of the process when $t = 17.5$.
(b) Test the hypothesis $\beta_3 = 0$ in two ways:
  (i) use (14.4.2);
  (ii) use the additional sum of squares method.

3.† Thirteen sets of observations were taken on the variables y, x_1, x_2, and x_3. Here y is the percentage of bacteria surviving a treatment, and x_1, x_2, x_3 are the concentrations of three chemicals used in the treatment. The model

$$E(Y_i) = 1 \cdot \beta_1 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 \qquad 1 \le i \le 13$$

was fitted to the data by least squares, with the following results:

$$\hat{\beta} = \begin{bmatrix} 39.16 \\ 1.02 \\ -1.86 \\ -0.34 \end{bmatrix}; \qquad (X'X)^{-1} = \begin{bmatrix} 8.06 & -0.08 & -0.09 & -0.79 \\ -0.08 & 0.008 & 0.002 & 0.003 \\ -0.09 & 0.002 & 0.017 & 0.002 \\ -0.79 & 0.003 & 0.002 & 0.087 \end{bmatrix}$$

The residual sum of squares was 38.7.

(a) Obtain a 95% confidence interval for $\beta_2 - \beta_3$.
(b) Test the hypothesis $\beta_3 = 0$.
(c) The model was refitted with $\beta_2 = \beta_4 = 0$, and the new residual sum of squares was found to be 167.6. Test the hypothesis $\beta_2 = \beta_4 = 0$.

4. An experiment is performed to investigate the dependence of burst strength (y) on crack length (x_1) and operating temperature (x_2) in pressure tubes. Cracks of three different lengths are cut in specimens, and half of these are cycled to sharpen the cracks. The specimens are then tested at three different temperatures, and the burst strength is determined. The following table gives the results (simplified by changes of origin and scale).

	Blunt cracks (not cycled)			Sharp cracks (cycled)		
	$x_1 = -1$	$x_1 = 0$	$x_1 = 1$	$x_1 = -1$	$x_1 = 0$	$x_1 = 1$
$x_2 = -1$	10	10	8	5	1	2
$x_2 = 0$	16	14	13	8	7	6
$x_2 = 1$	17	16	11	16	12	8

It is assumed that burst strength is a linear function of x_1 and x_2, that the effect of cycling is the same at all levels of crack length and temperature, and that errors are independent $N(0, \sigma^2)$.

(a) Using an indicator variable $x_3 = \pm 1$ for crack type, set up a linear model corresponding to the above assumptions.
(b) Fit the model by least squares, and compute the variance estimate. (If you did (a) correctly, $X'X$ will be a diagonal matrix, and the calculations are easy.)
(c) Test the hypothesis that cycling has no effect on mean burst strength.
(d) Obtain a 95% confidence interval for the mean burst strength of pressure tubes with sharp cracks of length $+1$ at the lowest operating temperature.

5. Two random samples of 11 lambs each were used in an experiment to assess the effect of a treatment on body weight. One sample received the treatment, and the other sample served as the control group (no treatment). The body weight y (in pounds) and age x (in days) were recorded for each animal at the end of the experiment.

Control	y	35	34	34	35	26	32	24	33	23	20	15
	x	83	81	80	78	73	72	72	70	70	65	54
Treated	y	45	44	44	46	42	39	38	40	38	31	23
	x	90	83	80	80	79	74	72	70	66	54	50

(a) Plot the data. Fit different straight line models for the two samples, and compute the total residual sum of squares.
(b) Fit parallel straight lines to the two samples and calculate the residual sum of squares.
(c) Use the additional sum of squares method to test the hypothesis of equal slopes.
(d) Repeat (c) using a t-test.
(e) Assuming the parallel line model, find a 95% confidence interval for the increase in mean weight due to the treatment.

14.5. Checking the Model

Example 13.5.3 demonstrates the importance of plotting the data to check that a straight line model is reasonable. This is doubly important with more complex linear models. In this section we briefly describe some procedures for checking the assumptions which underly the normal linear model. Most of

these involve looking for patterns in plots of residuals or standardized residuals. For a more detailed discussion, see Chapter 3 of N. Draper and H. Smith, *Applied Regression Analysis*, 2nd Edition, Wiley (1981).

The residual $\hat{\varepsilon}_i = y_i - \hat{\mu}_i$ can be regarded as an estimate of the error $\varepsilon_i \equiv Y_i - \mu_i$. Since the Y_i's are assumed to be independent $N(\mu_i, \sigma^2)$, the ε_i's are independent $N(0, \sigma^2)$. Thus, if the model is correct, we would expect the $\hat{\varepsilon}_i$'s to look like independent observations from $N(0, \sigma^2)$.

In fact, the $\hat{\varepsilon}_i$'s have unequal variances less than σ^2. To see this we note that, by (14.2.1) and (14.2.4),

$$\hat{\mu} = X\hat{\beta} = XX^L y = My,$$

where $M = (m_{ij})$ is an $n \times n$ symmetric matrix:

$$M = XX^L = X(X^tX)^{-1}X^t.$$

Thus $\hat{\mu}_i$ and $\hat{\varepsilon}_i$ are linear combinations of $y_1, y_2, \ldots, y_n$:

$$\hat{\mu}_i = m_{i1}y_1 + m_{i2}y_2 + \cdots + m_{in}y_n; \tag{14.5.1}$$

$$\hat{\varepsilon}_i = y_i - \hat{\mu}_i = y_i - m_{i1}y_1 - m_{i2}y_2 - \cdots - m_{in}y_n. \tag{14.5.2}$$

Using (5.5.5), (5.5.6), and properties of the matrix M, it can be shown that $E(\hat{\varepsilon}_i) = 0$ and $\text{var}(\hat{\varepsilon}_i) = (1 - m_{ii})\sigma^2$. Thus $\hat{\varepsilon}_i \div \sigma\sqrt{1 - m_{ii}}$ has a standardized normal distribution.

Usually σ will be unknown, and we replace it by the estimate s to obtain the *standardized residual*

$$r_i = \hat{\varepsilon}_i \div s\sqrt{1 - m_{ii}}.$$

If the model is satisfactory, then $r_1, r_2, \ldots, r_n$ should look like observations from $N(0, 1)$. About 95% of the r_i's should lie in the interval $(-2, 2)$, and about 68% in the interval $(-1, 1)$. For most purposes, it is better to plot standardized residuals, although usually a graph of the residuals will show a similar pattern. The r_i's, like the $\hat{\varepsilon}_i$'s, are slightly correlated, but this does not seem to create difficulties provided that n, the number of observations, is much larger than q, the number of β_j's in the linear model.

Calculation of Leverages

The quantity m_{ii} which appears in the expression for the standardized residual r_i is called the *leverage* of the ith point. The leverages are the diagonal elements of the $n \times n$ matrix $M = XX^L$. Since $X^LX = I$, it follows that $MM = M$, and from this result it can be shown that $0 \le m_{ii} \le 1$.

When the matrix methods of Section 14.2 are used, the leverages may be calculated by multiplying X and $(X^L)^t$ term by term to obtain a new $n \times q$ matrix and then finding its row totals.

When algebraic formulas are used as in Chapter 13, we write $\hat{\mu}_i$ as a linear

combination of $y_1, y_2, \ldots, y_n$ and then find m_{ii} as the coefficient of y_i. For instance, in the one-sample problem we have

$$\hat{\mu}_i = \bar{y} = \frac{1}{n}(y_1 + \cdots + y_i + \cdots + y_n)$$

and so $m_{ii} = \frac{1}{n}$. In the k-sample problem, $\hat{\mu}_i$ is the sample mean $\bar{y}_j$ for the sample which contains the ith observation. The leverage is $m_{ii} = 1/n_j$ where n_j is the number of observations in this sample. Thus in Example 13.4.2, the 8 observations at 21°C have leverages $\frac{1}{8}$, and the other 4 observations at 30°C have leverages $\frac{1}{4}$.

Residual Plots

It is a good idea to plot the standardized residuals r_i or residuals $\hat{\varepsilon}_i$ versus the fitted values $\hat{\mu}_i$. Four such plots are shown in Figure 14.5.1. In the first of these, the n points $(r_i, \hat{\mu}_i)$ lie in a band of roughly constant width about zero, and the model appears to be satisfactory. In (ii), there is more scatter in the residuals as $\hat{\mu}$ increases. A possible explanation of this is that the error variance σ^2 is not constant. A nonlinear transformation of the y_i's, such as a logarithmic transformation, may help to remedy the problem.

In the third diagram there is an outlying point. If this point is removed and the model is refitted, the new residual plot should resemble case (i). An explanation for the outlier should be sought, and both the outlier and the revised analysis should be reported.

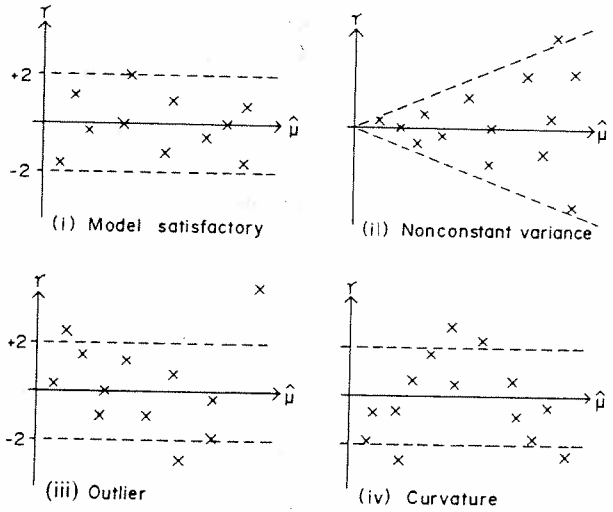

Figure 14.5.1. Patterns in residual plots.

In (iv) there is a pattern in the residual plot which suggests that the means μ_i have not been modelled correctly. This is also a possible explanation in (ii). With a straight line model, addition of a quadratic term to the model should fix the problem, but some detective work may be needed to find a remedy with more complex models.

For straight line models, a plot of r_i or $\hat{\varepsilon}_i$ versus $\hat{\mu}_i$ gives essentially the same information as a plot of y_i versus x_i. However, patterns will show up more clearly in the residual plot which shows only deviations from the fitted line.

It can be shown that $\operatorname{cov}(\hat{\varepsilon}_i, \hat{\mu}_i) = 0$, but that $\operatorname{cov}(\hat{\varepsilon}_i, y_i) = \operatorname{var}(\hat{\varepsilon}_i) = (1 - m_{ii})\sigma^2$. A plot of $\hat{\varepsilon}_i$ or r_i versus observed values y_i will generally *not* be helpful, because we expect it to show a pattern even if the model is correct.

Various other residual plots may be useful, depending upon the situation. For instance, we might wish to plot the residuals in Example 13.5.1 against the weights of the women if these were available. In the plastic gear example, we might plot the residuals in the order that the corresponding measurements were made, with the purpose of checking whether there was a systematic change in laboratory conditions.

Although residual plots are very useful in statistical analysis, a word of caution is necessary. Even if the model is correct, random variation will produce patterns in the residuals rather more often than most people would expect. Many beginners spot an "unusual" pattern in almost every residual plot that they examine!

To help judge whether an observed pattern is "real", it is helpful to generate n observations from $N(0, 1)$ on the computer and plot them in the same way as the residuals. By repeating this several times one gets a feeling for the amount of random variation, and is in a better position to interpret the observed graph.

Checking the Independence Assumption

The assumption that the Y_i's are independent is crucial to the analysis. Clusters of points in a plot of the data or the residuals may indicate a lack of independence. It is important to learn as much as possible about how the data were collected in order that such difficulties can be anticipated and explained. In the plastic gear example, knowledge of the way the experiment was run is enough to suggest that repeat observations at the same temperature are not independent replicates (see Example 14.3.3).

If the y_i's are observed sequentially in time, there may be some carryover effect from one time period to the next. For instance, monthly expenditure figures may show a zigzag pattern because month-end expenses are sometimes included in the completed month and sometimes in the new month. If y_{i-1} is large, then y_i will tend to be small. This sort of dependence is called a serial correlation, and it should show up as a trend in a scatter-plot of the $n-1$ points $(\hat{\varepsilon}_i, \hat{\varepsilon}_{i-1})$ or (r_i, r_{i-1}). Another possibility is to examine runs of positive and negative residuals, and compare the observed results with theoretical results for random sequences (see Section 2.7).

Looking for Influential Points

Sometimes just one or two of the observations will have a large influence on the analysis. The fourth data set in Example 13.5.3 shows an extreme case in which the estimation of the slope is entirely dependent upon one observation at $x = 19$. We would like to be able to detect influential points when they occur in more complex linear models.

By (14.2.4) we have $\hat{\beta} = X^L y$ where $X^L = (x_{ij}^L)$ is $q \times n$. It follows that

$$\hat{\beta}_i = x_{i1}^L y_1 + x_{i2}^L y_2 + \cdots + x_{in}^L y_n \qquad \text{for } i = 1, 2, \ldots, q.$$

Thus x_{ij}^L is the amount by which $\hat{\beta}_i$ would change as a result of a unit increase in y_j. By examining the elements in the ith row of X^L, we can determine whether there are one or two points which strongly influence the estimation of β_i.

Similarly, (14.5.1) shows that increasing y_i by one unit will change $\hat{\mu}_i$ by m_{ii} units. If the leverage m_{ii} is close to 1, then $\hat{\mu}_i$ is determined almost entirely by just the one observation y_i.

In standardizing residuals, we divide by $\sqrt{1 - m_{ii}}$. The effect of this is to increase the magnitudes of residuals for influential points relative to those for less influential points.

EXAMPLE 14.5.1. Consider the fuel consumption example of Section 14.2. The matrix $(X^L)^t$ is shown in Figure 14.2.1. Looking down the 2nd column (which is the 2nd row of X^L), we see that y_7 is the most influential observation in determining the estimated temperature effect $\hat{\beta}_2$. An increase of 1 unit in y_7 would change $\hat{\beta}_2$ from -0.6285 to -0.6796. Similarly, an increase of 1 unit in y_3 would produce a rather large change in $\hat{\beta}_3$, from 0.1298 to 0.1568.

To find the leverages m_{ii}, we multiply X and $(X^L)^t$ in Figure 14.2.1 term-by-term to obtain a new 10×3 matrix, and then calculate the row totals. The results are as follows:

i	1	2	3	4	5	6	7	8	9	10
m_{ii}	0.14	0.17	0.82	0.35	0.11	0.15	0.78	0.17	0.16	0.16

Two observations, y_3 and y_7, are highly influential in determining the fitted values. For instance, a unit increase in y_3 would increase $\hat{\mu}_3$ by 0.82. The fitted value at $t_3 = -10$, $v_3 = 41.2$ is determined almost entirely by just the one observation y_3, because there are no other points (t_i, v_i) close to $(-10, 41.2)$. One would be reluctant to put much faith in the model for values of (t, v) in this vicinity.

Checking the Normality Assumption

Failure of the normality assumption is less serious than lack of independence or incorrect modelling of the μ_i's. So long as the Y_i's are not too far from normal, it is still reasonable to estimate parameters by minimizing $\Sigma(y_i-\mu_i)^2$, and the analysis will give sensible results. The biggest effect of non-normality is on variance estimation. If the Y_i's are non-normal, the distribution of $\Sigma\hat{\varepsilon}_i^2/\sigma^2$ may be quite unlike $\chi^2_{(n-q)}$, and intervals for σ^2 based on (13.2.4) may be seriously in error. Tests and confidence intervals for the β_j's are much less severely affected by departures from normality.

Note that, by (14.5.1), $\hat{\varepsilon}_i$ is a linear combination of $Y_1, Y_2, \ldots, Y_n$. Because of the Central Limit Theorem, the distribution of $\hat{\varepsilon}_i$ may be close to normal even when the Y_i's are decidedly non-normal. The fact that the $\hat{\varepsilon}_i$'s or r_i's appear normal does not imply that the normality assumption is correct. One needs a large number of independent replicate measurements in order to have a good chance of detecting non-normality.

Many graphical procedures for checking normality have been proposed in the literature. A problem with all of these is the difficulty in judging whether a pattern in the graph is indicative of departures from normality rather than chance variation in the data.

EXAMPLE 14.5.2. Consider the analysis of the plastic gear data as a 9-sample problem in Example 14.3.3. We noted earlier in the section that under this model the leverages are $\frac{1}{8}$ for the 8 observations at 21°C and $\frac{1}{4}$ for the other 30 observations. Thus the standardized residuals are

$$r_{4j}=\frac{y_{4j}-\bar{y}_4}{s\sqrt{1-\frac{1}{8}}}; \qquad r_{ij}=\frac{y_{ij}-\bar{y}_i}{s\sqrt{1-\frac{1}{4}}} \quad \text{for } i\neq 4$$

where $s^2=0.02066$.

In Figure 14.5.2 the 40 standardized residuals are plotted out on a line to obtain what is essentially a histogram with a large number of classes. This picture shows reasonable symmetry about 0, and there do not appear to be too many very large or small observations. One could pool classes, compute expected frequencies from $N(0, 1)$, and test goodness of fit as in Section 12.5. However, this test is not usually relevant because it is departures in the extreme tails of the distribution which are of primary interest.

Alternatively, we can transform the r's via the probability integral transformation $u=F(r)$, where F is the c.d.f. of $N(0, 1)$ (see Section 6.3). If the r's are observations from $N(0, 1)$, then the u's are observations from $U(0, 1)$. If

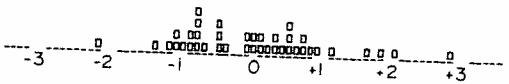

Figure 14.5.2. Histogram of standardized residuals for the plastic gear data (k-sample model).

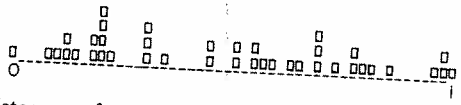

Figure 14.5.3. Histogram of transformed residuals $u=F(r)$ for the plastic gear data (k-sample model).

the normality assumption holds, the u's should be scattered randomly and uniformly between 0 and 1. Figure 14.5.3 shows a histogram of the 40 values $u_{ij}=F(r_{ij})$ for the plastic gear example. It is perhaps easier to judge whether there is an excess of large or small values from Figure 14.5.3 than from Figure 14.5.2.

A useful exercise is to generate sets of $n=40$ random numbers on the computer and plot them as in Figure 14.5.3. In this way one can get a feeling for how the graph should look if the normality assumption is satisfied. In the present example, the observed graph does not appear unusual, and there is no evidence against the normality assumption.

The same technique can be used to check other continuous probability models by taking F to be the appropriate cumulative distribution function.

PROBLEMS FOR SECTION 14.5

1. A straight line model $\mu=\alpha+\beta x$ is fitted to n observed points (x_i, y_i) by least squares.

   (a) Show that the leverage of the ith point is

   $$m_{ii}=\frac{1}{n}+(x_i-\bar{x})^2/S_{xx}.$$

   Which of the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ have the smallest variances?

   (b) Which observations will be the most influential in determining $\hat{\beta}$?

2. Calculate leverages m_{ii} and standardized residuals r_i for the four data sets in Table 13.5.2, and plot the standardized residuals versus the fitted values. (Note that r_i is undefined for the last point in data set 4. The fitted line must go through this point, and so this point gives no information about the adequacy of the model.)

3. Calculate residuals, leverages, and standardized residuals in Problem 13.5.1(a), and plot standardized residuals versus fitted values. Note that standardization increases the relative magnitude of the residual for the observation at $x=1.5$, because this point has a high leverage.

4. Calculate leverages and standardized residuals in Problem 13.5.4(a), and plot standardized residuals versus fitted values. Comment on your findings.

5. Check the standardized residuals for normality in Problems 13.3.7(a) and (b). Does it seem more reasonable to assume normality of the log counts?

6.†(a) Let M be a symmetric idempotent $n\times n$ matrix, so that $M^t=M$ and $MM=M$. Show that each diagonal element of M must lie between 0 and 1.

(b) Show that $X(X^tX)^{-1}X^t$ is symmetric and idempotent, and hence that the leverages m_{ii} lie between 0 and 1.
(c) Show that, if the ith point has leverage $m_{ii}=1$, then $\hat{\mu}_i = y_i$ and $\hat{\varepsilon}_i = 0$.

*14.6. Derivations

In this section we derive the distribution of the residual sum of squares in the normal linear model, and also the distribution of the additional sum of squares due to a linear hypothesis. We used these results in earlier sections to set up significance tests and confidence intervals.

We assume that $Y_1, Y_2, \ldots, Y_n$ are independent $N(\mu_i, \sigma^2)$, and that $\mu = X\beta$ where X is an $n \times q$ matrix of constants with linearly independent columns.

Let $U_i = (Y_i - \mu_i)/\sigma$, so that $Y_i = \mu_i + \sigma U_i$. In matrix notation we have

$$Y = \mu + \sigma U = X\beta + \sigma U$$

where U is an $n \times 1$ vector whose components $U_1, U_2, \ldots, U_n$ are independent $N(0, 1)$.

The vector of fitted values is

$$\hat{\mu} = X\hat{\beta} = XX^LY = XX^L(X\beta + \sigma U).$$

Since $X^LX = I$, we have

$$\hat{\mu} = X\beta + \sigma MU \tag{14.6.1}$$

where $M = XX^L$. The vector of residuals is

$$\begin{aligned}\hat{\varepsilon} &= Y - \hat{\mu} = (X\beta + \sigma U) - (X\beta + \sigma MU)\\ &= \sigma(I - M)U.\end{aligned} \tag{14.6.2}$$

In proving the theorems below, we shall construct an orthogonal matrix C and then consider the orthogonal transformation $Z = C^tU$. Then $Z = (Z_i)$ is $n \times 1$, and by Theorem 7.3.1, its components $Z_1, Z_2, \ldots, Z_n$ are independent $N(0, 1)$.

The following lemma will be used in constructing the required orthogonal transformation.

Lemma 14.6.1. *Let X be an $n \times q$ matrix with linearly independent columns. Then there exists an $n \times q$ matrix P and a nonsingular $q \times q$ matrix A such that $X = PA$ and $P^tP = I$.*

PROOF. Let $X_1, X_2, \ldots, X_q$ denote the q columns of X, and let $b = (b_j)$ be a $q \times 1$ vector of constants. The product Xb is $n \times 1$ and represents a linear combination of the columns of X:

$$Xb = X_1b_1 + X_2b_2 + \cdots + X_qb_q.$$

*This section may be omitted on first reading.

The set of all such linear combinations is a vector space $\mathscr{V}(X)$ called the column space of X. Since the X_j's are assumed to be linearly independent, $\mathscr{V}(X)$ has dimension q.

Let $P_1, P_2, \ldots, P_q$ be a set of normed orthogonal basis vectors for $\mathscr{V}(X)$. One way to construct the P_j's is by applying the Gram–Schmidt orthogonalization procedure to the columns of X. Let P be the $n \times q$ matrix with $P_1, P_2, \ldots, P_q$ as its columns. Since the P_j's are normed and orthogonal we have $P_j^tP_j = 1$ and $P_i^tP_j = 0$ for $i \neq j$. It follows that $P^tP = I$.

Since $P_1, P_2, \ldots, P_q$ is a set of basis vectors for $\mathscr{V}(X)$, every vector in $\mathscr{V}(X)$ can be written as a linear combination of $P_1, P_2, \ldots, P_q$. In particular, each of the X_j's can be written as a linear combination of $P_1, P_2, \ldots, P_q$,

$$X_j = P_1a_{1j} + P_2a_{2j} + \cdots + P_qa_{qj}$$

for some constants a_{ij}. Thus we have $X = PA$. The matrix A must be nonsingular because both X and P have rank q. □

Theorem 14.6.1. *Under the normal linear model assumptions stated above, $\Sigma\hat{\varepsilon}_i^2$ is distributed independently of $\hat{\beta}$, and*

$$\frac{1}{\sigma^2}\Sigma\hat{\varepsilon}_i^2 \sim \chi^2_{(n-q)}.$$

PROOF. By the Lemma we can write $X = PA$ where P is $n \times q$ with normed orthogonal columns and A is $q \times q$ nonsingular. Let $C = (P|R)$ be an $n \times n$ orthogonal matrix whose first q columns are the columns of P. Since $C^tC = CC^t = I$, we have $P^tP = I$, $R^tR = I$, and

$$PP^t + RR^t = I.$$

Since $P^tP = I$ and A is nonsingular, we have

$$\begin{aligned}M &= XX^L = X(X^tX)^{-1}X^t\\ &= PA(A^tP^tPA)^{-1}A^tP^t\\ &= PAA^{-1}(A^t)^{-1}A^tP^t = PP^t;\\ I - M &= I - PP^t = RR^t.\end{aligned}$$

It follows from (14.6.1) and (14.6.2) that

$$\hat{\mu} = X\beta + \sigma PP^tU; \qquad \hat{\varepsilon} = \sigma RR^tU.$$

Since $R^tR = I$, the residual sum of squares is

$$\Sigma\hat{\varepsilon}_i^2 = \hat{\varepsilon}^t\hat{\varepsilon} = \sigma^2U^tRR^tRR^tU = \sigma^2U^tRR^tU.$$

Now consider the orthogonal transformation $Z = C^tU$. Then $Z_1, Z_2, \ldots, Z_n$ are independent $N(0, 1)$ variates. Note that P^tU contains $Z_1, Z_2, \ldots, Z_q$ and R^tU contains $Z_{q+1}, Z_{q+2}, \ldots, Z_n$. Since $\hat{\mu}$ is a function of P^tU and $\hat{\varepsilon}$ is a function of R^tU, it follows that $\hat{\mu}$ and $\hat{\varepsilon}$ are distributed independently. Also we

have

$$\frac{1}{\sigma^2}\Sigma\hat{\varepsilon}_i^2 = (R^tU)^t(R^tU)$$
$$= \text{sum of squares of components of } R^tU$$
$$= Z_{q+1}^2 + Z_{q+2}^2 + \cdots + Z_n^2.$$

Now (6.9.9) implies that $\Sigma\hat{\varepsilon}_i^2/\sigma^2$ is distributed as $\chi^2_{(n-q)}$.

Since $X^t\hat{\varepsilon} = 0$ by (14.2.2), it follows that

$$\hat{\beta} = (X^tX)^{-1}X^tY = (X^tX)^{-1}X^t(\hat{\mu} + \hat{\varepsilon})$$
$$= (X^tX)^{-1}X^t\hat{\mu}.$$

Thus $\hat{\beta}$ is a function of $\hat{\mu}$. Since $\hat{\mu}$ and $\hat{\varepsilon}$ are distributed independently, so are $\hat{\beta}$ and $\Sigma\hat{\varepsilon}_i^2$. □

Theorem 14.6.2. *Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent $N(\mu_i, \sigma^2)$ with $\mu = X\beta$ as above. Let H be a linear hypothesis,*

$$H: \beta = A\gamma$$

where A is $q \times p$ with linearly independent columns. Let $\Sigma\tilde{\varepsilon}_i^2$ be the residual sum of squares under H, and let $Q = \Sigma\tilde{\varepsilon}_i^2 - \Sigma\hat{\varepsilon}_i^2$ be the additional sum of squares due to H. Then, if H is true, Q is distributed independently of $\Sigma\hat{\varepsilon}_i^2$, and $Q/\sigma^2 \sim \chi^2_{(q-p)}$.

PROOF. Under H, the model becomes $\mu = W\gamma$ where $W = XA$. The columns of W are linear combinations of the columns of X, and therefore $\mathscr{V}(W)$, the column space of W, is a subspace of $\mathscr{V}(X)$.

As in the preceding theorem, we consider an orthogonal transformation $Z = C^tU$. For the first p columns of C we take a normed orthogonal basis of the vector space $\mathscr{V}(W)$. To this we add $q-p$ columns so that the first q columns of C form a normed orthogonal basis of $\mathscr{V}(X)$. This is possible because $\mathscr{V}(W)$ is a subspace of $\mathscr{V}(X)$.

The argument in the preceding proof can now be used for both the original model $\mu = X\beta$ and the hypothesized model $\mu = W\gamma$, giving

$$\Sigma\hat{\varepsilon}_i^2/\sigma^2 = Z_{q+1}^2 + Z_{q+2}^2 + \cdots + Z_n^2;$$
$$\Sigma\tilde{\varepsilon}_i^2/\sigma^2 = Z_{p+1}^2 + Z_{p+2}^2 + \cdots + Z_n^2.$$

Subtracting gives

$$Q/\sigma^2 = (\Sigma\tilde{\varepsilon}_i^2 - \Sigma\hat{\varepsilon}_i^2)/\sigma^2$$
$$= Z_{p+1}^2 + Z_{p+2}^2 + \cdots + Z_q^2.$$

Since this is the sum of squares of $q-p$ independent $N(0,1)$ variates, it follows by (6.9.9) that $Q/\sigma^2 \sim \chi^2_{(q-p)}$. Also, since $Z_{p+1}, \ldots, Z_q$ are distributed independently of $Z_{q+1}, \ldots, Z_n$, it follows that Q is distributed independently of $\Sigma\hat{\varepsilon}_i^2$. □

CHAPTER 15

Sufficient Statistics and Conditional Tests

In this chapter we discuss some general principles of statistical inference, and their applications in the construction of significance tests and confidence intervals.

An important requirement of any valid statistical inference is that it should not depend upon any features of the data which are irrelevant to the question of interest. The sufficiency principle attempts to formalize this requirement. Section 1 describes this principle and defines sufficient statistics. Some properties of sufficient statistics are derived in Section 2.

Significance levels and coverage probabilities are computed from sampling distributions in a series of imaginary repetitions of the experiment. These repetitions are purely hypothetical, and will not actually be carried out. Sections 3 and 4 are concerned with how to choose an appropriate series of repetitions for inferences about a parameter. In particular, it is argued that, when ancillary statistics are present, significance levels and coverage probabilities should be computed from a conditional distribution.

Section 5 considers difficulties which can arise in testing composite hypotheses. Sometimes a satisfactory test can be obtained by conditioning on the observed values of sufficient statistics for the unknown parameters. Some examples of conditional tests are given in Section 6.

15.1. The Sufficiency Principle

An important requirement of any valid statistical inference is that it should not be affected by features of the data which are irrelevant to the question of interest. The sufficiency principle is an attempt to formalize this requirement.

Let y, y' be two possible (mutually exclusive) outcomes of an experiment whose probability model involves an unknown parameter θ. Suppose that we wish to make inferences about the value of θ. Roughly speaking, the sufficiency principle states that, if the choice between y and y' is a purely random one not depending upon the value of θ, then inferences about θ should be the same if y is observed as they would be if y' were observed.

For instance, consider $n = 3$ Bernoulli trials, and suppose that we wish to make inferences about $\theta = P(\text{success})$. Consider the three outcomes $y = \text{SSF}$, $y' = \text{SFS}$, and $y'' = \text{FSS}$. Each of these outcomes has probability $\theta^2(1-\theta)$. No matter what the value of θ is, the three outcomes are equally probable. The choice among them is purely random, and does not depend in any way on the value of θ. The sufficiency principle states that inferences about θ should be the same no matter which of these outcomes is observed.

The conditional probability of observing outcome y given that either y or y' has occurred is

$$P(y|y \text{ or } y') = \frac{P(y;\theta)}{P(y;\theta)+P(y';\theta)} = \frac{\text{Odds}}{\text{Odds}+1} \tag{15.1.1}$$

where the ratio of probabilities,

$$\text{Odds} = P(y;\theta)/P(y';\theta), \tag{15.1.2}$$

is the fair betting odds for outcome y versus outcome y'. If the odds do not depend upon θ, then the choice between outcomes y and y' is purely random and is unrelated to the value of θ.

The sufficiency principle states that, *if the odds* (15.1.2) *do not depend upon* θ, *then outcomes* y *and* y' *should lead to the same inferences concerning* θ. An equivalent requirement is that the conditional probability (15.1.1) does not depend upon θ.

Sufficiency and the Likelihood Function

The likelihood function of θ based on outcome y is proportional to $P(y;\theta)$:

$$L(\theta; y) = k(y)\cdot P(y;\theta), \tag{15.1.3}$$

where $k(y)$ is positive and does not depend upon θ. The odds (15.1.2) are independent of θ if and only if $L(\theta; y)$ is proportional to $L(\theta; y')$. Another way of stating the sufficiency principle is that *outcomes of the same experiment which give rise to proportional likelihood functions for* θ *should lead to the same inferences about* θ. Indeed, this is the reason that the likelihood function is defined only up to a multiplicative constant, and that two likelihood functions which are proportional to one another are regarded as equivalent.

In Chapters 9–14 we restricted discussion almost exclusively to methods based on the likelihood function or likelihood ratio statistic. For these methods, observations which give rise to proportional likelihood functions will lead to the same inferences, and therefore the sufficiency principle is automatically satisfied.

Sufficient Statistics

A statistic T is a random variable whose value $T(y)$ can be computed from the data without knowledge of the value of θ. T is called a *sufficient statistic for* θ if knowledge of the observed value of T is sufficient to determine $L(\theta; y)$ up to a constant of proportionality. In other words, T is a sufficient statistic for θ if $L(\theta; y)$ can be written as a function of y only times a function of T and θ:

$$L(\theta; y) = C(y)\cdot H(T(y);\theta). \tag{15.1.4}$$

Two outcomes y, y' such that $T(y) = T(y')$ will give rise to proportional likelihood functions for θ, and by the sufficiency principle, they should lead to the same inferences concerning θ. All that we require from the data for inferences about θ is the observed value of a sufficient statistic T.

Even when θ is one dimensional, we may need two or more functions of the data to fully determine the likelihood function. If knowledge of the observed values of k statistics $T_1, T_2, \ldots, T_k$ is sufficient to determine $L(\theta; y)$ up to a proportionality constant, then $T = (T_1, T_2, \ldots, T_k)$ is called a *set of sufficient statistics*. Two outcomes y, y' such that $T_i(y) = T_i(y')$ for $i = 1, 2, \ldots, k$ will give rise to proportional likelihood functions for θ.

The existence of a set of sufficient statistics $T_1, T_2, \ldots, T_k$ enables us to condense or reduce the data to k numbers $T_1(y), T_2(y), \ldots, T_k(y)$ without losing information about θ. A set of sufficient statistics which gives the greatest possible reduction of the data is called *minimally sufficient for* θ. If T is minimally sufficient for θ, then $T(y) = T(y')$ if and only if $L(\theta; y)$ and $L(\theta; y')$ are proportional.

The sufficiency principle states that outcomes which give rise to proportional likelihood functions for θ should lead to the same inferences concerning θ. An equivalent statement of the sufficiency principle is that *outcomes which imply the same value of a minimally sufficient statistic or set of statistics* T *should lead to the same inferences concerning* θ. If T is minimally sufficient for θ, then T carries all of the relevant information for inferences about θ. Inferences about θ should depend only on T and not on the remainder of the data.

EXAMPLE 15.1.1. Consider n Bernoulli trials, and suppose that we wish to make inferences about $\theta = P(\text{success})$. An outcome of the experiment may be written as a sequence $y = (y_1, y_2, \ldots, y_n)$, where $y_i = 1$ if the ith trial produces a success and $y_i = 0$ otherwise. Since $P(y_i = 1) = \theta$ and $P(y_i = 0) = 1 - \theta$, we have

$$f(y_i) = \theta^{y_i}(1-\theta)^{1-y_i} \qquad \text{for } y_i = 0, 1.$$

Since trials are independent, the probability of outcome y is

$$P(y;\theta)=\prod_{i=1}^{n} f(y_i)=\theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i}.$$

The likelihood function is a constant times $P(y,\theta)$,

$$L(\theta;y)=k(y)\cdot\theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i} \qquad \text{for } 0<\theta<1.$$

Usually we would take $k(y)=1$ for convenience.

Let $y'=(y'_1, y'_2, \ldots, y'_n)$ be another possible outcome. Then

$$P(y;\theta)/P(y';\theta)=\theta^{\Sigma y_i-\Sigma y'_i}(1-\theta)^{\Sigma y'_i-\Sigma y_i},$$

which is independent of θ if and only if $\Sigma y_i=\Sigma y'_i$. This is also the condition under which y and y' give rise to proportional likelihood functions for θ. By the sufficiency principle, outcomes y, y' such that $\Sigma y_i=\Sigma y'_i$ should lead to the same inferences for θ.

The random variable $T\equiv\Sigma Y_i$ is a sufficient statistic for θ in this example. Outcomes y, y' such that $T(y)=T(y')$ (that is, $\Sigma y_i=\Sigma y'_i$) give rise to proportional likelihood functions. In fact, T is minimally sufficient because if $T(y)\neq T(y')$, then $L(\theta;y)$ is not proportional to $L(\theta;y')$.

The sufficient statistic $T\equiv\Sigma Y_i$ carries all of the information from the data concerning the value of θ. The remainder of the data (i.e. information about the order in which the 0's and 1's occurred) is not relevant to inferences about θ under the model assumed. This additional information is what would be used to check the assumptions of independent trials and equal success probabilities which underly the Bernoulli trials model.

EXAMPLE 15.1.2. Let $Y_1, Y_2, \ldots, Y_n$ be independent Poisson variates with the same mean μ. Then the probability of outcome $y=(y_1, y_2, \ldots, y_n)$ is

$$P(y;\mu)=\prod_{i=1}^{n}\mu^{y_i}e^{-\mu}/y_i!=\mu^{\Sigma y_i}e^{-n\mu}/(y_1!\,y_2!\ldots y_n!)$$

where $y_i=0, 1, 2, \ldots$. The likelihood function of μ is

$$L(\mu;y)=C(y)\cdot\mu^{t}e^{-n\mu} \qquad \text{for } \mu>0,$$

where $t=\Sigma y_i$. The variate $T\equiv\Sigma Y_i$ is a sufficient statistic for μ. In fact, T is minimally sufficient because $L(\mu;y')$ is not proportional to $L(\mu;y)$ unless $\Sigma y'_i=\Sigma y_i$.

Under the model assumed, all of the information relevant to inferences about μ is carried by the sufficient statistic $T\equiv\Sigma y_i$. We can replace the n-dimensional observation vector y by the single number $t=\Sigma y_i$ without losing information about μ. The individual y_i's are not needed for inferences about μ, although they would be required if we wished to check the assumptions of the model.

In both this example and the preceding one, the sample size n is regarded as fixed and known in advance. For this reason we have not included n in the

sufficient statistic, although its value would be required for inferences about the parameter.

EXAMPLE 15.1.3(a). Let $Y_1, Y_2, \ldots, Y_n$ be independent exponential variates with the same mean θ. Their joint p.d.f. is

$$f(y_1, y_2, \ldots, y_n)=\prod_{i=1}^{n}\frac{1}{\theta}e^{-y_i/\theta}=\theta^{-n}e^{-\Sigma y_i/\theta}$$

for $0<y_i<\infty$. If the measurement intervals are small (see Section 9.4), the likelihood function of θ is

$$L(\theta;y)=C(y)\cdot\theta^{-n}e^{-t/\theta} \qquad \text{for } \theta>0$$

where $t=\Sigma y_i$. Assuming n to be known in advance, the total $T\equiv\Sigma Y_i$ is a sufficient statistic for θ.

(b). A more complicated situation was considered in Section 9.5. The lifetimes of n specimens were assumed to be independent exponential variates, but censoring of lifetimes at predetermined times was permitted. The likelihood function then has the form

$$L(\theta)=C\cdot\theta^{-m}e^{-s/\theta} \qquad \text{for } \theta>0$$

where m is the number of specimens which fail, and s is the sum of m failure times and $n-m$ censoring times. We would not know m or s until after the experiment. Thus, in this case, we need the observed values of two statistics, M (the number of failures) and S (the total time on test), before we can write down $L(\theta)$. Under the exponential model with censoring, the pair (M, S) is minimally sufficient. Neither M nor S by itself is a sufficient statistic for θ.

EXAMPLE 15.1.4. Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent variates having a uniform distribution on the interval $[0, \theta]$ where $\theta>0$. From Problem 9.4.11, the likelihood function of θ is

$$L(\theta;y)=\begin{cases}C(y)\theta^{-n} & \text{for } \theta\geq y_{(n)};\\ 0 & \text{otherwise}\end{cases}$$

where $y_{(n)}$ is the largest sample value. Two samples y, y' will give rise to proportional likelihood functions for θ if and only if $y_{(n)}=y'_{(n)}$, so that the range of $L(\theta)$ is the same for both samples. Hence $Y_{(n)}$ is a minimally sufficient statistic for θ.

Special care is required in examples like this where the range of y depends upon θ. A set of sufficient statistics must determine not only the functional form of $L(\theta;y)$, but also the range of possible values for θ.

EXAMPLE 15.1.5. Let $Y_1, Y_2, \ldots, Y_n$ be independent variates having a Cauchy distribution centred at θ. The p.d.f. of this distribution is

$$f(x)=\frac{1}{\pi[1+(x-\theta)^2]} \qquad \text{for } -\infty<x<\infty.$$

The joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$ is

$$f(y_1, y_2, \ldots, y_n) = \pi^{-n} \prod_{i=1}^{n} [1 + (y_i - \theta)^2]^{-1}$$

and the likelihood function of θ is

$$L(\theta; y) = C(y) \cdot \prod_{i=1}^{n} [1 + (y_i - \theta)^2]^{-1} \qquad \text{for } -\infty < \theta < \infty.$$

This is more complicated than the likelihood functions in the preceding examples. We cannot find one or two statistics which determine the likelihood function in this case.

Note that reordering the y_i's will not change the likelihood function. Thus, by the sufficiency principle, inferences concerning θ should not depend upon the order in which the y_i's were recorded. Let $y_{(i)}$ denote the ith smallest of the y_i's, so that $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$. Then

$$L(\theta; y) = C(y) \cdot \prod_{i=1}^{n} [1 + (y_{(i)} - \theta)^2]^{-1} \qquad \text{for } -\infty < \theta < \infty,$$

and $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ is a set of sufficient statistics for θ. It can be shown that this set of statistics is in fact minimally sufficient. The best that we can do in this example is an n-dimensional set of sufficient statistics.

In general, when y consists of the observed values of n independent and identically distributed random variables, the order in which the observations were recorded is irrelevant for inferences about θ. Thus $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ is a set of sufficient statistics for θ. Whether or not these statistics are minimally sufficient depends upon the form of the distribution assumed. In the preceding four examples it was possible to further reduce the data to only one or two statistics without losing information about θ, but this is not possible when a Cauchy distribution is assumed.

EXAMPLE 15.1.6 (The Exponential Family). Suppose that Y is a variate whose p.f. or p.d.f. depends upon a single parameter θ, and is of the form

$$f(y; \theta) = A(\theta) \cdot B(y) \cdot e^{c(\theta) \cdot d(y)} \qquad \text{for } -\infty < y < \infty$$

where A, B, c, and d are known functions. The distribution of Y is then said to belong to the *exponential family of distributions*. Note that the range of Y does not depend upon θ.

Several of the one-parameter distributions which we have considered are members of the exponential family. For example, the binomial p.f. can be written

$$f(y; \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} = (1-\theta)^n \binom{n}{y} e^{y \log\{\theta/(1-\theta)\}},$$

which is of the exponential form with

$$A(\theta) = (1-\theta)^n; \; c(\theta) = \log \frac{\theta}{1-\theta}; \; d(y) = y;$$

$B(y) = \binom{n}{y}$ for $y = 0, 1, \ldots, n$ and $B(y) = 0$ otherwise.

Similarly, the Poisson, exponential, and χ^2 distributions are members of the exponential family.

If $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed variates whose distribution belongs to the exponential family, their joint p.f. or p.d.f. is

$$f(y_1, y_2, \ldots, y_n) = [A(\theta)]^n \left[\prod_{i=1}^{n} B(y_i) \right] \exp\left\{ c(\theta) \cdot \sum_{i=1}^{n} d(y_i) \right\}.$$

The likelihood function is then

$$L(\theta) = k(y) \cdot [A(\theta)]^n \exp\{c(\theta) \cdot \Sigma d(y_i)\}.$$

Since the range of the Y_i's does not depend upon θ, the set of possible values for θ does not depend upon the data. Hence the statistic $T \equiv \Sigma d(y_i)$ is minimally sufficient for θ. Because of this, statistical inference is more straightforward for distributions belonging to the exponential family.

The definition of the exponential family can be extended to include distributions which depend upon several parameters $\theta_1, \theta_2, \ldots, \theta_r$. Details may be found in Chapter 2 of *Theoretical Statistics* by D.R. Cox and D.V. Hinkley.

EXAMPLE 15.1.7 (Normal Linear Model). Let $Y_1, Y_2, \ldots, Y_n$ be independent $N(\mu_i, \sigma^2)$ variates with $\mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iq}\beta_q$, where the x_{ij}'s are known constants and the β_j's are unknown parameters. We shall show that the parameter estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q$ and residual sum of squares $\Sigma\hat{\varepsilon}_i^2$ form a set of sufficient statistics for the unknown parameters $\beta_1, \beta_2, \ldots, \beta_q$ and σ.

From Section 13.2, the log likelihood function is

$$l(\beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \Sigma (y_i - \mu_i)^2.$$

Using matrix notation as in Sections 14.1 and 14.2, we have

$$\Sigma (y_i - \mu_i)^2 = (y - \mu)^t (y - \mu).$$

Now since $\mu = X\beta$, $\hat{\mu} = X\hat{\beta}$, and $\hat{\varepsilon} = y - \hat{\mu}$, we have

$$y - \mu = y - \hat{\mu} + \hat{\mu} - \mu = \hat{\varepsilon} + X(\hat{\beta} - \beta)$$

and therefore

$$\begin{aligned} \Sigma (y_i - \mu_i)^2 &= [\hat{\varepsilon} + X(\hat{\beta} - \beta)]^t [\hat{\varepsilon} + X(\hat{\beta} - \beta)] \\ &= \hat{\varepsilon}^t \hat{\varepsilon} + \hat{\varepsilon}^t X(\hat{\beta} - \beta) + (\hat{\beta} - \beta)^t X^t \hat{\varepsilon} + (\hat{\beta} - \beta)^t X^t X (\hat{\beta} - \beta). \end{aligned}$$

Since $X^t\hat{\varepsilon} = 0$ by (14.2.2), we have $\hat{\varepsilon}^t X = (X^t\hat{\varepsilon})^t = 0$. Thus both cross-product terms are zero, and

$$\Sigma (y_i - \mu_i)^2 = \Sigma \hat{\varepsilon}_i^2 + (\hat{\beta} - \beta)^t X^t X (\hat{\beta} - \beta).$$

It follows that

$$l(\beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} [\Sigma \hat{\varepsilon}_i^2 + (\hat{\beta} - \beta)^t X^t X (\hat{\beta} - \beta)].$$

Two samples y, y' for which $\hat{\beta}$ and $\Sigma\hat{\varepsilon}_i^2$ are the same will give rise to the same log likelihood function for β and σ. Therefore $\hat{\beta}$ and $\Sigma\hat{\varepsilon}_i^2$ form a set of sufficient statistics for the unknown parameters.

In the above argument, the x_{ij}'s and n are treated as constants whose values are known prior to the experiment, and the vector of y_i's is the experimental outcome. The only functions of the y_i's which we require for inferences about β and σ are $\hat{\beta}$ and $\Sigma\hat{\varepsilon}_i^2$. We would also need to know n and X^tX, but these are not included as part of the set of sufficient statistics.

Problems for Section 15.1

1. Show that $T \equiv Y_1 + Y_2 + \cdots + Y_n$ is a sufficient statistic for λ in Problem 9.2.2(a), and find the probability distribution of T.

2.† Suppose that we observe a single measurement Y from $N(0, \sigma^2)$. Is Y a sufficient statistic for σ? Is Y minimally sufficient?

3. Bacteria are distributed randomly and uniformly throughout river water at the rate of λ bacteria per unit volume. n test tubes containing volumes $v_1, v_2, \ldots, v_n$ of river water are prepared.
   (a) Suppose that the number of bacteria in each of the n test tubes is determined. Find a sufficient statistic for λ.
   (b) Suppose that the n samples are combined to give a single sample of volume $v = \Sigma v_i$, and the total number of organisms is determined. Find a sufficient statistic for λ. Does combining the samples result in a loss of information concerning λ?

4. Show that $\bar{X}$ is a sufficient statistic for μ in Problem 9.1.13.

5.† Suppose that Y has a binomial (n, θ) distribution where n is known and θ is unknown. Is the pair of statistics $T_1 \equiv Y$, $T_2 \equiv n - Y$ minimally sufficient for θ?

6. Let $X_1, X_2, \ldots, X_n$ be independent variates having a continuous uniform distribution on the interval $(\theta, \theta + 1)$. Show that $X_{(1)}$ and $X_{(n)}$ form a pair of sufficient statistics for θ.

7.† Let $X_1, X_2, \ldots, X_n$ be independent variates having a continuous uniform distribution on the interval $(-\theta, \theta)$. Find a sufficient statistic for θ.

8. Let $Y_1, Y_2, \ldots, Y_n$ be independent $N(\mu, \sigma^2)$ random variables. Show the following:
   (a) $\bar{Y}$ is a sufficient statistic for μ when σ is known;
   (b) $\Sigma(Y_i - \mu)^2$ is a sufficient statistic for σ when μ is known;
   (c) $\bar{Y}$ and $\Sigma(Y_i - \bar{Y})^2$ form a set of sufficient statistics for μ and σ when both parameters are unknown.
   (d) ΣY_i and ΣY_i^2 also form a set of sufficient statistics for μ and σ when both parameters are unknown.

9.† A scientist makes n measurements $X_1, X_2, \ldots, X_n$ of a constant μ using an apparatus of known variance σ^2, and m additional measurements $Y_1, Y_2, \ldots, Y_m$ of μ using a second apparatus of known variance $k\sigma^2$. Assume that all measurements are independent and normally distributed. Show that $T \equiv nk\bar{X} + m\bar{Y}$ is a sufficient statistic for μ, and find its distribution.

10. Suppose that $X_1, X_2, \ldots, X_n$ are $N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m$ are $N(\mu_2, \sigma^2)$, all independent. Show that $\bar{X}$, $\bar{Y}$, and $V \equiv \Sigma(X_i - \bar{X})^2 + \Sigma(Y_i - \bar{Y})^2$ form a set of sufficient statistics for μ_1, μ_2, and σ.

11. Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent and exponentially distributed random variables, with $E(Y_i) = (\alpha + \beta x_i)^{-1}$. Here $x_1, x_2, \ldots, x_n$ are known constants, and α, β are unknown parameters. Find a pair of sufficient statistics for α and β.

12. Show that the Poisson, exponential, and χ^2 distributions are members of the exponential family.

13. Show that the normal distributions $N(0, \sigma^2)$ and $N(\mu, 1)$ are members of the exponential family.

14. Suppose that the distribution of X belongs to the exponential family, and Y is a one-to-one function of X. Show that the distribution of Y also belongs to the exponential family.

15.† Suppose that the distribution of X belongs to the exponential family. The parameter $\phi = c(\theta)$ is called the *natural parameter* of the distribution. Find the natural parameter for the binomial, Poisson, and exponential distributions.

15.2. Properties of Sufficient Statistics

In this section we discuss some properties of a sufficient statistic or set of sufficient statistics T.

Let y be a typical outcome of an experiment for which the probability model depends upon an unknown parameter θ. In the preceding section we defined T to be a sufficient statistic (or set of sufficient statistics) for θ if knowledge of $T(y)$, the observed value of T, is sufficient to determine $L(\theta; y)$ up to a proportionality constant. It follows by (15.1.3) and (15.1.4) that, if T is sufficient for θ, then

$$P(y; \theta) = c(y) \cdot H(T(y); \theta) \tag{15.2.1}$$

for all y, where $c(y)$ does not depend upon θ.

Property 1. If T is sufficient for θ, then the likelihood function for θ based on the distribution of T is proportional to $L(\theta; y)$.

This result is to be expected because T carries all of the sample information concerning θ. Thus we should get the same information about θ from observing just the value of T as we would get from the complete sample y.

To find the probability of the event $T = t$, we sum or integrate (15.2.1) over all y such that $T(y) = t$. Since the second factor on the right hand side is constant in this sum or integral, we obtain

$$P(T = t; \theta) = \left[\sum_{T(y)=t} c(y) \right] \cdot H(t; \theta) = d(t) \cdot H(t; \theta) \tag{15.2.2}$$

where $d(t)$ is not a function of θ. The likelihood function based on (15.2.2) will be the same up to a proportionality constant as that based on (15.2.1).

Property 2. If T is sufficient for θ, the conditional distribution of outcomes y given the observed value of T does not depend upon θ.

This property is often used to define sufficient statistics. It is closely related to the fact that if T is sufficient for θ, then (15.1.1) is independent of θ whenever $T(y) = T(y')$. We shall use this result in Section 15.5 to construct exact significance tests for composite hypotheses.

It follows from (3.4.1) that

$$P(Y = y|T = t) = P(Y = y, T = t)/P(T = t).$$

The numerator in this expression equals $P(Y = y)$ whenever $T(y) = t$, and it equals zero otherwise. It follows that

$$P(Y = y|T = t) = P(Y = y)/P(T = t) \qquad \text{if } T(y) = t, \tag{15.2.3}$$

and $P(Y = y|T = t) = 0$ otherwise. Thus by (15.2.1) and (15.2.2) we have

$$P(Y = y|T = t) = c(y) \Big/ \sum_{T(y)=t} c(y) = c(y)/d(t) \tag{15.2.4}$$

which does not depend upon θ.

EXAMPLE 15.2.1. Consider n Bernoulli trials with success probability θ. Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a zero-one vector indicating the observed sequence of failures and successes as in Example 15.1.1. Then

$$P(Y = y; \theta) = \theta^{\Sigma y_i}(1 - \theta)^{n - \Sigma y_i} = \theta^t(1 - \theta)^{n-t}$$

where $t = \Sigma y_i$. Here $T \equiv \Sigma Y_i$ is a sufficient statistic for θ. Since T is the total number of successes in n Bernoulli trials, it has a binomial (n, θ) distribution, and

$$P(T = t; \theta) = \binom{n}{t}\theta^t(1 - \theta)^{n-t} \qquad \text{for } t = 0, 1, \ldots, n.$$

The likelihood function based on observing $T = t$ and that based on the full sample y are both proportional to $\theta^t(1 - \theta)^{n-t}$.

The conditional probability of outcome y given that $T(y) = t$ is

$$P(Y = y)/P(T = t) = \begin{cases} 1 \Big/ \binom{n}{t} & \text{for } \Sigma y_i = t; \\ 0 & \text{otherwise.} \end{cases}$$

All $\binom{n}{t}$ possible sequences of t successes and $n - t$ failures are equally probable. This distribution does not depend upon θ, and could be used for testing the adequacy of the Bernoulli model.

EXAMPLE 15.2.2. Let $Y_1, Y_2, \ldots, Y_n$ be independent Poisson variates with the same mean μ. We showed in Example 15.1.2 that

$$P(Y = y; \mu) = \mu^{\Sigma y_i}e^{-n\mu}/y_1!\, y_2! \ldots y_n!,$$

and that $T \equiv \Sigma Y_i$ is a sufficient statistic for μ. Since T is a sum of independent Poisson variates, it has a Poisson distribution with mean $\Sigma E(Y_i) = n\mu$. See the Corollary to Example 4.6.1. Thus we have

$$P(T = t; \mu) = (n\mu)^t e^{-n\mu}/t! = \mu^t e^{-n\mu}(n^t/t!).$$

The likelihood function based on observing just the total $t = \Sigma y_i$ and that based on the full sample $y_1, y_2, \ldots, y_n$ are both proportional to $\mu^t e^{-n\mu}$.

The conditional probability of outcome y given that $T(y) = t$ is

$$P(Y = y)/P(T = t) = \frac{t!}{y_1!\, y_2! \ldots y_n!}\left(\frac{1}{n}\right)^t \qquad \text{for } \Sigma y_i = t.$$

This is a multinomial distribution with index t and equal probability parameters $p_1 = p_2 = \cdots = p_n = 1/n$. As expected, the conditional distribution of outcomes given the sufficient statistic does not depend upon the parameter μ.

Property 3. Applying a one-to-one transformation to a set of sufficient statistics produces another set of sufficient statistics.

Let $T_1, T_2, \ldots, T_k$ be a set of sufficient statistics for θ. Suppose that $U_1, U_2, \ldots, U_k$ are functions of $T_1, T_2, \ldots, T_k$, and that the transformation from $(T_1, T_2, \ldots, T_k)$ to $(U_1, U_2, \ldots, U_k)$ is invertible. Since the U_i's and T_i's can be deduced from one another, they have the same information content. Given the values of the U_i's, we can calculate the values of the T_i's and hence determine $L(\theta; y)$. Thus the U_i's also form a set of sufficient statistics for θ.

EXAMPLE 15.2.3 (Normal Linear Model). We showed in Example 15.1.7 that $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q$ and $\Sigma\hat{\varepsilon}_i^2$ form a set of sufficient statistics for the parameters $\beta_1, \beta_2, \ldots, \beta_q$, and σ in the normal linear model. One can show, by a similar argument to that in Example 15.1.7, that

$$\Sigma\hat{\varepsilon}_i^2 = \Sigma Y_i^2 - \hat{\beta}^t X^t X\hat{\beta}.$$

Note also that

$$\hat{\beta} = (X^tX)^{-1}X^tY = (X^tX)^{-1}T$$

where $T = X^tY$ is $q \times 1$ with jth component

$$T_j = \sum_{i=1}^{n} x_{ij}Y_i.$$

As in Example 15.1.7, we treat X as a matrix of constants which, like n, is known prior to the experiment.

Given observed values of T and ΣY_i^2, we can calculate $\hat{\beta}$ and $\Sigma\hat{\varepsilon}_i^2$. Conversely, T and ΣY_i^2 can be computed from $\hat{\beta}$ and $\Sigma\hat{\varepsilon}_i^2$. Thus T and ΣY_i^2 have the same information content as $\hat{\beta}$ and $\Sigma\hat{\varepsilon}_i^2$. Since $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q$, and $\Sigma\hat{\varepsilon}_i^2$ form a set of sufficient statistics, the same is true of $T_1, T_2, \ldots, T_q$ and ΣY_i^2.

Property 4. The maximum likelihood estimate $\hat{\theta}$ is part of any set of sufficient statistics $T_1, T_2, \ldots, T_k$ in the sense that its value can be computed from just the T_i's. This follows because the T_i's determine $L(\theta; y)$ up to a proportionality constant, and $\hat{\theta}$ does not depend upon this constant.

In some simple examples, $\hat{\theta}$ is itself a sufficient statistic which carries all of the information relevant to estimation of θ. For instance, in Example 15.1.1 we have $\hat{\theta} \equiv T/n$ where $T \equiv \Sigma Y_i$ is a sufficient statistic. Since n is given, the values of $\hat{\theta}$ and T can be deduced from one another, and they have the same information content. Similar comments apply in Examples 2, 3(a), 4, 6, and 7 of Section 15.1.

In more complex examples, $\hat{\theta}$ is not by itself a sufficient statistic for θ. For instance, in Example 15.1.3(b) we have a pair of sufficient statistics (S, M), and $\hat{\theta} \equiv S/M$. The likelihood function is

$$L(\theta) = c\theta^{-m}e^{-s/\theta} = c\theta^{-m}e^{-m\hat{\theta}/\theta} \qquad \text{for } \theta > 0.$$

Knowledge of $\hat{\theta}$ alone is not enough to determine $L(\theta)$ up to a constant of proportionality. We also need to know the observed value of M, the number of failures out of the n specimens tested. Similarly, in Example 15.1.5, knowledge of $\hat{\theta}$ alone is not sufficient to determine $L(\theta; y)$. In this case, we need $n-1$ additional statistics for a sufficient set.

Under suitable regularity conditions (see Section 9.7) we can expand the log relative likelihood function in a Taylor's series about $\theta = \hat{\theta}$:

$$r(\theta) = -\tfrac{1}{2}(\theta-\hat{\theta})^2\mathcal{I}(\hat{\theta}) + \frac{(\theta-\hat{\theta})^3}{3!}l^{(3)}(\hat{\theta}) + \frac{(\theta-\hat{\theta})^4}{4!}l^{(4)}(\hat{\theta}) + \cdots.$$

Here $l^{(i)}(\hat{\theta})$ denotes the ith derivative of $l(\theta)$ with respect to θ, evaluated at $\theta = \hat{\theta}$. In general, $\mathcal{I}(\hat{\theta})$ and $l^{(i)}(\hat{\theta})$ will depend not only on $\hat{\theta}$ but also on other functions of the data.

Since the log relative likelihood function is determined completely by

$$\hat{\theta}, \mathcal{I}(\hat{\theta}), l^{(3)}(\hat{\theta}), l^{(4)}(\hat{\theta}), \ldots,$$

it follows that this set of statistics is sufficient for θ. We can think of $\hat{\theta}$ as the primary source of information concerning θ, while the remaining statistics in the list supply supplementary information. The information $\mathcal{I}(\hat{\theta})$ indicates the precision of the experiment with respect to $\hat{\theta}$, since if $\mathcal{I}(\hat{\theta})$ is large, then likelihood intervals or approximate confidence intervals for θ will be narrow. The remaining statistics $l^{(3)}(\hat{\theta}), l^{(4)}(\hat{\theta}), \ldots$ give information about the shape of the likelihood function, and hence indicate the appropriate form for likelihood and confidence regions.

In Section 11.3 we noted that, in large samples, the cubic and higher terms

in the series expansion will be negligible with high probability. Thus we have $r(\theta) \approx -\frac{1}{2}(\theta-\hat{\theta})^2\mathcal{I}(\hat{\theta})$ in large samples, so that $\hat{\theta}$ and $\mathcal{I}(\hat{\theta})$ form a set of approximate sufficient statistics for θ. In this case we can summarize nearly all of the information concerning θ by giving the most likely value $\hat{\theta}$ and a measure of precision $\mathcal{I}(\hat{\theta})$.

Problems for Section 15.2

1. Suppose that X has a binomial distribution with parameters (n, θ), and that Y is independent of X and has a binomial distribution with parameters (m, θ), where m and n are known. Show that $T \equiv X + Y$ is a sufficient statistic for θ, and verify that the conditional distribution of X and Y given T does not depend upon θ.

2. Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables, with

$$P(Y_i = y) = \frac{1}{N} \qquad \text{for } y = 1, 2, \ldots, N,$$

where N is an unknown positive integer.

(a) Show that $Y_{(n)}$ is a sufficient statistic for N, and derive its probability function.
(b) Derive the conditional probability function of $Y_1, Y_2, \ldots, Y_n$ given $Y_{(n)}$.

3.† A manufacturing process produces fibers of varying lengths. The length X of such a fiber is assumed to be a continuous variate with p.d.f.

$$f(x) = 2\lambda x e^{-\lambda x^2} \qquad \text{for } x > 0$$

where $\lambda > 0$. Suppose that n fibers are selected at random and their lengths $X_1, X_2, \ldots, X_n$ are determined.

(a) Show that $T \equiv \Sigma X_i^2$ is a sufficient statistic for λ.
(b) Show that $2\lambda X^2$ has a χ^2 distribution with 2 degrees of freedom, and hence that $2\lambda T \sim \chi^2_{(2n)}$. Find the p.d.f. of T, and show that it gives rise to the same likelihood function for λ as the original sample.

4. Show that $X_{(1)}$ is a sufficient statistic for c in Problem 9.4.10. Derive the probability density function of $X_{(1)}$, and verify that the likelihood function of c is proportional to this p.d.f.

5.† Consider Problems 1 through 11 of Section 15.1. In which of these problems are the maximum likelihood estimates *not* sufficient statistics?

15.3. Exact Significance Levels and Coverage Probabilities

Consider an experiment for which the probability model depends upon an unknown parameter θ. To test the simple hypothesis $H: \theta = \theta_0$, we define a test statistic D such that large values of D indicate evidence against H. The

significance level is then

$$\mathrm{SL} = P\{D \geq D_{\mathrm{obs}}\}$$

where D_{obs} is the observed value of D. This probability is calculated under the assumption that $\theta = \theta_0$.

Similar calculations are required to find coverage probabilities. Suppose that we have a procedure for constructing a region R within which the true parameter value is thought to lie. For instance, R might be the 10% likelihood region or significance region for θ. Then the coverage probability of the region at $\theta = \theta_0$ is

$$\mathrm{CP}(\theta_0) = P\{\theta_0 \in R\}.$$

This probability is again calculated under the assumption that $\theta = \theta_0$.

A difficult problem, which we have ignored until now, is how to choose the probability distribution from which SL and CP should be calculated. In this section we consider two special cases in which it is clear how this distribution should be chosen. Some more general discussion of the problem will be given in the next section.

Case 1. $\hat{\theta}$ a Sufficient Statistic for θ

In Section 15.2 we noted that the MLE $\hat{\theta}$ is itself a sufficient statistic for θ in some simple examples. Then, by the sufficiency principle, inferences about θ should depend only on $\hat{\theta}$. Thus any valid test statistic D or region R will depend upon the data only through the value of $\hat{\theta}$. Significance levels and coverage probabilities can therefore be found from the (marginal) distribution of $\hat{\theta}$. We find the range of $\hat{\theta}$-values for which $D \geq D_{\mathrm{obs}}$, or for which $\theta_0 \in R$, and then sum or integrate the distribution of $\hat{\theta}$ over this range to obtain SL or CP. Equivalently, we can work with the distribution of any convenient one-to-one function of $\hat{\theta}$, which will also be a sufficient statistic for θ.

EXAMPLE 15.3.1. Suppose that $X_1, X_2, \ldots, X_n$ are independent exponential variates with the same mean θ. From Example 11.3.3, the log RLF of θ is

$$r(\theta) = -n\left[\frac{\hat{\theta}}{\theta} - 1 - \log\frac{\hat{\theta}}{\theta}\right] \quad \text{for } \theta > 0.$$

Here $\hat{\theta} \equiv \Sigma X_i/n$ is a sufficient statistic for θ. Any reasonable method for testing $H\colon \theta = \theta_0$ or for generating regions of plausible parameter values will depend only on $\hat{\theta}$. Significance levels and coverage probabilities can thus be calculated from the distribution of $\hat{\theta}$ when $\theta = \theta_0$. Equivalently, we can use the distribution of $U \equiv 2n\hat{\theta}/\theta_0$ which is also sufficient for θ, and has a $\chi^2_{(2n)}$ distribution when $\theta = \theta_0$ (see Problem 6.9.7). We used this result to find coverage probabilities of likelihood regions in Example 11.3.3.

Suppose, for instance, that $n = 10$, $\hat{\theta} = 3$, and that we wish to test $H\colon \theta = 2$.

For any reasonable choice of the test statistic D, the significance level can be found from the fact that $U \equiv 10\hat{\theta}$ has a $\chi^2_{(20)}$ distribution when $\theta = 2$. In particular, the likelihood ratio statistic for testing $H\colon \theta = \theta_0$ is

$$D \equiv -2r(\theta_0) \equiv 2n\left[\frac{\hat{\theta}}{\theta_0} - 1 - \log\frac{\hat{\theta}}{\theta_0}\right] \equiv 2n\left[\frac{U}{2n} - 1 - \log\frac{U}{2n}\right].$$

The observed value of U is $10\hat{\theta} = 30$, and thus

$$D_{\mathrm{obs}} = 20\left[\frac{30}{20} - 1 - \log\frac{30}{20}\right] = 1.8907.$$

By Newton's method, we find that $D \geq D_{\mathrm{obs}}$ for $U \geq 30$ and for $U \leq 12.516$. Hence the exact significance level is

$$\begin{aligned}\mathrm{SL} &= P\{U \leq 12.516\} + P\{U \geq 30\} = 1 - P\{12.516 < U < 30\} \\ &= 1 - P\{12.516 < \chi^2_{(20)} < 30\} = 0.1726.\end{aligned}$$

For comparison, the large-sample result is

$$\mathrm{SL} \approx P\{\chi^2_{(1)} \geq 1.8907\} = 0.1691,$$

which agrees closely with the exact result.

Case 2. Ancillary Statistics

Except in some simple examples, $\hat{\theta}$ will not itself be a sufficient statistic. In general, it will be necessary to supplement $\hat{\theta}$ with one or more additional statistics T in order to obtain a set of sufficient statistics $(\hat{\theta}, T)$ for θ. The definition of exact significance levels and coverage probabilities will depend upon the sort of information carried by the supplementary statistics T.

Because of Property 3 in Section 15.2, there will be many different ways to select T. Suppose that it is possible to find a statistic or set of statistics T such that

(i) $(\hat{\theta}, T)$ is minimally sufficient for θ; and
(ii) the distribution of T does not depend upon θ.

Then T is called an *ancillary statistic* (or set of ancillary statistics) for θ. We shall argue that, in this situation, exact significance levels and coverage probabilities should be computed from the conditional distribution of $\hat{\theta}$ given T.

An ancillary statistic T gives no direct information about the value of θ because its marginal distribution $f_2(t)$ does not depend upon θ. Observing just the value of T would therefore tell us nothing about the value of θ. The primary information about θ is carried by $\hat{\theta}$, with T providing supplementary or ancillary information.

Let $f(\hat{\theta}, t)$ denote the joint p.f. or p.d.f. of $\hat{\theta}$ and T. We can write $f(\hat{\theta}, t)$ as a

product,

$$f(\hat{\theta}, t) = f_1(\hat{\theta}|t) f_2(t),$$

where the second factor does not depend upon θ. Since $(\hat{\theta}, T)$ is a set of sufficient statistics, $L(\theta)$ is proportional to $f(\hat{\theta}, t)$ (see Property 1 in Section 15.2). Since $f_2(t)$ does not depend upon θ, it follows that $L(\theta)$ is proportional to $f_1(\hat{\theta}|t)$, the conditional p.f. or p.d.f. of $\hat{\theta}$ given the observed value of T. This conditional distribution carries all of the information concerning θ. The marginal distribution of T does not depend upon θ, and so it is not used in making inferences about θ.

Often one can interpret an ancillary statistic T as a measure of the precision with which it is possible to estimate θ. The various possible outcomes of an experiment may differ greatly in the amount of information about θ which they are capable of yielding. If we are fortunate, we observe an outcome which permits the value of θ to be determined quite precisely. If we are unlucky, we may obtain an uninformative outcome from which we can learn relatively little about θ.

In problems of inference, it is necessary to take into account the informativeness of the data actually obtained. The fact that we might obtain a more informative or less informative result if the experiment were repeated should be considered in designing future experiments, but it is irrelevant to the interpretation of the data at hand. The observed value of the ancillary statistic indicates the informativeness of the data actually obtained. It is therefore appropriate to base inferences about θ on the conditional distribution of $\hat{\theta}$ given the observed value of the ancillary statistic T.

EXAMPLE 15.3.2 (Random Sample Size). Suppose that the experiment involves n Bernoulli trials as in Example 15.1.1. For instance, we might examine n subjects for tuberculosis with the intention of estimating θ, the proportion of the population having this disease. In Example 15.1.1 we assumed that n, the sample size, was fixed and known prior to the experiment. However it may be that n itself is subject to variation, and could be modelled as an observed value of a random variable N. For instance, the sample size might depend upon the amount of money and laboratory space, and the number of personnel available for the study, and perhaps none of these is under the strict control of the experimenter. Or perhaps unforeseen circumstances unrelated to the incidence of tuberculosis could cause the experiment to be terminated after 150 people have been examined, although it was originally planned to examine 200.

Suppose, then, that the sample size N is a random variable with probability function $g(n)$ not depending upon θ. The experiment produces n, the observed value of N, and a sequence $y = (y_1, y_2, \ldots, y_n)$ where $y_i = 1$ or 0 according to whether the ith subject does or does not have tuberculosis. Given n, the probability of the sequence y is

$$P(y|n) = \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i}$$

as in Example 15.1.1. Hence the joint probability of y and n is

$$P(y, n) = \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i} g(n).$$

The likelihood function of θ is then

$$L(\theta) = \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i} = \theta^{n\hat{\theta}}(1-\theta)^{n(1-\hat{\theta})} \qquad \text{for } 0 < \theta < 1$$

where $\hat{\theta} = \Sigma y_i / n$.

It follows that $(\hat{\theta}, N)$ is a pair of minimally sufficient statistics for θ, and that N is ancillary. Hence inferences about θ will be based on the conditional distribution of $\hat{\theta}$ given the observed value of N. Equivalently, inferences about θ may be based on the conditional (binomial) distribution of $n\hat{\theta}$ given the observed sample size n. Coverage probabilities and significance levels will thus be calculated as in Example 11.2.1 and 12.2.1. Although the sample size N is random, we take it as fixed in inferences about θ. The fact that we might get a different sample size in repetitions of the experiment is irrelevant to the interpretation of the data actually obtained.

In the above discussion, we assumed that the distribution of N did not depend upon θ. Of course, if the distribution of N depended upon θ, conditioning on the observed value of N would entail a loss of information concerning θ, because $L(\theta)$ would no longer be proportional to $f(\hat{\theta}|n)$. For instance, one might decide to keep examining subjects until three with tuberculosis had been found and then stop. Then the distribution of N would depend upon θ, and it would not be appropriate to condition upon its observed value.

EXAMPLE 15.3.3. A total of n clouds are to be observed in an experiment to determine the effectiveness of cloud seeding in producing rain. For each cloud it is decided whether or not to seed by flipping a balanced coin. Hence Z, the number of clouds to be seeded, has a binomial $(n, \frac{1}{2})$ distribution.

Let X be the number of seeded clouds which produce rain, and let Y be the number of unseeded clouds which produce rain. We assume that clouds are independent, and that the probability of rain is p_1 for a seeded cloud and p_2 for an unseeded cloud. Then, given that z clouds are seeded, X has a binomial (z, p_1) distribution, and Y has a binomial $(n-z, p_2)$ distribution independently of X. We observe (x, y, z) and wish to make inferences about p_1 and p_2. In particular, we might want to test the hypothesis that $p_1 = p_2$.

The joint probability function of X, Y, and Z is

$$\begin{aligned} f(x, y, z) &= f(x, y|z) f(z) = f(x|z) f(y|z) f(z) \\ &= \binom{z}{x} p_1^x (1-p_1)^{z-x} \binom{n-z}{y} p_2^y (1-p_2)^{n-z-y} \binom{n}{z} \left(\frac{1}{2}\right)^n. \end{aligned}$$

The likelihood function of p_1 and p_2 is thus

$$\begin{aligned} L(p_1, p_2) &= p_1^x (1-p_1)^{z-x} p_2^y (1-p_2)^{n-z-y} \\ &= p_1^{z\hat{p}_1} (1-p_1)^{z(1-\hat{p}_1)} p_2^{(n-z)\hat{p}_2} (1-p_2)^{(n-z)(1-\hat{p}_2)} \end{aligned}$$

where $\hat{p}_1 = x/z$ and $\hat{p}_2 = y/(n-z)$. Here $\hat{p}_1$, $\hat{p}_2$, and Z are jointly minimally sufficient for p_1 and p_2, and Z is ancillary. Inferences about p_1 and p_2 will therefore be based on the conditional distribution of $\hat{p}_1$ and $\hat{p}_2$ given the observed value of Z, or equivalently, on the conditional distribution of X and Y given the observed value of Z. Since z is to be treated as fixed, a test of $H: p_1 = p_2$ can be carried out as in Example 12.4.1, with $n_1 = z$ and $n_2 = n - z$. See Section 15.6 for discussion of an exact test of this hypothesis.

The relationship between the value of Z and precision is easily seen in this example. If one should get $z = 0$ (improbable but still possible), then one would not seed any clouds, and thus would obtain no information about p_1. Similarly, with $z = n$, one would obtain no information about p_2. In both cases, the experiment would be incapable of giving evidence against the hypothesis $p_1 = p_2$. However, if one obtained $z \approx n/2$, the experiment would give a reasonable amount of information about p_1 and p_2, and hence would be capable of showing that they are different. The observed value of Z thus indicates the precision which is possible in inferences about p_1 and p_2. Although Z is a random variable, we regard Z as fixed at its observed value in the analysis.

In this case, the existence of the ancillary statistic Z shows up a defect in the design of the experiment. It would be better to set up the experiment so that the value of Z was fixed in advance near $n/2$. This could be done by drawing balls at random without replacement from an urn containing $n/2$ white balls and $n/2$ black balls, and seeding a cloud if a white ball is drawn.

EXAMPLE 15.3.4 (Cauchy Distribution). Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent variates having a Cauchy distribution centered at θ. From Example 15.1.5, the complete set of order statistics $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$ is minimally sufficient for θ.

In this example it is possible to find a set of $n-1$ ancillary statistics. To see this we note that the distribution of $U_i \equiv Y_i - \theta$ does not depend upon θ. In fact, U_i has a Cauchy distribution centred at zero, with p.d.f.

$$f(u) = \frac{1}{\pi}(1+u^2)^{-1} \qquad \text{for } -\infty < u < \infty.$$

Now consider the $n-1$ statistics

$$A_i \equiv Y_{(i+1)} - Y_{(i)} \qquad \text{for } i = 1, 2, \ldots, n-1.$$

Since $Y_{(i)} \equiv \theta + U_{(i)}$, we have

$$A_i \equiv [\theta + U_{(i+1)}] - [\theta + U_{(i)}] \equiv U_{(i+1)} - U_{(i)}.$$

The distribution of the U_i's does not depend upon θ, and so neither does the distribution of $A_1, A_2, \ldots, A_{n-1}$.

Now let T be any statistic such that the transformation from $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ to $T, A_1, A_2, \ldots, A_{n-1}$ is one-to-one. For instance, we could take $T \equiv Y_{(i)}$ for any i, or $T \equiv \bar{Y}$, or $T \equiv \hat{\theta}$. Then $(T, A_1, A_2, \ldots, A_{n-1})$ is a set of minimally

sufficient statistics for θ, and $A_1, A_2, \ldots, A_{n-1}$ are ancillary. All of the information about θ is carried by the conditional distribution of T given the observed values of the ancillary statistics. This distribution, which may be found by numerical integration, would be used for calculating exact significance levels or coverage probabilities.

In this example, the ancillary statistics give information about the shape of the likelihood function. For instance if $n = 2$, $L(\theta)$ has a unique maximum at $\bar{y}$ when a_1 is small, but is bimodal with a relative minimum at $\bar{y}$ when a_1 is large. The observed value of A_1 indicates the shape of $L(\theta)$, and hence the appropriate form of likelihood and confidence regions. However A_1 itself tells us nothing about the magnitude of θ.

PROBLEMS FOR SECTION 15.3

1.† Suppose that patients arrive for treatment according to a Poisson process in time, with 20 arrivals per year on average. The treatment is successful for a fraction θ of patients. Let X be the number of successful treatments and Y the number of unsuccessful treatments in a one-year period. Then X and Y are independent Poisson variates with means 20θ and $20(1-\theta)$. Find an ancillary statistic T such that $\hat{\theta}$ and T are jointly sufficient for θ, and derive the appropriate conditional distribution for inferences about θ.

2. Let $X_1, X_2, \ldots, X_n$ be independent random variables having a continuous uniform distribution on the interval $[\theta, \theta+1]$.

   (a) Show that $\hat{\theta} \equiv X_{(n)} - 1$, and that $T \equiv X_{(n)} - X_{(1)}$ is an ancillary statistic.
   (b) Show that the value of θ must lie in the interval $[\hat{\theta}, \hat{\theta}+t]$, where t is the observed value of T.
   (c) Show that the interval $[\hat{\theta}, \hat{\theta}+\frac{1}{2}]$ has (unconditional) coverage probability $1-(\frac{1}{2})^n$.
   (d) If $n = 3$, then $[\hat{\theta}, \hat{\theta}+\frac{1}{2}]$ is an 87.5% confidence interval for θ. Explain why this interval might not give a satisfactory summary of the information provided by the data concerning the value of θ.

3. Let $Y_1, Y_2, \ldots, Y_n$ be independent variates having a continuous uniform distribution on the interval $(\theta, 2\theta)$, where $\theta > 0$.

   (a) Show that $Y_{(1)}$ and $Y_{(n)}$ together are sufficient for θ, and that $\hat{\theta}$ is not a sufficient statistic.
   (b) Show that $A \equiv Y_{(n)}/Y_{(1)}$ is ancillary, and that $\hat{\theta}$ and A are jointly sufficient for θ.

4.* Let Y_1, Y_2 be independent variates having a Cauchy distribution centered at θ, and define $A_1 \equiv Y_{(2)} - Y_{(1)}$ as in Example 15.3.4.

   (a) Show that, if $A_1 \le 2$, the likelihood equation $l'(\theta) = 0$ has just one real root, and that $\hat{\theta} = \bar{y}$.
   (b) Show that, if $A_1 > 2$, the likelihood equation has three real roots, and that there is a relative minimum at $\theta = \bar{y}$.

15.4. Choosing the Reference Set

To evaluate the significance level for a test of $H: \theta = \theta_0$, it is necessary to imagine a series of repetitions of the experiment with θ fixed at θ_0. At each repetition the value of the test statistic D is to be computed and compared with D_{obs}. The significance level is the fraction of the time that D would be greater than or equal to D_{obs} in infinitely many repetitions. Coverage probabilities are dependent on a similar imaginary set of repetitions. The series of repetitions with respect to which SL and CP are defined is sometimes called the *reference set* for inferences about θ.

Even if the experiment were actually going to be repeated over and over again, care would be required in choosing the reference set for inferences about θ. The planned series of repetitions will not necessarily be the appropriate set for inferences about θ! For instance, in the cloud seeding experiment (Example 15.3.3), the number Z of clouds seeded would vary in future repetitions. However significance levels and coverage probabilities should be computed from the conditional distribution of X and Y, with the ancillary statistic Z held fixed at its observed value.

Most real experiments do not get repeated over and over again, and so the reference set (or series of repetitions) is purely hypothetical. Usually all that we have is a set of data from which we wish to extract information about θ and a description of how it was collected. It may be possible to imagine many different ways in which the experiment could be repeated. Except in some simple examples it is not obvious what set of repetitions is appropriate for inferences about θ.

Significance levels and coverage probabilities are dependent on the choice of a reference set. Since it is often unclear how the reference set should be chosen, there is an unavoidable fuzziness about the definitions of exact significance levels and coverage probabilities.

In this section we consider two examples which illustrate the dependence of SL and CP on the choice of the reference set. These examples also illustrate an important property of the likelihood ratio statistic: that its distribution is remarkably stable under different possible choices of the reference set. Thus, if likelihood ratio tests are used, it generally matters very little how the reference set is chosen. Similarly, intervals constructed from the likelihood function or from likelihood ratio tests will have practically the same coverage probability under a variety of different choices for the reference set. This is an important advantage of likelihood-based methods.

EXAMPLE 15.4.1. Suppose that $X = 15$ successes and $Y = 35$ failures are observed in successive Bernoulli trials with $P(\text{success}) = \theta$. Consider a test of $H: \theta = \theta_0$ using some test statistic $D(X, Y)$, and let $D_{\text{obs}} = D(15, 35)$ be the observed value of D. Then the significance level is the sum of the probabilities of pairs (x, y) for which $D(x, y) \geq D_{\text{obs}}$.

One could imagine repeating this experiment in many different ways, three

of which are as follows:

(1) Repeat with $X + Y$ fixed at 50.
(2) Repeat with X fixed at 15, so that Y is the number of failures before the 15th success.
(3) Repeat with Y fixed at 35, so that X is the number of successes before the 35th failure.

Under H, the probability of pair (x, y) in the three cases is

$$f_1(x, y) = \binom{x+y}{x} \theta_0^x (1-\theta_0)^y \qquad \text{for } x + y = 50;\ x = 0, 1, \ldots, 50;$$

$$f_2(x, y) = \binom{x+y-1}{x-1} \theta_0^x (1-\theta_0)^y \qquad \text{for } x = 15;\ y = 0, 1, 2, \ldots;$$

$$f_3(x, y) = \binom{x+y-1}{y-1} \theta_0^x (1-\theta_0)^y \qquad \text{for } y = 35;\ x = 0, 1, 2, \ldots .$$

We have three different reference sets depending upon what sequence of repetitions we imagine.

In case (1), we calculate SL by summing $f_1(x, y)$ over all pairs (x, y) for which $x + y = 50$ and $D(x, y) \geq D_{\text{obs}}$. In (2), we sum $f_2(x, y)$ over all (x, y) with $x = 15$ and $D(x, y) \geq D_{\text{obs}}$. And in (3) we sum $f_3(x, y)$ over all (x, y) with $y = 35$ and $D(x, y) \geq D_{\text{obs}}$. The significance level will in general be different for the three cases. Two observers who see the same sequence of 15 successes and 35 failures might therefore calculate different significance levels (or confidence intervals) because they imagine different ways in which the experiment might be repeated. And of course it is entirely possible that there is no intention of actually repeating the experiment anyway!

It is a bit unsettling that inferences should depend upon an imaginary set of future repetitions which will not actually be carried out. However, this is unavoidable if we wish to consider frequency characteristics such as significance levels and coverage probabilities. What we can do is attempt to lessen the importance of choosing the reference set by using methods closely related to the likelihood function.

In all three cases above, the log likelihood function of θ is

$$l(\theta) = x \log \theta + y \log(1-\theta) \qquad \text{for } 0 < \theta < 1,$$

and the MLE is $\hat{\theta} = x/(x+y)$. The likelihood ratio statistic for testing $H: \theta = \theta_0$ is

$$D(x, y) = -2r(\theta_0) = 2\left[x \log \frac{x}{(x+y)\theta_0} + y \log \frac{y}{(x+y)(1-\theta_0)}\right].$$

In all three situations $D \approx \chi^2_{(1)}$ if H is true, and the approximate significance level is $P\{\chi^2_{(1)} \geq D_{\text{obs}}\}$. If we are content to use this large-sample approximation, it does not matter which of the three reference sets is chosen.

Table 15.4.1. Exact Significance Levels for Three Possible Reference Sets

θ_0	Approx. SL	Exact significance levels (1)	(2)	(3)
0.15	0.0073	0.0082	0.0081	0.0077
0.16	0.0136	0.0186	0.0151	0.0226
0.17	0.0237	0.0372	0.0262	0.0238
0.18	0.0393	0.0403	0.0433	0.0489
0.19	0.0619	0.0685	0.0678	0.0871
0.20	0.0933	0.1087	0.0995	0.0904
0.40	0.1416	0.1528	0.1407	0.1560
0.42	0.0798	0.0877	0.0907	0.0879
0.44	0.0421	0.0471	0.0528	0.0489
0.46	0.0208	0.0235	0.0263	0.0243
0.48	0.0096	0.0108	0.0104	0.0121
0.50	0.0041	0.0066	0.0057	0.0049

The exact significance level in the likelihood ratio test depends on the choice of the reference set, but the dependence is slight. For instance, consider a test of H: $\theta = 0.2$, for which $D_{\text{obs}} = 2.82$ and $\text{SL} \approx P\{\chi^2_{(1)} \geq 2.82\} = 0.0933$. In (1) we find that $D(x, y) < D_{\text{obs}}$ for $6 \leq x \leq 14$, and thus

$$\text{SL}_1 = 1 - \sum_{x=6}^{14} f_1(x, 50 - x) = 0.1087.$$

In (2) we have $D < D_{\text{obs}}$ for $36 \leq y \leq 93$, and

$$\text{SL}_2 = 1 - \sum_{y=36}^{93} f_2(15, y) = 0.0995.$$

In (3) we have $D < D_{\text{obs}}$ for $4 \leq x \leq 14$, and

$$\text{SL}_3 = 1 - \sum_{x=4}^{14} f_3(x, 35) = 0.0904.$$

Similarly close agreement is found for other hypothesized values (see Table 15.4.1).

For reasons similar to those given in Example 11.2.1, the significance level is a discontinuous function of θ_0, and the discontinuities will occur at different parameter values in (1), (2), and (3). This accounts almost entirely for the differences among SL_1, SL_2, and SL_3.

When the likelihood ratio test is used, it matters very little whether (1), (2), or (3) is assumed. This will generally not be the case for other choices of the test statistic D.

EXAMPLE 15.4.2. Suppose that there are two different techniques for determining the log concentration (in standard units) of a chemical in solution. The

first technique gives a reading $X \sim N(\mu, 1)$ where μ is the true log concentration, while the second gives $X \sim N(\mu, 100)$. A solution is assigned to either the first technique or the second by flipping an unbiased coin, and a single measurement is taken. We wish to obtain a confidence interval for the true log concentration μ of this particular solution.

Define $T = 0$ if the first technique is used, and $T = 1$ otherwise. The experiment yields a pair of values (x, t). Given t, X has standard deviation 10^t, and p.d.f.

$$f(x|t) = \frac{1}{\sqrt{2\pi}\,10^t} \exp\{-\tfrac{1}{2}(x-\mu)^2/10^{2t}\} \qquad \text{for } -\infty < x < \infty$$

and the joint distribution of X and T is

$$f(x, t) = f(x|t) \cdot f_2(t) = \tfrac{1}{2} f(x|t) \qquad \text{for } -\infty < x < \infty;\ t = 0, 1.$$

Hence the likelihood function of μ is

$$L(\mu) = \exp\{-\tfrac{1}{2}(x-\mu)^2/10^{2t}\} \qquad \text{for } -\infty < \mu < \infty.$$

The MLE is $\hat{\mu} = x$, and $(\hat{\mu}, T)$ is a pair of minimally sufficient statistics for μ. Note that T is an ancillary statistic because its distribution does not depend upon μ.

Because of the symmetry, it is natural to consider symmetric intervals $X \pm a$.

(a) *Conditional reference set.* Since T is ancillary, the arguments of Section 15.3 imply that coverage probabilities should be calculated from the conditional distribution of X (or $\hat{\mu}$) given the observed value of T. Thus the coverage probability of $X \pm a$ is

$$\begin{aligned} \text{CP}(\mu_0) &= P\{\mu_0 \in X \pm a | T = t\} = P\{|X - \mu_0| \leq a | T = t\} \\ &= P\{|Z| \leq a/10^t\} \end{aligned}$$

where $Z \sim N(0, 1)$. For instance, if $a = 3$, the coverage probability is $P\{|Z| \leq 3\} = 0.997$ when $t = 0$, and $P\{|Z| \leq 0.3\} = 0.236$ when $t = 1$. The 95% confidence interval for μ is $X \pm 1.96$ when $t = 0$, and $X \pm 19.6$ when $t = 1$.

(b) *Unconditional reference set.* The unconditional coverage probability of the interval $X \pm a$ is

$$\begin{aligned} \text{CP}(\mu_0) &= P\{\mu_0 \in X \pm a\} = P\{|X - \mu_0| \leq a\} \\ &= P\{|X - \mu_0| \leq a | T = 0\} P\{T = 0\} + P\{|X - \mu_0| \leq a | T = 1\} P\{T = 1\} \\ &= \tfrac{1}{2} P\{|Z| \leq a\} + \tfrac{1}{2} P\left\{|Z| \leq \frac{a}{10}\right\}, \end{aligned}$$

where $Z \sim N(0, 1)$. For instance, the coverage probability of $X \pm 3$ is

$$\tfrac{1}{2} P\{|Z| \leq 3\} + \tfrac{1}{2} P\{|Z| \leq 0.3\} = \tfrac{1}{2}(0.997 + 0.236) = 0.617$$

for all μ_0, and so $X \pm 3$ is a 61.7% confidence interval for μ. Similarly, we find that $X \pm 16.45$ is a 95% confidence interval for μ. The 95% coverage probability is achieved by including μ_0 with probability 1 whenever the precise technique is used ($t = 0$), and with probability 0.9 whenever $t = 1$.

Clearly it is the conditional reference set which is appropriate in this example. If it is known that the measurement was made with the more precise technique, then the narrower interval $x \pm 1.96$ should be given. The fact that half of future measurements would be made with the less precise technique is irrelevant in so far as inferences about μ are concerned.

(c) *Likelihood ratio statistic.* The likelihood ratio statistic for testing H: $\mu = \mu_0$ is

$$D = -2r(\mu_0) = 2[l(\hat{\mu}) - l(\mu_0)] = (X - \mu_0)^2/10^{2T}.$$

When $T = 0$, D is the square of the $N(0, 1)$ variate $X - \mu_0$, and when $T = 1$, D is the square of the $N(0, 1)$ variate $(X - \mu_0)/10$. Thus the conditional distribution of D given $T = t$ is $\chi^2_{(1)}$ for $t = 0$ and for $t = 1$. It follows that the unconditional distribution of D is also $\chi^2_{(1)}$.

In Chapter 11 we suggested that confidence intervals be constructed from the likelihood function. Since $P\{\chi^2_{(1)} \leq 3.841\} = 0.95$, we take $D \leq 3.841$ to obtain the 95% confidence interval

$$X \pm 1.96 \times 10^T.$$

This interval has coverage probability 0.95 *both conditionally and unconditionally*:

$$P\{\mu_0 \in X \pm 1.96 \times 10^T | T = t\} = P\{\mu_0 \in X \pm 1.96 \times 10^T\} = 0.95.$$

Similarly, we have

$$P\{D \geq D_{\text{obs}} | T = t\} = P\{D \geq D_{\text{obs}}\},$$

so *we get the same significance level whether or not we condition on* T. When the likelihood ratio statistic is used, we get the correct answer even if we use the wrong (unconditional) reference set!

We noted in Section 15.3 that, if $(\hat{\theta}, T)$ is minimally sufficient for θ and T is ancillary, then $L(\theta)$ is proportional to $f(\hat{\theta}|T = t)$. Because of this, significance tests and confidence intervals based on the likelihood ratio statistic will automatically reflect the presence of ancillary statistics, and conditional significance levels and coverage probabilities will usually differ only slightly from the unconditional values. Choice of the appropriate reference set for inferences about θ is less important when we work with the likelihood ratio statistic.

15.5. Conditional Tests for Composite Hypotheses

H is called a composite hypothesis if, under H, there remains an unknown parameter or vector of parameters θ. Most of the examples in Chapters 12, 13, and 14 involved tests of composite hypotheses.

A special feature of the normal distribution examples of Chapters 13 and 14 is that the exact distribution of the likelihood ratio statistic D does not depend upon the values of any unknown parameters. For instance, the likelihood ratio statistic for testing hypotheses about the slope β in a straight line model is

$$D \equiv n \log\left[1 + \frac{1}{n-2} T^2\right] \quad \text{where } T \equiv \frac{\hat{\beta} - \beta}{\sqrt{s^2 c}} \sim t_{(n-2)}$$

and $c = 1/S_{xx}$ (see Sections 14.4 and 13.6). The distribution of T does not depend upon the values of the unknown intercept α and variance σ^2, and so neither does the distribution of D. Thus $P\{D \geq D_{\text{obs}}\}$ does not depend on α or σ^2.

Usually, the exact distribution of the test statistic D *does* depend upon the value of any unknown parameter θ not specified by the hypothesis. Then $P\{D \geq D_{\text{obs}}\}$ will be a function of θ rather than a numerical value.

One way around this problem is to compute the significance level from an appropriate conditional distribution which does not depend upon θ. Suppose that, under H, T is a sufficient statistic or set of sufficient statistics for θ. Then, by (15.2.1), we can write the probability of a typical outcome y as

$$P(Y = y; \theta) = c(y) \cdot H(t; \theta) \tag{15.5.1}$$

where $t = T(y)$, and c does not depend upon θ. By (15.2.4), the conditional probability of y given that $T = t$ is

$$P(Y = y | T = t) = c(y)/d(t) \tag{15.5.2}$$

where $d(t)$ is the sum of $c(y)$ over all y for which $T(y) = t$.

Suppose that we compute the significance level from the conditional distribution of Y given the observed value of T:

$$\text{SL} = P\{D \geq D_{\text{obs}} | T = t\}. \tag{15.5.3}$$

Then, since this conditional distribution does not depend upon θ, we shall obtain a numerical value for the significance level.

An example follows which illustrates this conditional procedure, and some general comments are given at the end of the section. Additional examples of conditional tests for composite hypotheses are considered in Section 15.6.

The Hardy–Weinberg Law

In some simple cases, the inheritance of a characteristic such as flower color is governed by a single gene which occurs in two forms, R and W say. Each individual has a pair of these genes, one obtained from each parent, so there are three possible genotypes: RR, RW, and WW.

Suppose that, in both the male and female populations, a proportion θ of the genes are of type R and the other $1 - \theta$ are of type W. Suppose further that

mating occurs at random with respect to this gene pair. Then the proportions of individuals with genotypes RR, RW, and WW in the next generation will be

$$p_1 = \theta^2,\ p_2 = 2\theta(1-\theta),\ p_3 = (1-\theta)^2. \tag{15.5.4}$$

Furthermore, if random mating continues, these proportions will remain nearly constant for generation after generation. This famous result from Genetics is called the Hardy–Weinberg Law.

Suppose that n individuals (e.g. pea plants) are selected at random and are classified according to genotype. Let y_1 be the number with genotype RR (red flowers), y_2 the number with genotype RW (pink flowers), and y_3 the number with genotype WW (white flowers), where $y_1 + y_2 + y_3 = n$. We wish to test whether these observed frequencies are consistent with the Hardy–Weinberg Law (15.5.4).

Note that, under (15.5.4), there remains an unknown parameter θ to be estimated from the data. Thus the hypothesis to be tested is composite. Following the procedure described above, we shall calculate the significance level from the conditional distribution of the Y_i's given the observed value of a sufficient statistic T.

Under the hypothesis, the distribution of the Y_i's is trinomial with probability parameters as given in (15.5.4):

$$P(Y = y;\theta) = \binom{n}{y_1\, y_2\, y_3}[\theta^2]^{y_1}[2\theta(1-\theta)]^{y_2}[(1-\theta)^2]^{y_3}$$

$$= \binom{n}{y_1\, y_2\, y_3} 2^{y_2}\theta^t(1-\theta)^{2n-t}$$

where $t = 2y_1 + y_2$. Here $T \equiv 2Y_1 + Y_2$ is a sufficient statistic for θ, and we have

$$P(Y = y;\theta) = c(y)\cdot H(t;\theta)$$

where $H(t;\theta) = \theta^t(1-\theta)^{2n-t}$, and

$$c(y) = \binom{n}{y_1\, y_2\, y_3} 2^{y_2}.$$

By (15.5.2), the conditional probability of outcome y given that $T = t$ is $c(y)/d(t)$ where $d(t)$ is the sum of $c(y)$ over all y such that $2y_1 + y_2 = t$, and, of course, $y_1 + y_2 + y_3 = n$.

The MLE of θ is $\hat\theta = t/2n$, and the estimated expected frequencies for the three genotypes are

$$e_1 = n\hat\theta^2,\ e_2 = 2n\hat\theta(1-\hat\theta),\ e_3 = n(1-\hat\theta)^2.$$

By (12.5.1), the likelihood ratio statistic for testing the hypothesis (15.5.4) is

$$D(y) = 2\Sigma y_j \log(y_j/e_j).$$

If n is large, then D has approximately a χ^2 distribution with one degree of freedom, and

$$\text{SL} \approx P\{\chi^2_{(1)} \geq D_{\text{obs}}\}.$$

The unconditional probability of the event $D \geq D_{\text{obs}}$ would be computed by summing the trinomial probabilities $P(y;\theta)$ over all y_1, y_2, y_3 such that $D \geq D_{\text{obs}}$. This probability will depend upon what value is taken for the unknown parameter θ. Instead, we compute the conditional probability of $D \geq D_{\text{obs}}$ given the observed value of T. This conditional probability is found by summing $P(Y = y|T = t)$, and it will not depend upon θ.

Since $\hat\theta = t/2n$, conditioning on the observed t is equivalent to restricting attention to those outcomes for which $\hat\theta$ equals its observed value. Hence the expected frequencies e_1, e_2, e_3 will be the same for all outcomes considered in the conditional test.

To compute the exact conditional significance level, we list all possible outcomes (y_1, y_2, y_3) with $y_1 + y_2 + y_3 = n$ and $2y_1 + y_2 = t$. For each we calculate $D(y)$ and $c(y)$. We sum $c(y)$ over all these outcomes to obtain $d(t)$, and divide to get the conditional probabilities $P(Y = y|T = t)$. Finally, we sum these probabilities over all outcomes such that $D(y) \geq D_{\text{obs}}$. This procedure is illustrated in the following example.

EXAMPLE 15.5.1. Suppose that $n = 20$ individuals were observed, and that the observed frequencies were as follows:

Genotype	RR	RW	WW	Total
Obs. freq. y_i	5(2.8)	5(9.4)	10(7.8)	20

Here $t = 2y_1 + y_2 = 15$, and $\hat\theta = t/2n = 0.375$. The expected frequencies are as shown in parentheses, and

$$D_{\text{obs}} = 2\left[5\log\frac{5}{2.8} + 5\log\frac{5}{9.4} + 10\log\frac{10}{7.8}\right] = 4.45.$$

All possible outcomes (y_1, y_2, y_3) with $y_1 + y_2 + y_3 = 20$ and $2y_1 + y_2 = 15$ are listed in Table 15.5.1 together with the corresponding values of $D(y)$ and $c(y)$. Summing $c(y)$ gives $d(15) \times 10^{-10} = 4.0225$, and we divide by this value to get the conditional probabilities $P(Y = y|T = 15) = c(y)/d(15)$. There are four outcomes in the table such that $D \geq D_{\text{obs}}$, and summing their probabilities gives

$$\text{SL} = 0.0126 + 0.0370 + 0.0028 + 0.0001 = 0.0525.$$

For comparison, the large-sample approximation gives

$$\text{SL} \approx P\{\chi^2_{(1)} \geq 4.45\} = 0.035.$$

The agreement is not too bad in view of the small expected frequencies.

Table 15.5.1. Evaluation of the Exact Conditional Significance Level in a Test of the Hardy–Weinberg Law

y_1	y_2	y_3	$D(y)$	$c(y) \times 10^{-10}$	$P(Y=y\|T=15)$
0	15	5	9.57	0.0508	0.0126
1	13	6	3.22	0.4445	0.1105
2	11	7	0.60	1.2383	0.3078
3	9	8	0.04	1.4189	0.3527
4	7	9	1.30	0.7095	0.1764
*5	5	10	4.45	0.1490	0.0370
6	3	11	9.86	0.0113	0.0028
7	1	12	18.69	0.0002	0.0001
Total			–	4.0225	0.9999

Note that only one of the y_i's is "free to vary" in Table 15.5.1, the other two then being determined by the constraints $y_1 + y_2 + y_3 = 20$ and $2y_1 + y_2 = 15$. This is directly related to the single degree of freedom in the χ^2 approximation.

It is possible to obtain an algebraic formula for $P(Y=y|T=t)$ in this case. Since $T \equiv 2Y_1 + Y_2$ represents the total number of R-genes out of $2n$ genes selected at random, the distribution of T is binomial $(2n, \theta)$. It follows that

$$P(T=t) = \binom{2n}{t}\theta^t(1-\theta)^{2n-t} \qquad \text{for } t = 0, 1, \ldots, 2n$$

and (15.2.3) gives

$$P(Y=y|T=t) = \frac{P(Y=y)}{P(T=t)} = \binom{n}{y_1\, y_2\, y_3} 2^{y_2} \Big/ \binom{2n}{t}.$$

In the example we have

$$P(Y=y|T=15) = \binom{20}{y_1\, y_2\, y_3} 2^{y_2} \Big/ \binom{40}{15}$$

and this formula could have been used to calculate the last column of Table 15.5.1.

Discussion

The conditional test is based on a factorization of the distribution of Y:

$$P(Y=y;\theta) = P(T=t;\theta) \cdot P(Y=y|T=t).$$

Since T is sufficient for θ, the first factor carries all of the information about θ,

the unknown parameter under H. The second factor does not depend upon θ, and is used for testing the hypothesis H.

Often T can be thought of as a measure of precision, and there are good reasons for conditioning on its observed value. For instance, $T \equiv 2Y_1 + Y_2$ indicates the amount of information available for testing the Hardy–Weinberg Law. If T is close to $2n$, then almost all individuals must necessarily fall in the RR class, whether or not the Hardy–Weinberg Law holds, and it will not be possible to obtain evidence against this hypothesis. A similar comment applies when T is close to 0. The prospect of obtaining evidence against the hypothesis is much better when T is close to n. Thus T is a measure of the experiment's precision, and one can argue, as in Section 15.4, that inferences should be made conditional on its observed value.

Conditioning on a set of sufficient statistics will not always give satisfactory results, because in so doing we may discard some of the information relevant to assessing the hypothesis. This information loss can be substantial in some examples. As a general rule, it seems dangerous to use this conditional procedure unless $\hat{\theta}$ is sufficient for θ and T is a one-to-one function of $\hat{\theta}$, as in the Hardy–Weinberg example. If $\hat{\theta}$ is not sufficient, it is probably better to use the conditional distribution of Y given $\hat{\theta}$, even though this distribution will not be completely independent of θ.

Again, there are advantages in taking D to be the likelihood ratio statistic for testing H. In large samples, D and $\hat{\theta}$ are distributed independently of one another. Significance levels computed from the χ^2 approximation (12.3.2) can therefore be regarded as either conditional (given $\hat{\theta}$) or unconditional. Except in very small samples, the conditional distribution of D given $\hat{\theta}$ will be almost the same as the unconditional distribution of D. As we noted in Section 15.4, the distribution of the likelihood ratio statistic is remarkably stable under different possible choices for the reference set. With likelihood ratio tests it usually doesn't matter much whether the significance level is computed conditionally (given $\hat{\theta}$) or unconditionally.

15.6. Some Examples of Conditional Tests

A conditional procedure for testing composite hypotheses was described in Section 15.5. In this section, we give some additional examples of conditional tests.

Comparison of Binomial Proportions

Suppose that Y_1 and Y_2 are independent with $Y_1 \sim$ binomial (n_1, p_1) and $Y_2 \sim$ binomial (n_2, p_2), and that we wish to test the composite hypothesis $H: p_1 = p_2 = p$, say, where p is unknown. Under H, the joint probability function

of Y_1 and Y_2 is

$$P(Y=y;p)=\binom{n_1}{y_1}p^{y_1}(1-p)^{n_1-y_1}\cdot\binom{n_2}{y_2}p^{y_2}(1-p)^{n_2-y_2}$$
$$=\binom{n_1}{y_1}\binom{n_2}{y_2}p^t(1-p)^{n_1+n_2-t}$$

where $t=y_1+y_2$. Thus $T\equiv Y_1+Y_2$ is a sufficient statistic for p, and the test of H will be based on the conditional distribution of Y_1 and Y_2 given the observed value of T.

The distribution of T is binomial (n_1+n_2, p), and so

$$P(T=t;p)=\binom{n_1+n_2}{t}p^t(1-p)^{n_1+n_2-t}.$$

By (15.2.3), the conditional distribution of Y given $T=t$ is

$$P(Y=y|T=t)=\frac{P(Y=y;p)}{P(T=t;p)}=\binom{n_1}{y_1}\binom{n_2}{y_2}\Big/\binom{n_1+n_2}{t}$$

where $y_1+y_2=t$. This conditional distribution is hypergeometric, and it does not depend upon the unknown parameter p.

Under H, the MLE of p is $\tilde{p}=t/(n_1+n_2)$. From Section 12.4, the likelihood ratio statistic for testing H is

$$D(y)=2\Sigma y_i\log\frac{y_i}{n_i\tilde{p}}+2\Sigma(n_i-y_i)\log\frac{n_i-y_i}{n_i(1-\tilde{p})}.$$

The exact conditional significance level is found by summing $P(Y=y|T=t)$ over all y such that $y_1+y_2=t$ and $D\geq D_{\text{obs}}$. Note that, since $\tilde{p}=t/(n_1+n_2)$, the estimated expected frequencies $n_i\tilde{p}$ and $n_i(1-\tilde{p})$ will be the same for all y considered in the conditional test.

EXAMPLE 15.6.1. For the data of Example 12.4.1, we have $n_1=n_2=44$, and the observed value of T is $t=14+4=18$. Hence $\tilde{p}=18/88$, and the estimated expected frequencies are $n_i\tilde{p}=9$, and $n_i(1-\tilde{p})=35$. The likelihood ratio statistic for testing $H\colon p_1=p_2$ is

$$D(y_1,y_2)=2\Sigma y_i\log(y_i/9)+2\Sigma(44-y_i)\log((44-y_i)/35)$$

with observed value $D_{\text{obs}}=D(4,14)=7.32$. The conditional probability function of (Y_1, Y_2) given that $T=18$ is

$$g(y_1,y_2)=\binom{44}{y_1}\binom{44}{y_2}\Big/\binom{88}{18}$$

where $y_1+y_2=18$. If $p_1=p_2$, then the 18 rats with tumors are a random sample without replacement from the 88 rats in the study, and $g(y_1,y_2)$ is the probability that y_1 of the rats with tumors received the low dose and the other $y_2=18-y_1$ received the high dose.

Table 15.6.1. Calculation of Conditional Significance Level in Example 15.6.1

y_1	y_2	$g(y_1,y_2)$	$D(y_1,y_2)$	y_1	y_2	$g(y_1,y_2)$	$D(y_1,y_2)$
0	18	0.0000	29.63	10	8	0.1818	0.28
1	17	0.0000	20.92	11	7	0.1215	1.13
2	16	0.0002	15.21	12	6	0.0616	2.55
3	15	0.0013	10.80	13	5	0.0233	4.60
*4	14	0.0065	7.32	14	4	0.0065	7.32
5	13	0.0233	4.60	15	3	0.0013	10.80
6	12	0.0616	2.55	16	2	0.0002	15.21
7	11	0.1215	1.13	17	1	0.0000	20.92
8	10	0.1818	0.28	18	0	0.0000	29.63
9	9	0.2078	0.00	Total		1.0002	–

The 19 possible outcomes (y_1, y_2) with $y_1+y_2=18$ are listed in Table 15.6.1. There are 10 outcomes with $D\geq D_{\text{obs}}$, and we sum their conditional probabilities to obtain the exact conditional significance level, SL = 0.0160. For comparison, the large-sample approximation gives

$$\text{SL}\approx P\{\chi^2_{(1)}\geq 7.32\}=0.0068.$$

The agreement with the exact result is not very good, although the general conclusion (strong evidence that $p_1\neq p_2$) is the same in either case.

When there is only one degree of freedom, the accuracy of the large-sample approximation to the exact conditional significance level can often be improved by using a continuity correction (see Section 6.8). In this example, we replace $y_1=4$ by $y_1=4.5$ and $y_2=14$ by $y_2=13.5$ before computing D. We then obtain $D_{\text{obs}}=5.87$, and

$$\text{SL}\approx P\{\chi^2_{(1)}\geq 5.87\}=0.0154$$

which is much closer to the exact result.

Exact Test for Independence

As in Section 12.6, we consider an $a\times b$ contingency table (f_{ij}) with row totals r_i and columns totals c_j where $\Sigma r_i=\Sigma c_j=n$. The f_{ij}'s have a multinomial distribution, and the independence hypothesis is $H\colon p_{ij}=\alpha_i\beta_j$ where the α_i's and β_j's are unknown parameters. Under H, the probability of the f_{ij}'s is

$$P(f;\alpha,\beta)=(f_{11}f_{12}^{n}\cdots f_{ab})\prod_{i=1}^{a}\prod_{j=1}^{b}(\alpha_1\beta_j)^{f_{ij}}$$
$$=(f_{11}f_{12}^{n}\cdots f_{ab})\Big(\prod_{i=1}^{a}\alpha_i^{r_i}\Big)\Big(\prod_{j=1}^{b}\beta_j^{c_j}\Big).$$

The r_i's and c_j's are sufficient statistics for the unknown parameters.

The r_i's have a multinomial distribution with class probabilities $\alpha_1, \alpha_2, \ldots, \alpha_a$ and the c_j's are multinomial with class probabilities $\beta_1, \beta_2, \ldots, \beta_b$. Under the independence hypothesis, the r_i's are distributed independently of the c_j's. Hence the probability function of the f_{ij}'s given the sufficient statistics is

$$P(f|r, c) = (f_{11} f_{12} \cdots f_{ab}) \Big/ \binom{n}{r_1 \ldots r_a} \binom{n}{c_1 \ldots c_b}.$$

The exact conditional significance level will be computed from this conditional distribution.

By (12.6.1) and (12.6.2), the likelihood ratio statistic for testing the independence hypothesis is

$$D = 2\Sigma\Sigma f_{ij} \log (f_{ij}/e_{ij})$$

where $e_{ij} = r_i c_j / n$. Note that the estimated expected frequencies e_{ij} will be the same for all f_{ij}'s considered in a conditional test.

To carry out an exact test of the independence hypothesis, we list all tables (f_{ij}) having the same row and column totals as the observed table. The conditional probability and value of D are computed for each such table. The exact conditional significance level is then found by summing $P(f|r, c)$ over all such tables for which $D \geq D_{\text{obs}}$. Except in very small examples, a computer will be needed for the calculations.

EXAMPLE 15.6.2. In Example 12.6.1 we carried out an approximate test for independence in the following 2 by 2 table:

44 (39.56)	9 (13.44)	53
9 (13.44)	9 (4.56)	18
53	18	71

Expected frequencies under the independence hypothesis are shown in parentheses.

For an exact test, we need to list all tables having the same row and column totals as the observed table. The general form of such tables is

x	$53-x$	53
$53-x$	$x-35$	18
53	18	71

where $x = 35, 36, \ldots, 53$. Only one of the frequencies is "free to vary", corresponding to the single degree of freedom for the approximate χ^2 test.

The conditional p.f. of such a table is

$$g(x) = \binom{71}{x \quad 53-x \quad 53-x \quad x-35} \Big/ \binom{71}{53 \quad 18} \binom{71}{53 \quad 18}$$

Table 15.6.2. Conditional Exact Test for Independence in a 2 × 2 Table

x	$g(x)$	$D(x)$	x	$g(x)$	$D(x)$
35	0.0021	12.47	45	0.0012	10.69
36	0.0187	6.16	46	0.0002	14.97
37	0.0731	2.92	47	0.0000	20.05
38	0.1641	1.02	48	0.0000	26.00
39	0.2367	0.13	49	0.0000	32.96
40	0.2320	0.07	50	0.0000	41.12
41	0.1594	0.78	51	0.0000	50.81
42	0.0781	2.21	52	0.0000	62.75
43	0.0275	4.33	53	0.0000	80.40
*44	0.0069	7.15	Total	1.0000	–

which simplifies to a hypergeometric distribution:

$$g(x) = \binom{53}{x}\binom{18}{53-x} \Big/ \binom{71}{53} \qquad \text{for } x = 35, 36, \ldots, 53.$$

The likelihood ratio statistic is

$$D(x) = 2\left[x \log \frac{x}{39.56} + \cdots + (x-35) \log \frac{x}{4.56}\right],$$

with observed value $D_{\text{obs}} = D(44) = 7.15$.

From Table 15.6.2 we see that $D(x) \geq D_{\text{obs}}$ for $x = 35$ and for $x \geq 44$. Hence the exact significance level is

$$\text{SL} = g(35) + g(44) + g(45) + \cdots + g(53) = 0.0104,$$

and the observed table gives strong evidence against the hypothesis of independence.

In this example the row and column totals are modelled as random variables, but we condition on their observed values in the exact test for independence. The independence test would be the same if some or all of the marginal totals had been fixed prior to the experiment. See the note following Example 12.6.1.

EXAMPLE 15.6.3. Is the following 2 × 3 contingency table consistent with the hypothesis that the row and column classifications are independent?

	B_1	B_2	B_3	Total
A_1	1 (1.8)	1 (3.0)	7 (4.2)	9
A_2	2 (1.2)	4 (2.0)	0 (2.8)	6
Total	3	5	7	15

SOLUTION. The expected frequencies under the hypothesis of independence are shown above in parentheses. Since these are small, it is advisable to carry out an exact test for independence. For this we require a list of all 2×3 tables having the same marginal totals as the observed table. The general form of these tables is

x	y	$9-x-y$	9
$3-x$	$5-y$	$x+y-2$	6
3	5	7	15

Just two of the frequencies are "free to vary", corresponding to the two degrees of freedom for the χ^2 approximation.

There are 24 pairs (x, y) with $0 \le x \le 3$ and $0 \le y \le 5$, but three of these have $x+y<2$ and would give a negative entry in the table. Thus there are only 21 allowable pairs (x, y) (see Table 15.6.3). The conditional probability function is

$$g(x, y) = \binom{15}{x \quad y \quad 9-x-y \quad 3-x \quad 5-y \quad x+y-2} \Big/ \binom{15}{9 \quad 6} \binom{15}{3 \quad 5 \quad 7}$$

$$= \frac{725.035}{x!\,y!(9-x-y)!(3-x)!(5-y)!(x+y-2)!}$$

and the likelihood ratio statistic is

$$D(x, y) = 2\left[x \log \frac{x}{1.8} + y \log \frac{y}{3.0} + \cdots + (x+y-2) \log \frac{x+y-2}{2.8}\right].$$

From Table 15.6.3 we see that $D_{\text{obs}} = D(1, 1) = 11.37$. There are 5 tables for which $D \ge 11.37$, and the exact significance level is the sum of their

Table 15.6.3. Conditional Exact Test for Independence in a 2×3 Table

x	y	$g(x)$	$D(x)$	x	y	$g(x)$	$D(x)$
0	2	0.0020	13.46	2	2	0.1259	1.27
0	3	0.0140	7.72	2	3	0.2098	0.08
0	4	0.0210	6.81	2	4	0.1049	1.81
0	5	0.0070	10.63	2	5	0.0126	8.00
*1	1	0.0030	11.37	3	0	0.0014	14.45
1	2	0.0420	3.90	3	1	0.0210	6.81
1	3	0.1259	1.27	3	2	0.0699	3.90
1	4	0.1049	1.81	3	3	0.0699	3.90
1	5	0.0210	6.81	3	4	0.0210	6.81
2	0	0.0006	16.37	3	5	0.0014	14.45
2	1	0.0210	5.63	Total		1.0002	–

conditional probabilities:

$$\text{SL} = g(0, 2) + g(1, 1) + g(2, 0) + g(3, 0) + g(3, 5) = 0.0084.$$

The observed table gives strong evidence against the hypothesis of independence. □

PROBLEMS FOR SECTION 15.6

1. In a pilot study, a new deodorant was found to be effective for 2 of 10 men tested and for 4 of five women tested. Carry out an exact conditional test of the hypothesis that the deodorant is equally effective for men and women.

2.† Two manufacturing processes produce defective items with probabilities p_1 and p_2, respectively. It was decided to examine four items from the first process and sixteen items from the second. In each case, two defectives were found. Perform an exact conditional test of the hypothesis $p_1 = p_2$.

3.† Two manufacturing processes produce defective items with probabilities p_1 and p_2, respectively. Items were examined from the first process until the rth defective had been obtained, by which time there had been x_1 good items. The second process gave x_2 good items before the rth defective.

   (a) Write down the joint probability function of X_1 and X_2. Show that, if $p_1 = p_2 = p$, then $T \equiv X_1 + X_2$ is a sufficient statistic for p.
   (b) For each process, items were examined until $r = 2$ defectives had been found. Process 1 gave 2 good items, and process 2 gave 14 good items. Carry out an exact conditional test of the hypothesis $p_1 = p_2$, and compare the significance level with that obtained in Problem 2.

4. Twelve pea plants were observed, and there were four of each of the genotypes RR, RW, and WW. Use a conditional test to determine whether these results are consistent with the Hardy–Weinberg law (Section 15.5).

5. A study of the effect of Interferon on the severity of chicken pox was carried out with 44 childhood cancer victims who had developed chicken pox. Doctors gave Interferon to 23 children, and the other 21 received an inactive placebo. The disease was fatal or life-threatening in 2 of those who received Interferon, and in 6 of those who did not. Test the hypothesis that disease severity is independent of the treatment.

6.† An investigator wishes to learn whether the tendency to crime is influenced by genetic factors. He argues that, if there is no genetic effect, the incidence of criminality among identical twins should be the same as that among fraternal twins. Accordingly, he examines the case histories of 30 criminals with twin brothers, of whom 13 are identical and 17 are fraternal. He finds that 12 of the twin brothers have also been convicted of crime, but only two of these are fraternal twins. Perform an exact conditional test of the hypothesis of no genetic effect.

7. (a) Suppose that X and Y are independent and have Poisson distributions with means μ and ν, respectively. Derive the appropriate conditional distribution for a test of $H: \mu = k\nu$, where k is a given constant.

(b) There were 13 accidents in a large manufacturing plant during the two weeks prior to the introduction of a new safety program. There were only 3 accidents in the week following its introduction. Test the hypothesis that the accident rate has not changed.

8. A likelihood ratio test for the hypothesis of marginal homogeneity in a 2 by 2 table was described in Section 12.8.

(a) Show that the significance level in an exact conditional test of this hypothesis will be computed from the binomial distribution

$$\binom{f_{12}+f_{21}}{f_{12}}\left(\frac{1}{2}\right)^{f_{12}+f_{21}}.$$

(b) Carry out a conditional exact test using the data of Section 12.8.

9. Articles coming off a production line may be classified as acceptable, repairable, or useless. If n items are examined let X_1, X_2, and X_3 be the number of acceptable, repairable, and useless items found. Suppose that it is twice as probable that an item is acceptable as it is that it is repairable.

(a) Show that $X_1 + X_2$ is a sufficient statistic for p, the probability of a repairable item.

(b) Of six items examined, one is acceptable, four are repairable, and one is useless. Use an exact conditional test to assess the agreement of these observations with the model.

10.† In a certain factory there are three work shifts: days (#1), evenings (#2), and nights (#3). Let X_i denote the number of accidents in the ith shift ($i = 1, 2, 3$). The X_i's are assumed to be independent Poisson variates with means μ_1, μ_2, and μ_3. There are only half as many workers on the night shift as on the other two. Hence, if the accident rate is constant over the three shifts, we should have $\mu_1 = \mu_2 = 2\mu_3$. Set up an exact conditional test for this hypothesis.

11. Suppose that n families each with three children are observed. Let X_i be the number of such families which contain i boys and $3 - i$ girls ($i = 0, 1, 2, 3$). If births are independent, the probability that a family of 3 has i boys will be given by

$$p_i = \binom{3}{i}\theta^i(1-\theta)^{3-i} \qquad \text{for } i = 0, 1, 2, 3$$

where θ is the probability of a male child.

(a) Show that $T \equiv X_1 + 2X_2 + 3X_3$ is a sufficient statistic for θ and has a binomial distribution with parameters $(3n, \theta)$.

(b) In 8 families there were 3 with three boys, 2 with one boy, and 3 with no boys. Use an exact conditional test to investigate whether these results are consistent with the model.

12.† In an experiment to detect linkage of genes, there are four possible types of offspring. According to theory, these four types have probabilities $p/2$, $1 - p/2$, $1 - p/2$, and $p/2$, where p is an unknown parameter called the recombination fraction. Let X_1, X_2, X_3, and X_4 be the frequencies of the four offspring types in n independent repetitions.

(a) Find a sufficient statistic for p.

(b) If the genes are not linked, they lie on different chromosomes, and $p = \frac{1}{2}$. Evidence against the hypothesis $p = \frac{1}{2}$ is thus evidence that the genes are linked. Describe an exact test for this hypothesis.

(c) Describe exact and approximate tests of the model when p is unknown.

13. A lethal drug is administered to n rats at each of k doses $d_1, d_2, \ldots, d_k$. Let the numbers of deaths be $Y_1, Y_2, \ldots, Y_k$. According to the logistic model (Section 10.5), the probability of death at dose d_i is

$$p(d_i) = e^{\alpha+\beta d_i}/(1 + e^{\alpha+\beta d_i}).$$

(a) Show that $S \equiv \Sigma Y_i$ and $T \equiv \Sigma d_i Y_i$ are sufficient statistics for the unknown parameters α and β.

(b) Show that the conditional probability function of the Y_i's given S and T is

$$c\binom{n}{y_1}\binom{n}{y_2}\cdots\binom{n}{y_k},$$

where c is chosen so that the total conditional probability is 1.

(c) In an experiment with 10 rats at each of 3 doses $-1, 0, 1$, the numbers of deaths observed were 3, 0, and 10, respectively. Perform an exact conditional test of the logistic model.

(d) In an experiment with 10 rats at each of the 4 doses $-3, -1, 1, 3$, the numbers of deaths observed were 1, 6, 4, and 10, respectively. Are these frequencies consistent with the logistic model?

CHAPTER 16*

Topics in Statistical Inference

In Chapters 9–15 we have used likelihood methods, confidence intervals, and significance tests in making inferences about an unknown parameter θ. In Sections 1 and 2 below, we consider two additional methods for making inferences about an unknown parameter. With both the fiducial argument and Bayesian methods, information concerning θ is summarized in a probability distribution defined on the parameter space. For Bayesian methods one requires prior information about θ which is also in the form of a probability distribution. For the fiducial argument, θ must be completely unknown before the experiment.

In Section 3, we consider the problem of predicting a value of a random variable Y whose probability distribution depends upon an unknown parameter θ. When a Bayesian or fiducial distribution for θ is available, one can obtain a predictive distribution for Y which does not depend upon θ. Section 4 considers the use of predictive distributions in statistical inference, with particular reference to the Behrens–Fisher problem. Finally, in Section 5 we illustrate how a test of a true hypothesis can be used to obtain intervals of reasonable values for a future observation or an unknown parameter.

16.1. The Fiducial Argument

Suppose that we have obtained data from an experiment whose probability model involves a real-valued parameter θ which is completely unknown. We shall see that, under certain conditions, it is possible to deduce the probability

*This chapter may be omitted on first reading.

that $\theta \leq k$ for any specified parameter value k. The procedure for obtaining this probability is called the *fiducial argument*, and the probability is called a *fiducial probability* to indicate the method by which it was obtained.

Probability Distributions of Constants

In the fiducial argument, the probability distribution of a variate U is regarded as a summary of all the available information about U. This distribution continues to hold until such time as additional information about U becomes available. If U has a certain distribution before an experiment is performed, and if the experiment provides no information about U, then U has the same distribution after the experiment as before.

For example, consider a lottery in which there are N tickets numbered $1, 2, \ldots, N$, one of which is selected at random. Let U denote the number on the winning ticket. Then

$$P(U = u) = \frac{1}{N} \qquad \text{for } u = 1, 2, \ldots, N. \tag{16.1.1}$$

Now suppose that the winning ticket has been chosen, but that the number U has not been announced. A value of U has now been determined, but we have no more information concerning what that value is than we had before the draw. A ticket-holder would presumably feel that he had the same chance of winning as he had before the draw was made. *The fiducial argument is based on the assertion that* (16.1.1) *summarizes the uncertainty about U even after the draw has been made, provided that no information concerning the outcome of the draw is available.* After the draw, U is no longer subject to random variation, but is fixed at some unknown value. Now (16.1.1) summarizes all the available information concerning the unknown constant U, and may be called its *fiducial distribution.*

The fiducial argument does not involve any new "definition" of probability. Instead, it enlarges the domain of application of the usual (long-run relative frequency) notion of probability. Of course, one could take the position (as some people have) that (16.1.1) applies only before the draw, and that, after the draw, no probability statements whatsoever can be made. This position seems unnecessarily restrictive, and if adopted, would rule out many important applications of probability.

Before proceeding with the general discussion, we illustrate the fiducial argument in two examples.

EXAMPLE 16.1.1. A deck of N cards numbered $1, 2, \ldots, N$ is shuffled and one card is drawn. Let U denote the number on this card. Then U has probability distribution (16.1.1). To this number is added a real number θ which is completely unknown to us. We are not told the value of U or the value of θ, but only the value of their total $T \equiv \theta + U$. What can be said about θ in the light of an observed total t?

The observed total t could have arisen in N different ways:

$$(U = 1, \theta = t - 1), (U = 2, \theta = t - 2), \ldots, (U = N, \theta = t - N).$$

Given t, there is a one-to-one correspondence between values of U and possible values of θ. If we knew the value of θ, we could determine which value of U had been obtained. If we knew that θ was an even integer, then we could deduce whether an odd or even value of U had been obtained. However, if we know nothing about θ, then the experiment will tell us nothing about U; the state of uncertainty concerning the value of U will be the same after the experiment as before. Hence we assume that (16.1.1) also holds when t is known. But, given t, θ has N possible values $t-1, t-2, \ldots, t-N$ in one-to-one correspondence with the possible values of U, and we may write

$$P(\theta = t - u) = P(U = u) = \frac{1}{n}, \qquad u = 1, 2, \ldots, N.$$

This probability distribution over the possible values of θ is called the *fiducial distribution of* θ.

For instance, suppose that $N = 13$, and that the observed total is $t = 20$. Then θ has 13 possible values 19, 18, 17, ..., 7, each with probability $\frac{1}{13}$. The probability of any subset of θ values is now obtained by addition. For example,

$$P(\theta \leq 11) = P(\theta = 11) + P(\theta = 10) + \cdots + P(\theta = 7) = \frac{5}{13}.$$

Alternately, we may note that if $\theta \leq 11$, then the observed total 20 must have resulted from a value of U greater than or equal to 9, and hence

$$P(\theta \leq 11) = P(U \geq 9) = \frac{5}{13}.$$

EXAMPLE 16.1.2. Suppose that $T \sim N(\theta, 1)$ where θ is completely unknown, and that the experiment yields an observed value t. We define $U \equiv T - \theta$, so that U has a standardized normal distribution. The observed value t arose from some pair of values $(U = u, \theta = t - u)$. Given t, there is a one-to-one correspondence between possible values of U and possible values of θ. Since θ is unknown, the experiment will tell us nothing about which value of U was actually obtained. Consequently, we assume that $U \sim N(0, 1)$ even after t has been observed.

We can now compute probabilities of statements about θ by transforming them into statements about U. For instance, $\theta \leq k$ if and only if $U \geq t - k$, and hence

$$P(\theta \leq k) = P(U \geq t - k) = 1 - F(t - k) = F(k - t) \qquad (16.1.2)$$

where F is the $N(0, 1)$ cumulative distribution function. For any k, the fiducial probability of $\theta \leq k$ can be obtained from $N(0, 1)$ tables. For example, if we observe $t = 10$, the fiducial probability of $\theta \leq 11$ is

$$P(\theta \leq 11) = F(11 - 10) = 0.841.$$

Note that probability statements obtained from (16.1.2) are the same as would be obtained if θ were a random variable having a normal distribution with mean t and variance 1. We say that given $T = t$, the *fiducial distribution* of θ is $N(t, 1)$. This does not mean that θ is a random variable, but rather that we know precisely as much about θ as we would about an observation to be taken at random from $N(t, 1)$.

From (16.1.2), the cumulative distribution function of the fiducial distribution of θ is $F(\theta - t)$, where F is the c.d.f. of $N(0, 1)$. Differentiation with respect to θ gives

$$\frac{\partial}{\partial \theta} F(\theta - t) = f(\theta - t) \frac{\partial(\theta - t)}{\partial \theta} = f(\theta - t)$$

where f is the p.d.f. of $N(0, 1)$. Hence the fiducial p.d.f. of θ is

$$f(\theta; t) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(\theta - t)^2\right\} \qquad \text{for } -\infty < \theta < \infty.$$

This is the p.d.f. of a normal distribution with mean t and variance 1. As a result of the fiducial argument, θ and T have switched roles, with the observed t now appearing as a "parameter" in the fiducial distribution of θ.

Sufficient Conditions for the Fiducial Argument

In the preceding two examples, we made use of a quantity U which was a function of both the data and the parameter, and whose probability distribution did not depend upon θ. Such a function is called a *pivotal quantity*.

The following conditions are sufficient to permit application of the fiducial argument in the one-parameter case:

C1. There is a single real-valued parameter θ which is completely unknown.
C2. There exists a statistic T which is minimally sufficient for θ.
C3. There exists a pivotal quantity $U \equiv U(T, \theta)$ such that
   (a) for each value of θ, $U(t, \theta)$ is a one-to-one function of t;
   (b) for each value of t, $U(t, \theta)$ is a one-to-one function of θ.

If the variate T is continuous, we also require that U be continuous (and hence monotonic) in both t and θ.

The purpose of conditions C2 and C3(a) is to ensure that inferences about θ are based on all of the relevant information contained in the data. C2 can be replaced by the weaker condition that there exists a set of minimally sufficient statistics (T, A) where T is real-valued and A is a vector of ancillary statistics (see Section 15.3). We then use the conditional distributions of T and U given the observed value of A.

Given $T = t$, there is a one-to-one correspondence between possible values

of θ and possible values of U by C3(b). Since θ is completely unknown, observing t will give us no information about which value of U was actually obtained. Hence we assume that the distribution of U is the same after t has been observed as it was before observing t. Given t, we can convert statements about θ into statements about U and hence obtain their (fiducial) probabilities.

The above conditions are quite restrictive. In particular, C3(a) and (b) imply a one-to-one correspondence between values of T given θ, and values of θ given T, which will very rarely exist if T is discrete. Example 16.1.1 is exceptional in that, when t is known, there are only finitely many possible values for θ.

If the sufficient statistic T is continuous, one can usually take $U \equiv F(T; \theta)$, where F is the cumulative distribution function of T. From Section 6.3, U has a uniform distribution between 0 and 1 for each value of θ, and hence is a pivotal quantity. Since $F(t; \theta) = P(T \leq t)$ is an increasing function of t, C3(a) will also be satisfied, and only C3(b) needs to be checked. If C3(b) holds, then $P(\theta \leq k)$ will be equal to either $F(t; k)$ or $1 - F(t; k)$, depending upon whether $F(t; \theta)$ is an increasing or decreasing function of θ, and the fiducial p.d.f. of θ is given by

$$f(\theta; t) = \left| \frac{\partial}{\partial \theta} F(t; \theta) \right|.$$

EXAMPLE 16.1.3. Suppose that the MLE $\hat{\alpha}$ is a sufficient statistic for the unknown parameter α, and that $\hat{\alpha} \sim N(\alpha, c)$ where c is a known constant. Then the standardized variable

$$Z \equiv (\hat{\alpha} - \alpha)/\sqrt{c}$$

is pivotal and is distributed as $N(0, 1)$. It satisfies conditions 3(a) and 3(b). To obtain the fiducial distribution of α, we assume that Z is still distributed as $N(0, 1)$ when the variate $\hat{\alpha}$ is replaced by its observed value. Then we have

$$\alpha \equiv \hat{\alpha} - Z\sqrt{c}$$

where $\hat{\alpha}$ and c are known constants, and (6.6.6) gives

$$\alpha \sim N(\hat{\alpha}, c).$$

Given $\hat{\alpha}$, the fiducial distribution of α is normal with mean $\hat{\alpha}$ and variance c.

EXAMPLE 16.1.4. Let $X_1, X_2, \ldots, X_n$ be independent variates having an exponential distribution with unknown mean θ. Then $T \equiv \Sigma X_i$ is sufficient for θ, and

$$U \equiv 2T/\theta \sim \chi^2_{(2n)}$$

is a pivotal quantity satisfying conditions 3(a) and 3(b). To obtain the fiducial distribution of θ, we replace T by its observed value t and assume that U is

still distributed as $\chi^2_{(2n)}$. Statements about θ can now be converted into statements about U, and their probabilities can be obtained from tables of the χ^2 distribution.

The fiducial p.d.f. of θ can be obtained from the p.d.f. of U by standard change of variables methods. By (6.9.1), the p.d.f. of U is

$$f(u) = ku^{n-1}e^{-u/2} \qquad \text{for } u > 0$$

where $k = 1/2^n\Gamma(n)$. The fiducial p.d.f. of θ is thus

$$g(\theta; t) = f(u) \cdot \left| \frac{du}{d\theta} \right| = k\left(\frac{2t}{\theta}\right)^{n-1} e^{-t/\theta} \cdot \frac{2t}{\theta^2}$$

$$= \frac{1}{\theta\Gamma(n)}\left(\frac{t}{\theta}\right)^n e^{-t/\theta} \qquad \text{for } \theta > 0.$$

EXAMPLE 16.1.5. Consider the situation described in Example 16.1.1, but now suppose that n cards are drawn at random with replacement from the deck. The same unknown θ is added to the number on each card, and we are told the n totals $x_1, x_2, \ldots, x_n$. We wish to make inferences about θ on the basis of the data.

Each X_i can take N equally probable values $\theta + 1, \theta + 2, \ldots, \theta + N$, so that the probability function of X_i is

$$f(x) = P(X_i = x) = N^{-1} \qquad \text{for } x = \theta + 1, \theta + 2, \ldots, \theta + N.$$

Under random sampling with replacement, the X_i's are independent, and hence their joint probability function is

$$f(x_1)f(x_2)\ldots f(x_n) = N^{-n} \qquad \text{for } \theta + 1 \leq x_1, x_2, \ldots, x_n \leq \theta + N.$$

The likelihood function of θ is thus constant over the range of possible parameter values. We must have $\theta + 1 \leq x_{(1)}$ and $\theta + N \geq x_{(n)}$, where $x_{(1)}$ and $x_{(n)}$ are the smallest and largest sample values, so that

$$L(\theta) = 1 \qquad \text{for } x_{(n)} - N \leq \theta \leq x_{(1)} - 1.$$

It follows that $X_{(1)}$ and $X_{(n)}$ are jointly minimally sufficient for θ.

The number of possible parameter values is

$$x_{(1)} - 1 - [x_{(n)} - N - 1] = N - a$$

where $A \equiv X_{(n)} - X_{(1)}$ is the sample range. The larger the value of A obtained, the more precisely we may determine the value of θ. If we observe $A = 0$, there are N equally likely values for θ, but if $A = N - 1$, the value of θ can be determined exactly without error. Thus A is a measure of the experiment's informativeness, and is in fact an ancillary statistic. To see this, we write $X_i \equiv \theta + U_i$, where U_i is the number on the ith card drawn ($i = 1, 2, \ldots, n$). Then $X_{(1)} \equiv \theta + U_{(1)}$ and $X_{(n)} \equiv \theta + U_{(n)}$, so that

$$A \equiv X_{(n)} - X_{(1)} \equiv U_{(n)} - U_{(1)}.$$

The distribution of A thus depends only on the range of numbers which appear on the n cards drawn, and does not depend upon θ.

We now define a statistic T such that the transformation from $X_{(1)}, X_{(n)}$ to T, A is one-to-one; for instance, we could take $T \equiv X_{(1)}$. Then T, A are jointly sufficient for θ and A is ancillary. Inferences about θ will be based on the conditional distribution of T given the observed value of A. To obtain this distribution, we could first derive the joint probability function of $X_{(1)}$ and $X_{(n)}$ as in Problem 7.2.11, change variables, sum out T to get the probability function of A, and divide to get the required conditional probability function,

$$f(t|a;\theta) = \frac{1}{N-a} \qquad \text{for } t = \theta+1, \theta+2, \ldots, \theta+N-a.$$

Given that $A = a$, the n totals must fall in a range of length a which lies entirely between $\theta+1$ and $\theta+N$. There are $N-a$ such ranges, with lower limits $\theta+1, \theta+2, \ldots, \theta+N-a$, and these will be equally probable.

Now define $U \equiv T-\theta$. The conditional distribution of U given that $A = a$ is uniform,

$$P(U = u|a) = \frac{1}{N-a} \qquad \text{for } u = 1, 2, \ldots, N-a, \tag{16.1.3}$$

and does not depend upon θ. Given A and θ, there is a one-to-one correspondence between possible values of U and T. Given A and T, there is a one-to-one correspondence between possible values of U and θ. Thus, when A is given, the sufficient conditions for the fiducial argument are satisfied. The fiducial distribution of θ is obtained by assuming that (16.1.3) continues to hold when T is replaced by its observed value t, and this gives

$$P(\theta = k) = \frac{1}{N-a} \qquad \text{for } k = t-1, t-2, \ldots, t-N+a. \tag{16.1.4}$$

For example, suppose that $N = 13$, and that we observe the $n = 4$ totals 17, 11, 14, 23. Then $t = x_{(1)} = 11$, $x_{(n)} = 23$, and $a = 23-11 = 12$. Now (16.1.4) implies that $\theta = 10$ with probability 1. In this case the experiment is very informative and completely determines the value of θ. If we were less fortunate, we might observe totals such as 13, 17, 19, 13. Then $t = 13$ and $a = 6$, so that now (16.1.4) gives

$$P(\theta = k) = \tfrac{1}{7} \qquad \text{for } k = 12, 11, 10, \ldots, 6.$$

There are now seven equally probable values of θ. In the worst possible case, we observe equal totals, 18, 18, 18, 18. Then $t = 18$, $a = 0$, and (16.1.4) gives

$$P(\theta = k) = \tfrac{1}{13} \qquad \text{for } k = 17, 16, 15, \ldots, 5$$

so that there are 13 equally probable values of θ.

Two-Parameter Fiducial Distributions

Sometimes a double application of the one-parameter fiducial argument can be used to obtain a two-parameter fiducial distribution. However, there are examples where this can be done in two or more different ways, leading to different two-parameter distributions. There are serious difficulties in extending the fiducial argument beyond the one-parameter case, and the precise conditions under which this can be done are not known.

16.2. Bayesian Methods

In all of the procedures discussed so far, only the information provided by the experimental data is formally taken into account. However, in some situations we may wish to incorporate information about θ from other sources as well. If this additional information is in the form of a probability distribution for θ, it can be combined with the data using Bayes's Theorem (3.6.1).

Suppose that the probability model for the experiment depends on a parameter θ, and that an event E with probability $P(E;\theta)$ is observed to occur. In addition, suppose that θ is itself a random variable with a known probability distribution, called the *prior distribution* of θ, with probability or probability density function g, say. The conditional distribution of θ given that E has occurred is called the *posterior distribution* of θ. The posterior distribution has probability or probability density function given by

$$f(\theta|E) = P(E;\theta)g(\theta)/P(E) \tag{16.2.1}$$

where $P(E)$ is a normalizing constant:

$$P(E) = \begin{cases} \sum_{\theta\in\Omega} P(E;\theta)g(\theta) & \text{if } \theta \text{ is discrete;} \\ \int_{-\infty}^{\infty} P(E;\theta)g(\theta)\,d\theta & \text{if } \theta \text{ is continuous.} \end{cases} \tag{16.2.2}$$

The posterior distribution combines the information about θ provided by the experimental data with the information contained in the prior distribution.

The likelihood function of θ based on the observed event E is given by

$$L(\theta; E) = kP(E;\theta)$$

where k does not depend upon θ. Hence we may write

$$f(\theta|E) = cL(\theta; E)g(\theta) \tag{16.2.3}$$

where c is a constant with respect to θ, and is chosen so that the total probability in the posterior distribution is 1:

$$\frac{1}{c} = \begin{cases} \sum_{\theta\in\Omega} L(\theta; E)g(\theta) & \text{if } \theta \text{ is discrete;} \\ \int_{-\infty}^{\infty} L(\theta; E)g(\theta)\,d\theta & \text{if } \theta \text{ is continuous.} \end{cases} \tag{16.2.4}$$

The posterior p.f. or p.d.f. is thus proportional to the product of the likelihood function and the prior p.f. or p.d.f. of θ.

EXAMPLE 16.2.1. Consider the inheritance of hemophilia as discussed previously in Example 3.6.3. Suppose that a woman has n sons, of whom x are hemophilic and $n-x$ are normal. The probability of this event is

$$P(x;\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{16.2.5}$$

where θ is the probability that a particular son will be hemophilic. The problem is to make inferences about θ.

Given no additional information about θ, inferences would be based on (16.2.5). One could graph the relative likelihood function of θ, or compute confidence intervals. However, it may be possible to extract some information about θ by examining the woman's family tree. For instance, suppose that the woman had normal parents, but she had a brother who was hemophilic. Then her mother must have been a carrier, and she therefore had a 50% chance of inheriting the gene for hemophilia. If she did inherit the gene, then there is a 50% chance that a particular son will inherit the disease ($\theta=\frac{1}{2}$), and if she did not, all of her sons will be normal ($\theta=0$). (The possibility of a mutation is ignored in order to simplify the example.) The prior probability distribution of θ is thus given by

$$g(0)=P(\theta=0)=\tfrac{1}{2};\ g(\tfrac{1}{2})=P(\theta=\tfrac{1}{2})=\tfrac{1}{2}.$$

With this additional information, it is now possible to base the analysis on Bayes's Theorem.

By (16.2.3), the posterior probability function of θ is given by

$$f(\theta|x)=c(x)\theta^x(1-\theta)^{n-x}\tfrac{1}{2} \qquad \text{for } \theta=0,\tfrac{1}{2}.$$

If $x>0$, then $\theta=0$ and $\theta=\frac{1}{2}$ have posterior probabilities 0 and 1, respectively. If $x=0$, the posterior probabilities are

$$P(\theta=0|X=0)=c/2;\ P(\theta=\tfrac{1}{2}|X=0)=c/2^{n+1}.$$

Since the sum of these must be 1, we find that $c=2^{n+1}/(2^n+1)$, and hence that

$$P(\theta=0|X=0)=\frac{2^n}{2^n+1};\ P\left(\theta=\frac{1}{2}\middle|X=0\right)=\frac{1}{2^n+1}.$$

If the woman has at least one hemophilic son ($x>0$), she must be a carrier. If she has only normal sons ($x=0$), the probability that she is a carrier decreases as n increases.

EXAMPLE 16.2.2. Suppose that components are received from a manufacturer in large batches, and let θ denote the proportion of defectives in a batch. A random sample of n items is chosen from the batch, and is found to contain x defectives. If n is small in comparison with the batch size, the probability of x defectives in the sample is

$$P(x;\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{16.2.6}$$

Given no additional information, inferences about θ would be based on (16.2.6).

It may be that similar batches are received at regular intervals from the same manufacturer. The value of θ will vary somewhat from batch to batch. If the manufacturing process is reasonably stable, one might expect the variation in θ to be random, and introduce the assumption that θ is a random variable with probability density function g, say. Data from past samples would be used to help determine the form of the prior density function g.

An assumption which makes the mathematics easy is that θ has a beta distribution with parameters p and q,

$$g(\theta)=k\theta^{p-1}(1-\theta)^{q-1} \qquad \text{for } 0<\theta<1, \tag{16.2.7}$$

where $k=\Gamma(p+q)/\Gamma(p)\Gamma(q)$. Then, by (16.2.3), the posterior p.d.f. of θ given x is

$$f(\theta|x)=c(x)\theta^{x+p-1}(1-\theta)^{n-x+q-1} \qquad \text{for } 0<\theta<1,$$

which is also a beta distribution with parameters $x+p$ and $n-x+q$. Probabilities can be computed by numerical integration, or from tables of the F-distribution (see Problem 6.10.12).

Of course, it would be unwise to assume (16.2.7) merely because it leads to simple mathematics. Data from past samples should be used to check the adequacy of (16.2.7), and to estimate the parameters p and q. As additional data accumulate, further checks of the model can be made, and more precise estimates of p and q can be obtained. Procedures such as this, in which data are used to give information about both the current value of θ and the prior distribution of θ, are called *empirical Bayes* methods. □

In the two preceding examples, it was natural to regard the value of θ as having been generated by a repeatable experiment. Prior probabilities for θ-values then correspond to the relative frequencies with which the various θ-values would be expected to arise in many repetitions of the experiment. It is possible, conceptually at least, to verify the prior distribution empirically by actually repeating the experiment to obtain a sample $\theta_1, \theta_2, \ldots, \theta_n$ of θ-values. These values could be compared with the assumed prior distribution. However, the analysis would usually be complicated by the fact that only estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ were available.

Applications such as these, in which the prior distribution is the probability model of a physical process which generates the value of θ, are not controversial. However, Bayesian methods are sometimes advocated in situations where θ is thought of as a constant. The prior distribution may be

an objective summary of the prior state of knowledge concerning θ, or it may be a statement of an individual's subjective beliefs about θ. There are differences of opinion among statisticians concerning the appropriateness of Bayesian methods in such situations.

Fiducial Prior Distributions

It may be that the conditions for the fiducial argument were satisfied in some previous experiment involving the same parameter θ. The fiducial distribution of θ from the previous experiment might then be used as the prior distribution of θ in the current experiment.

EXAMPLE 16.2.3. Suppose that, in a previous experiment, N components with exponentially distributed lifetimes were tested until failure. From Example 16.1.4, the fiducial distribution of the mean lifetime θ has p.d.f.

$$g(\theta) = \frac{1}{\theta\Gamma(N)}\left(\frac{t}{\theta}\right)^N e^{-t/\theta} \qquad \text{for } \theta > 0,$$

where t is the total of the observed lifetimes. In the current experiment, n additional components are tested simultaneously, and testing stops after a predetermined time period T. From Section 9.5, the likelihood function of θ based on the current experiment is

$$L(\theta) = \theta^{-m} e^{-s/\theta} \qquad \text{for } \theta > 0,$$

where m is the number of components which were observed to fail, and s is the total elapsed lifetime of all n components (including those whose failure times were censored). By (16.2.3), the p.d.f. of the posterior distribution of θ is

$$f(\theta) = c\theta^{-m} e^{-s/\theta} \cdot \theta^{-N-1} e^{-t/\theta}$$

$$= c\theta^{-(m+N+1)} e^{-(s+t)/\theta} \qquad \text{for } \theta > 0.$$

It can now be shown by change of variables that $2(s+t)/\theta$ has a χ^2 distribution with $2(m+N)$ degrees of freedom. Hence tables of the χ^2 distribution may be used to obtain the posterior probabilities of statements about θ.

Note that it would not be possible to derive a fiducial distribution for θ on the basis of the current experiment, or on the basis of the previous and current experiments combined. In each case the minimally sufficient statistic is two dimensional, and there exists no ancillary statistic.

If there were no censoring in the second experiment, the two experiments could be combined to give a single experiment in which $N+n$ components were tested to failure. A fiducial distribution for θ could then be derived as in

Example 16.1.4. The same result would be obtained by taking the fiducial distribution of θ from the previous experiment as the prior distribution in Bayes's Theorem. However, the latter procedure seems inappropriate because it violates the symmetry between the two experiments, and it may lead to unacceptable results in more complicated situations. For further discussion, see D.A. Sprott, "Necessary restrictions for distributions a posteriori", *Journal of the Royal Statistical Society, B*, 22(1960), pages 312–318.

Prior Distributions which Represent Ignorance

Various attempts have been made to formulate prior probability distributions which represent a state of total ignorance about the parameter (see H. Jeffreys, *Theory of Probability*, 3rd edition, Oxford: Clarendon Press, 1961). These are generally derived from arguments of mathematical symmetry and invariance.

Let us consider the simplest case, in which nothing is known about a parameter θ except that it must take one of a finite set of values $\{1, 2, \ldots, N\}$. It might be argued that, since there is no reason to prefer one of these values over another, they should be assigned equal probabilities (Laplace's Principle of Insufficient Reason). The statement that the N possible parameter values are equally probable is then supposed to represent a complete lack of knowledge of θ.

The above argument implicitly assumes that there exists *some* probability distribution which appropriately represents total ignorance. If this assumption is granted, then the assignment of equal probabilities seems inevitable. However, the assumption itself is questionable. It would seem more reasonable to represent prior ignorance by *equally likely*, rather than equally probable, parameter values. If the N parameter values are equally probable, then $P(\theta \neq 1) = (N-1)/N$, and this would seem to be an informative statement. However, no such statement is possible if they are assumed to be equally likely, because likelihoods are not additive.

Now consider a parameter θ which can take values in a real interval $0 < \theta < 1$, say. Great difficulties arise in trying to formulate a probability distribution of θ which represents total ignorance. If one assumes that the distribution of θ is uniform, then one-to-one functions of θ will generally not have uniform distributions because of the Jacobian involved in continuous change of variables. If θ is totally unknown, then presumably θ^3 is also totally unknown, but it is impossible to have a uniform distribution on both of them. This problem does not arise if prior ignorance is represented by equally likely parameter values, because likelihoods are invariant under one-to-one parameter transformations.

For further discussion, see Chapter 1 of *Statistical Methods and Scientific Inference* by R.A. Fisher (2nd edition, New York: Hafner, 1959).

Subjective Prior Distributions

In yet another approach to the use of Bayes's Theorem, the prior distribution is taken to be a summary of an individual's prior belief about θ. See, for example, H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*, Boston: Harvard Univ. Graduate School of Bus. Admin., 1961; and L.J. Savage, *The Foundations of Statistical Inference*, London: Methuen, 1962. According to the advocates of this approach, the prior distribution for θ is to be determined by introspection, and is a measure of personal opinion concerning what the value of θ is likely to be. Bayes's Theorem is then used to modify opinion on the basis of the experimental data.

Any statistical analysis involves some elements of subjective judgement — for instance, in the choice of the probability model. Nevertheless, this subjective input is open to public scrutiny and possible modification if poor agreement with the data is obtained. The same is not true of a subjective prior distribution, which is entirely a personal matter. A subjective prior distribution may be based on nothing more than hunches and feelings, and it seems a mistake to give it the same weight in the analysis as information obtained from the experimental data. The subjective Bayesian approach may prove to be valuable in personal decision problems, but it does not seem appropriate for problems of scientific inference.

16.3. Prediction

Suppose that we wish to predict the value of a random variable Y whose probability distribution depends upon a parameter θ. We assume that θ is unknown, but that a previous set of data gives some information about the value of θ. In predicting Y, we have two types of uncertainty to contend with: uncertainty due to random variation in Y, and uncertainty due to lack of knowledge of θ. We wish to make statements about Y which incorporate both types of uncertainty.

For example, suppose that the lifetimes of a certain type of rocket component are exponentially distributed with mean θ. We have tested n components, and have observed their lifetimes $x_1, x_2, \ldots, x_n$. We wish to predict the lifetime of another component, or perhaps the lifetime of a system made up of several such components. Even if we knew θ, we could not make exact predictions because lifetimes are subject to random variation; that is, components run under identical conditions will generally have different lifetimes. The problem is further complicated by the fact that we do not know the value of θ, but have only limited information obtained from the n components tested. Both the randomness of Y and the uncertainty about θ will influence predictive statements about Y.

Throughout the discussion, we assume that the probability model is

appropriate. Mathematical models are only approximate descriptions of reality, and predictions based on them may be wildly in error if they are poor approximations. Errors of this kind are potentially the most serious, and in many situations it is difficult to estimate how large they are likely to be. Although we can and should check the agreement of the model with the past data, we cannot check the agreement with the future values which we are trying to predict.

Prediction problems have tidy solutions in the special case where all of the available information about θ can be summarized in the form of a probability distribution for θ (fiducial or Bayesian posterior). Suppose that θ has probability density function f, and that Y has p.d.f. $g(y; \theta)$ depending upon θ. If we interpret the latter as the conditional p.d.f. of Y given θ, the joint p.d.f. of Y and θ is $g(y; \theta)f(\theta)$. We then integrate out θ to obtain the marginal p.d.f. of Y,

$$p(y) = \int_{-\infty}^{\infty} g(y; \theta) f(\theta) d\theta. \tag{16.3.1}$$

This distribution combines uncertainty due to random variation in Y with uncertainty due to lack of knowledge of θ, and is called the *predictive distribution* of Y.

Prediction problems are more difficult when there is no probability distribution for θ. A procedure which is sometimes useful in this situation will be discussed in Section 16.5.

Predicting an $(n+1)$st Observation from an Exponential Distribution

Suppose that n independent values are observed from an exponential distribution with unknown mean θ. We wish to predict the value of Y, an $(n+1)$st observation to be taken from the same exponential distribution.

The fiducial argument is applicable in this case. From Example 16.1.4, the fiducial p.d.f. of θ based on the observed sample is

$$f(\theta) = \frac{1}{\theta\Gamma(n)}\left(\frac{t}{\theta}\right)^n e^{-t/\theta} \qquad \text{for } \theta > 0,$$

where $t = \Sigma x_i$ is the observed sample total. Given θ, the p.d.f. of Y is

$$g(y; \theta) = \frac{1}{\theta} e^{-y/\theta} \qquad \text{for } y > 0.$$

By (16.3.1), the p.d.f. of the predictive distribution of Y is

$$p(y) = \int_0^{\infty} \frac{1}{\theta} e^{-y/\theta} \cdot \frac{1}{\theta\Gamma(n)}\left(\frac{t}{\theta}\right)^n e^{-t/\theta} d\theta \qquad \text{for } y > 0.$$

Upon substituting $u = (y + t)/\theta$ and simplifying, we obtain

$$p(y) = \frac{t^n}{\Gamma(n)(t+y)^{n+1}} \int_0^\infty u^n e^{-u} du.$$

The integral on the right equals $\Gamma(n+1)$, and hence by (2.1.14),

$$p(y) = \frac{t^n}{(t+y)^{n+1}} \cdot \frac{\Gamma(n+1)}{\Gamma(n)} = \frac{nt^n}{(t+y)^{n+1}} \qquad \text{for } y > 0.$$

Integrating with respect to y now gives

$$P(Y \le y) = \int_0^y p(v)dv = 1 - \left(\frac{t}{t+y}\right)^n \qquad \text{for } y > 0,$$

and probabilities of statements about Y can easily be obtained. These probabilities take into account both the random variation of Y and the available information about θ.

In Example 9.4.1 we considered $n = 10$ observed lifetimes with total $t = 288$, and in this case

$$P(Y \le y) = 1 - \left(\frac{288}{288+y}\right)^{10} \qquad \text{for } y > 0.$$

We use this to make predictive statements about the lifetime Y of another component of the same type. For instance, we obtain

$$P(Y \le 5) = 0.158, \qquad P(Y \ge 75) = 0.099$$

and so on. Also, we find that

$$P(Y \le 1.48) = P(Y \ge 100.6) = 0.05.$$

The interval $1.48 \le Y \le 100.6$ is called a 90% *predictive interval* for Y. As one might expect, the interval is quite wide, indicating that we cannot predict the lifetime of a single component Y with much precision.

It is of some interest to compare the above results with what we could obtain if we knew the value of θ. If we assume that θ is equal to its maximum likelihood estimate, we have

$$P(Y \le y | \theta = 28.8) = 1 - e^{-y/28.8} \qquad \text{for } y > 0.$$

From this we obtain

$$P(Y \le 1.48) = P(Y \ge 86.3) = 0.05.$$

The central 90% interval is $1.48 \le Y \le 86.3$, which is not much narrower than the 90% predictive interval. This indicates that most of the uncertainty in predicting Y is due to the random variation of Y rather than to lack of information about the value of θ.

Predicting a Future Value from a Normal Distribution

Suppose that we wish to predict a future value of Y, where $Y \sim N(\alpha, c_1)$ with c_1 known. Suppose further that α is unknown, but that all available information concerning α is summarized in the (fiducial or Bayesian) distribution $\alpha \sim N(\hat{\alpha}, c_2)$ where $\hat{\alpha}$ and c_2 are known. Then by (16.3.1), the predictive distribution of Y has p.d.f.

$$p(y) = \frac{1}{2\pi\sqrt{c_1 c_2}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2c_1}(y-\alpha)^2 - \frac{1}{2c_2}(\alpha - \hat{\alpha})^2\right\} d\alpha.$$

This integral may be evaluated by completing the square in the exponent to produce a normal integral. After a bit of algebra, we find that $p(y)$ is the p.d.f. of a normal distribution with mean $\hat{\alpha}$ and variance $c_1 + c_2$. Hence the predictive distribution is

$$Y \sim N(\hat{\alpha}, c_1 + c_2).$$

An easier way to obtain this result is to write $Y \equiv \alpha + \sqrt{c_1}Z_1$ where $Z_1 \sim N(0, 1)$, and $\alpha \equiv \hat{\alpha} + \sqrt{c_2}Z_2$ where $Z_2 \sim N(0, 1)$, independently of Z_1. Combining these gives

$$Y \equiv \hat{\alpha} + \sqrt{c_1}Z_1 + \sqrt{c_2}Z_2$$

where $\hat{\alpha}$, c_1 and c_2 are known constants. Now (6.6.6) and (6.6.7) give $Y \sim N(\hat{\alpha}, c_1 + c_2)$ as before.

EXAMPLE 16.3.1. Suppose that we have already observed n independent measurements $x_1, x_2, \ldots, x_n$ from $N(\mu, \sigma^2)$ with σ known, and that we wish to predict the average value $\bar{Y}$ of m future observations from the same distribution. From Example 16.1.3, the fiducial distribution of μ based on the x_i's is $\mu \sim N(\bar{x}, \sigma^2/n)$. The sampling distribution of $\bar{Y}$ is $\bar{Y} \sim N(\mu, \sigma^2/m)$. Hence by the discussion above, the predictive distribution is

$$\bar{Y} \sim N\left(\bar{x}, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right).$$

This distribution combines uncertainty due to lack of knowledge of μ with uncertainty due to random variation in $\bar{Y}$. If $n \to \infty$, then $\bar{x} \approx \mu$. The uncertainty due to lack of knowledge of μ is then negligible, and the predictive distribution becomes the sampling distribution of $\bar{Y}$. On the other hand, if $m \to \infty$, then uncertainty due to random variation in $\bar{Y}$ becomes negligible, and the predictive distribution becomes the fiducial distribution of μ.

If σ is also unknown, we can integrate over its fiducial distribution as well to obtain

$$\frac{\bar{Y} - \bar{x}}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{(n-1)}$$

where $s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$ (see Section 16.4).

EXAMPLE 16.3.2. Suppose that the straight line model (13.5.1) has been fitted to n observed pairs (x_i, y_i), $i = 1, 2, \ldots, n$. We now wish to predict the value Y of the dependent variable when the independent variable has value x. For instance, in Example 13.5.1 we might wish to predict the systolic blood pressure Y of a particular woman aged $x = 50$ years.

If σ is known, the argument preceding the last example may be applied. The sampling distribution of Y is $N(\mu, \sigma^2)$ where $\mu = \alpha + \beta x$. One can argue that $\hat{\mu} = \hat{\alpha} + \hat{\beta}x$ carries all of the relevant information about μ. From Section 13.6, we have $\hat{\mu} \sim N(\mu, c\sigma^2)$ where

$$c = \frac{1}{n} + (x - \bar{x})^2/S_{xx}.$$

Hence, from Example 16.1.3, the fiducial distribution of μ is $N(\hat{\mu}, c\sigma^2)$. It now follows that the predictive distribution is

$$Y \sim N(\hat{\mu}, (1 + c)\sigma^2).$$

If σ is unknown, we replace σ^2 by $s^2 = \frac{1}{n-2}\Sigma\hat{\varepsilon}_i^2$ to get

$$T \equiv \frac{Y - \hat{\mu}}{\sqrt{(1 + c)s^2}} \sim t_{(n-2)}.$$

A central 99% predictive interval for Y is then

$$Y \in \hat{\mu} \pm a\sqrt{s^2\left[1 + \frac{1}{n} + (x - \bar{x})^2/S_{xx}\right]}$$

where $P\{|t_{(n-2)}| \leq a\} = 0.99$.

For instance, in Example 13.5.1, the central 99% predictive interval for the blood pressure of an individual woman aged 50 years is

$$Y \in 137.68 \pm 23.18.$$

From Section 13.6, a 99% confidence interval for the mean blood pressure of all women aged 50 years is

$$\mu \in 137.68 \pm 6.55.$$

The interval for Y is much wider than the interval for μ, because there is considerable variability in systolic blood pressure among women of the same age. Even if we knew μ exactly, we could not predict the value of Y very precisely.

16.4. Inferences from Predictive Distributions

Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent $N(\mu_i, \sigma^2)$, and that the μ_i's are linear functions of q unknown parameters $\alpha, \beta, \gamma, \ldots$, where $q < n$. This is the normal linear model (see Sections 13.1 and 13.2).

It can be argued that, if σ is known, then $\hat{\alpha}$ carries all of the relevant information about α. The sampling distribution of $\hat{\alpha}$ is $N(\alpha, c\sigma^2)$ where c is a constant. If σ is known, inferences about α are based on this distribution.

The sampling distribution of $\hat{\alpha}$ depends on σ, and so we cannot use it for inferences about α when σ is unknown. Instead we shall derive a predictive distribution for $\hat{\alpha}$ which does not depend on σ, and then use the predictive distribution for inferences about α.

Let $V \equiv \Sigma\hat{\varepsilon}_i^2$ denote the residual sum of squares for the linear model. Then V carries all of the relevant information about σ, and

$$U \equiv V/\sigma^2 \sim \chi^2_{(n-q)},$$

independently of $\hat{\alpha}$.

U is a pivotal quantity which satisfies the conditions for the fiducial argument. To obtain the fiducial distribution of σ, we replace V by its observed value $v = (n - q)s^2$, giving

$$\sigma^2 \equiv (n - q)s^2/U \qquad \text{where } U \sim \chi^2_{(n-q)}.$$

Now, by (16.3.1), the p.d.f. of the predictive distribution of $\hat{\alpha}$ given s is

$$p(\hat{\alpha}; \alpha, s) = \int_0^\infty g(\hat{\alpha}; \alpha, \sigma) f(\sigma; s) d\sigma.$$

We can avoid having to evaluate this integral by using (6.10.1). We have

$$\hat{\alpha} \equiv \alpha + Z\sqrt{\sigma^2 c}$$

where $Z \sim N(0, 1)$, independently of U. Substituting for σ^2 gives

$$\hat{\alpha} \equiv \alpha + Z\sqrt{c(n - q)s^2/U} \equiv \alpha + T\sqrt{s^2 c}$$

where $T \equiv Z \div \sqrt{U/(n - q)} \sim t_{(n-q)}$ by (6.10.1). Hence predictive statements for $\hat{\alpha}$ given s are obtained from

$$T \equiv \frac{\hat{\alpha} - \alpha}{\sqrt{s^2 c}} \sim t_{(n-q)}. \tag{16.4.1}$$

This result appears to be identical to (13.2.7), but it is not. In (13.2.7), s^2 is a random variable such that $(n - q)s^2/\sigma^2 \sim \chi^2_{(n-q)}$, whereas in (16.4.1) s^2 is the particular observed variance estimate.

In this problem, s^2 plays the role of an ancillary statistic. Since its sampling distribution does not depend upon α, s^2 give no direct information about the magnitude of α. However, its observed value indicates the informativeness or precision of the experiment with respect to α. By the arguments of Section 15.3, s^2 should be held fixed at its observed value in making inferences about α. Thus it would seem appropriate to use the predictive distribution (16.4.1) rather than the sampling distribution (13.2.7) in making inferences about α.

In fact, one will obtain the same numerical values for significance levels and confidence intervals whether one uses (16.4.1) or (13.2.7), and so the

distinction does not matter in this case. It does matter in more complicated cases, such as the Behrens–Fisher problem to be considered below.

Note that (16.4.1) defines a pivotal quantity T which satisfies the conditions set out for the fiducial argument in Section 16.1. Thus (16.4.1) can be used to obtain a fiducial distribution for α when σ is unknown.

Behrens–Fisher Problem

Suppose that we have $n+m$ independent measurements made using two different techniques which may not be equally precise. The n measurements made with the first technique are modelled as $N(\mu_1, \sigma_1^2)$, and the m measurements made with the second technique as $N(\mu_2, \sigma_2^2)$. We wish to make inferences about $\mu_1 - \mu_2$.

If $\sigma_1 = \sigma_2$, we have just the two-sample model discussed in Section 13.4. A similar analysis to that in Section 13.4 is possible if $\sigma_1 = k\sigma_2$ where k is a known constant. However, if the ratio σ_1/σ_2 is unknown, the analysis becomes difficult and controversial. The problem of making inferences about $\mu_1 - \mu_2$ when σ_1/σ_2 is unknown is called the Behrens–Fisher problem.

The MLE of $\mu_1 - \mu_2$ is $\bar{y}_1 - \bar{y}_2$, the difference between the two samples means. Its sampling distribution is

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N(\mu_1 - \mu_2, c_1\sigma_1^2 + c_2\sigma_2^2)$$

where $c_1 = 1/n$ and $c_2 = 1/m$. If σ_1 and σ_2 were known, inferences about $\mu_1 - \mu_2$ would be based on this distribution.

Since the sampling distribution of $\hat{\mu}_1 - \hat{\mu}_2$ depends on σ_1 and σ_2, we cannot use it for inferences about $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown. Instead we shall derive a predictive distribution for $\hat{\mu}_1 - \hat{\mu}_2$ given the observed sample variances s_1^2 and s_2^2. This predictive distribution does not depend upon σ_1 or σ_2, and it can be used for inferences about $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown.

From (16.4.1), the predictive distributions of $\hat{\mu}_1$ and $\hat{\mu}_2$ are given by

$$\hat{\mu}_1 \equiv \mu_1 + T_1\sqrt{s_1^2 c_1}; \qquad \hat{\mu}_2 \equiv \mu_2 + T_2\sqrt{s_2^2 c_2}$$

where $T_1 \sim t_{(n-1)}$ and $T_2 \sim t_{(m-1)}$. T_1 and T_2 are independent because the first sample is assumed to be independent of the second. Hence the predictive distribution of $\hat{\mu}_1 - \hat{\mu}_2$ given s_1^2 and s_2^2 is given by

$$\begin{aligned}\hat{\mu}_1 - \hat{\mu}_2 &\equiv \mu_1 - \mu_2 + T_1\sqrt{s_1^2 c_1} - T_2\sqrt{s_2^2 c_2} \\ &\equiv \mu_1 - \mu_2 + T\sqrt{c_1 s_1^2 + c_2 s_2^2}\end{aligned}$$

where T is a linear combination of T_1 and T_2:

$$T \equiv T_1\sqrt{\frac{c_1 s_1^2}{c_1 s_1^2 + c_2 s_2^2}} - T_2\sqrt{\frac{c_2 s_2^2}{c_1 s_1^2 + c_2 s_2^2}}.$$

The distribution of a linear combination

$$T \equiv T_1 \cos\theta - T_2 \sin\theta,$$

where $T_1 \sim t_{(\nu_1)}$ and $T_2 \sim t_{(\nu_2)}$ are independent, is called the *Behrens–Fisher distribution*. It is tabulated in the Fisher and Yates *Statistical Tables for Biological, Agricultural and Medical Research*. In this case we have

$$\tan\theta = \frac{\sin\theta}{\cos\theta} = \sqrt{\frac{c_1 s_2^2}{c_2 s_1^2}}$$

so that θ is a function of the observed variance ratio s_2^2/s_1^2.

When σ_1 and σ_2 are unknown, inferences about $\mu_1 - \mu_2$ may be based on the pivotal quantity

$$T \equiv \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{c_1 s_1^2 + c_2 s_2^2}}$$

which is referred to tables of the Behrens–Fisher distribution with parameters ν_1, ν_2, and θ, where $\nu_1 = n-1$, $\nu_2 = m-1$, and

$$\theta = \tan^{-1}\sqrt{\frac{c_1 s_1^2}{c_2 s_2^2}}.$$

A similar result can be derived for inferences about any linear combination of μ_1 and μ_2.

Note that the Behrens–Fisher distribution arises in connection with the *predictive* distribution of $\hat{\mu}_1 - \hat{\mu}_2$, in which s_1 and s_2 are held fixed at their observed values. Alternatively, one might consider the *sampling* distribution of

$$T' \equiv \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{c_1 S_1^2 + c_2 S_2^2}}$$

where now S_1^2 and S_2^2 are independent random variables such that

$$(n-1)S_1^2/\sigma_1^2 \sim \chi^2_{(n-1)}; \qquad (m-1)S_2^2/\sigma_2^2 \sim \chi^2_{(m-1)}.$$

The distribution of T' in repeated sampling is *not* Behrens–Fisher, and it will depend upon the unknown variance ratio σ_1^2/σ_2^2.

Many statisticians are of the opinion that inferences should always be based on sampling distributions, so that probabilities can be interpreted directly as relative frequencies in repetitions of the experiment. As a result, they do not accept the above solution based on the predictive distribution of $\hat{\mu}_1 - \hat{\mu}_2$. On the other hand, it seems appropriate that S_1^2 and S_2^2 should be treated as ancillary statistics and held fixed at their observed values as is the case in the above solution. If one insists that S_1^2 and S_2^2 be treated as ancillary statistics, then inferences cannot be based on a sampling distribution.

16.5. Testing a True Hypothesis

Sometimes one can generate intervals of "reasonable" values for an unknown quantity y by the device of testing a true hypothesis H. One assumes a value for y, carries out a test of significance, and finds the significance level SL(y). A small significance level indicates an inconsistency, and if H is known to be true, doubt is cast on the value assumed for y. One can define a 95% interval or region as the set of all values of y such that SL(y) ≥ 0.05. Several examples will be given to illustrate this procedure.

EXAMPLE 16.5.1. Suppose that $\bar{X} \sim N(\mu_1, \sigma^2/n)$ and $\bar{Y} \sim N(\mu_2, \sigma^2/m)$ independently of $\bar{X}$, where σ is known. Given observed values of $\bar{X}$ and $\bar{Y}$, we can test H: $\mu_1 - \mu_2$ using the result that $\bar{X} - \bar{Y} \sim N\left(0, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$ under the hypothesis. The significance level will be 5% or more if and only if

$$-1.96 \leq \frac{\bar{x} - \bar{y}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \leq 1.96. \tag{16.5.1}$$

Now suppose that we don't know $\bar{y}$ but we do know that $\mu_1 = \mu_2$. This would be the case if $\bar{Y}$ were the average of m future observations to be taken from the same $N(\mu, \sigma^2)$ distribution as the original sample $x_1, x_2, \ldots, x_n$ (see Example 16.3.1). Now (16.5.1) yields a 95% interval for $\bar{y}$:

$$\bar{y} \in \bar{x} \pm 1.96\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}.$$

This interval consists of the values of $\bar{Y}$ such that a test of the true hypothesis $\mu_1 = \mu_2$ would produce a significance level of 5% or more. The same interval can be obtained as the central 95% interval in the predictive distribution of $\bar{y}$ (see Example 16.3.1).

EXAMPLE 16.5.2. Let Y_1 and Y_2 be independent variates, with $Y_i \sim$ binomial (n_i, p_i). The likelihood ratio statistic for testing H: $p_1 = p_2$ is

$$D = 2\Sigma y_i \log \frac{y_i}{n_i \tilde{p}} + 2\Sigma(n_i - y_i) \log \frac{n_i - y_i}{n_i(1 - \tilde{p})}$$

where $\tilde{p} = (y_1 + y_2)/(n_1 + n_2)$. Given observed values y_1 and y_2, we can test H as in Section 12.4. The distribution of D is approximately $\chi^2_{(1)}$ if H is true, and so

$$\text{SL} \approx P\{\chi^2_{(1)} \geq D_{\text{obs}}\}.$$

Alternatively, an exact conditional significance level can be calculated as in Section 15.6.

If we don't know y_2 but we do know that $p_1 = p_2$, we can use the above test to construct a range of plausible values for y_2. For instance, suppose that 12 successes are observed in 20 Bernoulli trials, and that we wish to predict y_2, the number of successes in 30 future trials with the same success probability. Taking $n_1 = 20$, $n_2 = 30$, and $y_1 = 12$, we can compute D for selected values of y_2. We find that $D \leq 3.841$, and hence SL ≥ 0.05, for $10 \leq y_2 \leq 25$. Values of y_2 outside this interval are implausible in that they would lead to a significance level less than 0.05 in a test of the true hypothesis $p_1 = p_2$.

EXAMPLE 16.5.3. Suppose that a lake contains n_1 tagged fish and n_2 untagged fish, where n_1 is known but n_2 is unknown. Fish are caught during a predetermined period of time, and the catch is observed to consist of x tagged fish and y untagged fish. What can be concluded about n_2?

We assume $n_1 + n_2$ independent trials, and let the probability that a particular fish is caught during the time period be p_1 for tagged fish and p_2 for untagged fish. Given a value for n_2, we can test H: $p_1 = p_2$ as in the preceding example. If we are willing to assume that H is true, then a small significance level casts doubt on the value chosen for n_2.

For instance, suppose that there are 110 tagged fish in the lake, and that the sample contains 20 tagged fish and 478 untagged fish. Taking $n_1 = 110$, $y_1 = 20$, and $y_2 = 478$, we can compute D for selected values of n_2. We find that $D \leq 3.841$, and hence SL ≥ 0.05, for $1817 \leq n_2 \leq 4097$.

Note that, in order to derive this range of values for n_2, it is necessary to assume that $p_1 = p_2$. This assumption may not be appropriate, because fish that have been caught and tagged may have a larger (or smaller) probability of being caught subsequently.

EXAMPLE 16.5.4. Suppose that the normal straight line model (13.5.1) has been fitted to n observed pairs (x_i, y_i), $i = 1, 2, \ldots, n$. Now we observe an additional value y, for which the corresponding value of the independent variable x is unknown. The problem is to make inferences about x.

Suppose that a value is given for x. Then from the straight line model, the estimated mean value of Y is

$$\hat{\mu} \equiv \hat{\alpha} + \hat{\beta}x \sim N(\mu, c\sigma^2)$$

where $c = c(x) = (1/n) + (x - \bar{x})^2/S_{xx}$. We take $Y \sim N(\mu', \sigma^2)$ and test H: $\mu' = \mu$. Under H, we have

$$Y - \hat{\mu} \sim N(0, (1 + c)\sigma^2)$$

and the significance level is 5% or more if and only if

$$-1.96 \leq \frac{y - \hat{\mu} - 0}{\sqrt{\sigma^2(1 + c)}} \leq 1.96.$$

Substituting for $\hat{\mu}$ and c and squaring gives

$$[y - \hat{\alpha} - \hat{\beta}x]^2 \leq 1.96^2\sigma^2\left[1 + \frac{1}{n} + (x - \bar{x})^2/S_{xx}\right].$$

The 95% interval for x is the set of x-values for which this quadratic inequality is satisfied. For any value of x outside this interval, a test of the hypothesis $E(Y) = \alpha + \beta x$ will give a significance level less than 5%.

For σ unknown, we replace $1.96^2\sigma^2$ by t^2s^2, where $s^2 = \frac{1}{n-2}\Sigma\hat{\varepsilon}_i^2$ and $P\{-t \le t_{(n-2)} \le t\} = 0.95$, giving

$$[y - \hat{\alpha} - \hat{\beta}x]^2 \le t^2 s^2 \left[1 + \frac{1}{n} + (x - \bar{x})^2/S_{xx}\right]. \tag{16.5.2}$$

If we take x as known and y as unknown in this inequality, we get the central 95% predictive interval for a future observation Y (see Example 16.3.2). For y known but x unknown, we get a 95% interval for x. This interval consists of all values of x such that the observed value of the observation Y belongs to its 95% predictive interval.

If the slope β were zero, then an observed y would not determine a value of x. Thus one might anticipate difficulties when the estimated slope $\hat{\beta}$ is not significantly different from zero. A test of H: $\beta = 0$ is based on

$$\frac{\hat{\beta} - 0}{\sqrt{s^2/S_{xx}}} \sim t_{(n-2)},$$

and the condition for $\hat{\beta}$ to be significantly different from zero at the 5% level is

$$\hat{\beta}^2 > t^2 s^2 / S_{xx}$$

where $P\{-t \le t_{(n-2)} \le t\} = 0.95$. If this condition is not satisfied, the 95% interval for x will be either the entire real line, $-\infty < x < \infty$, or else the entire real line with a finite interval removed. These results can be derived by examining the discriminant and the sign of the second-degree term in the quadratic inequality (16.5.2).

The problem of estimating the independent variable x given a value of the dependent variable y is called the *calibration problem*. It arises when the quantity x which is of interest cannot be measured directly, but we must make do with a measurement y which is related to x. The equation relating $E(Y)$ to x is called the calibration curve. In this case we have assumed a linear calibration curve $E(Y) = \alpha + \beta x$ with independent $N(0, \sigma^2)$ errors.

APPENDIX A

Answers to Selected Problems

9.1.1 $\hat{\lambda} = \Sigma x_i/n = 0.25$

9.1.4 $\hat{\theta} = (2x_1 + x_2)/2n = 0.55$; exp. freq. 30.25, 49.50, 20.25

9.1.7 $L(N) = N^{-n}$ for $N \ge$ largest sample value $x_{(n)}$, and $L(N) = 0$ otherwise. $L(N)$ decreases as N increases, so $\hat{N} = x_{(n)}$.

9.1.9 $L(\theta) = \prod p_i^{f_i}$ where $p_i = \theta^{i-1}(1-\theta)$ for $i = 1, 2, 3$ and $p_4 = \theta^3$.
$\hat{\theta} = (f_2 + 2f_3 + 3f_4)/(f_1 + 2f_2 + 3f_3 + 3f_4) = 0.5$
Exp. freq. 100, 50, 25, 25; poor agreement with obs. freq.

9.1.13 $L(\mu) = \prod p_i^{f_i}$ where $p_i = \mu^i e^{-\mu}/i!(1 - e^{-\mu})$ for $i = 1, 2, \ldots$. $l'(\mu) = 0$ gives equation for $\hat{\mu}$; $\hat{\mu} = 3.048$.

9.2.1 $L(\theta) = \theta^m(1-\theta)^{M-m}\theta^{2n}(1-\theta^2)^{N-n}$
$l'(\theta) = 0$ gives $(M + 2N)\theta^2 + (M - m)\theta - (m + 2n) = 0$. $\hat{\theta}_1 = 0.11$, $\hat{\theta}_2 = \sqrt{0.02}$, $\hat{\theta} = 0.12$.

9.3.1 50% LI (0.168, 0.355); 10% LI (0.116, 0.460)

9.3.4 $l(\theta) = 54 \log(\frac{1}{6} + \theta) + 46 \log(\frac{1}{6} - 2\theta)$; $l(\hat{\theta}) = -175.74$
$r(0) = l(0) - l(\hat{\theta}) = -3.44$; $R(0) = 0.032$. It is quite unlikely that $\theta = 0$.

9.3.7 $R(b) = 1$ for $b = 9, 10$ (b is not unique)
$R(b) \ge 0.5$ for $b = 6, 7, \ldots, 17$; $R(b) \ge 0.1$ for $b = 5, 6, \ldots, 32$.

9.3.10 $r(\lambda) = 51 \log \lambda - 20\lambda + 3.259$; 10% LI $1.86 \le \lambda \le 3.39$.
$r(\lambda) = 2 \log p + 18 \log(1 - p) + 6.502$ where $p = e^{-\lambda}$;
10% LI $1.21 \le \lambda \le 4.28$. Yes, since the 10% LI is much wider.

9.4.1 $\hat{\theta} = 76.8$, $r(\theta) = -27 \log \theta - \frac{2074}{\theta} + 144.218$; 10% LI $61.7 \le \theta \le 97.2$. Obs. freq. 11, 8, 6, 2; exp. freq. 12.92, 6.74, 5.35, 2.00. Yes, the agreement is good.

9.4.4 $L(\theta) = \theta^{-2n} \exp(-\Sigma x_i/\theta)$; $\hat{\theta} = \frac{1}{2n}\Sigma x_i$; $R(\theta) = L(\theta)/L(\hat{\theta})$.

9.4.6 $l(\mu) = -[\Sigma(x_i - \mu)^2 + \Sigma(y_i - \mu/2)^2]/2\sigma^2$; $l'(\mu) = 0$ gives
$\hat{\mu} = (n\bar{x} + \frac{1}{2}m\bar{y})/(n + \frac{1}{4}m)$.

Expt 2: Y_i is Poisson with mean $100p$;
$l(p) = \Sigma Y_i \log p - 100pn$; $\mathcal{I}_E(p) = \dfrac{100n}{p}$.

11.7.2 $\text{MSE}(T_1) = E(Y^4) - \mu^2(2+\mu^2)$ since $E(Y^2) = 1 + \mu^2$.
$\text{MSE}(T_2) = E(Y^4) - (1-\mu^2)^2 = \text{MSE}(T_1) - 1$.

11.7.7 $T \equiv (\Sigma w_i X_i)/(\Sigma w_i)$ where $w_i = \sigma_i^{-2}$.

12.1.1 $X \equiv$ number tall $\sim$ binomial (100, p); H: $p = 0.75$
$D \equiv |X - 75|$; $D_{\text{obs}} = 10$; SL $= P\{|X - 75| \geq 10 | p = 0.75\}$. Exact SL $= 0.0275$ from summing binomial probabilities; approx. SL $\approx P\{|Z| \geq 2.31\} = 0.0209$, or SL $\approx P\{|Z| \geq 2.19\} = 0.0282$ with continuity correction.

12.1.3 $X \equiv$ # times benzedrine is faster $\sim$ binomial (20, p)
H: $p = 0.5$; $D \equiv |X - 10|$; $D_{\text{obs}} = 4$
SL $= P\{|X - 10| \geq 4 | p = 0.5\} = 0.115$. Results do not strongly contradict H: $p = 0.5$.

12.1.5 $X \equiv$ # seedlings from one packet $\sim$ binomial (100, p)
H: $p = 0.8$; $D = |X - 80|$; $D_{\text{obs}} = 7$ for 1st customer.
SL $= P\{|X - 80| \geq 7 | p = 0.8\} = 0.103$. No customer has a strong case against merchant. $T \equiv$ Total # seedlings $\sim$ binomial (400, p)
H: $p = 0.8$; $D \equiv |T - 320|$; $D_{\text{obs}} = 20$
SL $= P\{|T - 320| \geq 20 | p = 0.8\} = 0.015$
Combined results give strong evidence that $p \neq 0.8$.

12.1.9 $f(x) = (5 - x)/15$; $E(X) = 4/3$; $\text{var}(X) = 14/9$. $E(\bar{X}) = E(X) = 4/3$;
$\text{var}(\bar{X}) = \text{var}(X)/100 = 14/900$; $D \equiv |\bar{X} - 4/3|$; $\bar{X}_{\text{obs}} = 1.6$; $D_{\text{obs}} = 4/15$
SL $= P\{|\bar{X} - 4/3| \geq 4/15\} \approx P\{|Z| \geq 2.138\} = 0.0325$
Yes, $\bar{X}_{\text{obs}}$ is significantly larger than expected.

12.1.11 $T \equiv$ # calls in 5 hours $\sim$ Poisson (5λ). H: $\lambda = 7.2 \Rightarrow T \sim$ Poisson (36). $T_{\text{obs}} = 21$.
$$\text{SL} = P\{T \leq T_{\text{obs}} | \lambda = 7.2\} = \sum_{x=0}^{21} 36^x e^{-36}/x! = 0.0049$$
Strong evidence that $\lambda < 7.2$.

12.2.1 $l(\theta) = 65 \log \theta + 35 \log(1 - \theta)$; $\hat{\theta} = 0.65$
$D_{\text{obs}} = 2[l(0.65) - l(0.75)] = 4.95$; SL $\approx P(\chi^2_{(1)} \geq 4.95) = 0.0261$

12.2.5 $l(p) = \Sigma f_j \log p_j$; $\hat{p}_j = f_j/556$; $l(\hat{p}) = -619.586$
$l(p_0) = -619.824$; $D_{\text{obs}} = 2[-619.586 + 619.824] = 0.475$
SL $\approx P(\chi^2_{(3)} \geq 0.475) = 0.92$; no evidence against hypothesis.

12.2.8 $l(\theta) = -n \log \theta - \Sigma X_i/\theta$; $\hat{\theta} = \bar{X}$; $D = 2[l(\hat{\theta}) - l(\theta_0)]$
$\hat{\theta} = 43.3$, $l(\hat{\theta}) = -47.705$
H: $\theta = 37.4$; $l(\theta_0) = -47.821$; $D_{\text{obs}} = 0.23$; SL ≈ 0.63
H: $\theta = 65.1$; $l(\theta_0) = -48.426$; $D_{\text{obs}} = 1.44$; SL ≈ 0.23
Data are consistent with both hypotheses.
95% CI (14.7% LI) $24.76 \leq \theta \leq 86.55$.

12.3.1 -105.493; $-105.743(\tilde{\theta} = 0.55)$; -106.745
$D_{\text{obs}} = 0.5$; d.f. $= 2 - 1 = 1$; SL ≈ 0.48. $D_{\text{obs}} = 2.00$; d.f. $= 1 - 0 = 1$; SL ≈ 0.157.
$D_{\text{obs}} = 2.50$; d.f. $= 2 - 0 = 2$; SL ≈ 0.286

12.3.5 $l(\mu) = \Sigma X_i \log \mu_1 - n\mu_1 + \Sigma Y_i \log \mu_2 - m\mu_2$; $\hat{\mu}_1 = \bar{X}$, $\hat{\mu}_2 = \bar{Y}$.
$D = 2[l(\hat{\mu}) - l(\tilde{\mu})]$ where $\tilde{\mu}_1 = \tilde{\mu}_2 = (n\bar{X} + m\bar{Y})/(n + m)$.
$n = 12$, $m = 15$, $\hat{\mu}_1 = \dfrac{11}{12}$, $\hat{\mu}_2 = \dfrac{31}{15}$, $\tilde{\mu}_1 = \tilde{\mu}_2 = \dfrac{42}{27}$

$D_{\text{obs}} = 5.98$; d.f. $= 2 - 1 = 1$; SL ≈ 0.014
Strong evidence against H: $\mu_1 = \mu_2$.

12.3.9 $\tilde{\theta} = 9.6706$; $D_{\text{obs}} = 0.854$; d.f. $= 3 - 1 = 2$; SL ≈ 0.65.
No evidence of heterogeneity.
95% CI $9.74 \pm 1.96/0.563^{1/2}$ etc.
Overall $9.6706 \pm 1.96/(0.563 + 0.345 + 0.695)^{1/2}$.

12.4.3 (a) $\tilde{p}_i = 82/400$, $n\tilde{p}_i = 20.5$, $n(1 - \tilde{p}_i) = 79.5$ for $1 \leq i \leq 4$, $D_{\text{obs}} = 12.897$, d.f. $= 4 - 1 = 3$, SL ≈ 0.005
(b) $\tilde{p}_i = 72/300$ for $i = 1, 2, 3$; $\tilde{p}_4 = 10/100$. $D_{\text{obs}} = 2.757$, d.f. $= 4 - 2 = 2$, SL ≈ 0.25
(c) $D_{\text{obs}} = 12.897 - 2.757 = 10.140$, d.f. $= 2 - 1 = 1$, SL ≈ 0.0015. No evidence of difference among drugs; strong evidence of difference between drugs and placebo.

12.4.6 $\tilde{\alpha} = -0.0002126$, $\tilde{\beta} = -0.03297$, $\tilde{p} = \exp(\tilde{\alpha} + d\tilde{\beta})$

$n\tilde{p}$	4537.07	3726.28	3374.24	2547.54	1933.03
$n(1 - \tilde{p})$	0.93	125.72	230.76	265.46	272.97

$D_{\text{obs}} = 3.53$, d.f. $= 5 - 2 = 3$, SL ≈ 0.32
No evidence against hypothesis.

12.5.2 Exp. freq. $26306 \binom{12}{i}\left(\frac{1}{3}\right)^i\left(\frac{2}{3}\right)^{12-i}$; Pool last 2 classes. $D_{\text{obs}} = 38.15$; d.f. $= (12 - 1) - 0 = 11$; SL < 0.001. MLE of $P(5 \text{ or } 6)$ is $\hat{p} = 0.3377$. Using this value rather than $p = \frac{1}{3}$ gives $D_{\text{obs}} = 11.13$ with 10 d.f. Dice appear to be biased in favor of 5 or 6.

12.5.6 $\tilde{\theta} = 0.55$; exp. freq. 30.25 49.5 20.25
$D_{\text{obs}} = 1.24$; d.f. $= (3 - 1) - 1 = 1$; SL ≈ 0.27
$D_{\text{obs}} = 4.94$; d.f. $= 1$; SL ≈ 0.026. The larger sample provides stronger evidence against the model.

12.5.10 (a) $\tilde{\mu} \approx 83/60$; exp. freq. 15.0 20.8 14.4 6.6 3.1 (≥ 4)
A good fit.
(b) $\tilde{\mu} \approx 136/60$; exp. freq. 6.2 14.1 16.0 12.1 6.8 3.1 1.7 (≥ 6)
$D_{\text{obs}} = 24.7$; d.f. $= (7 - 1) - 1 = 5$; SL < 0.001.
Injuries tend to occur in clusters, not individually.
(c) Exp. freq. 16.6; $D_{\text{obs}} = 9.67$; d.f. $= (5 - 1) - 0 = 4$; SL ≈ 0.046. Some evidence against hypothesis.

12.5.13 $P(i \text{ digits between 0's}) = (0.9)^i(0.1)$ for $i = 0, 1, 2, \ldots$

Class	0	1–2	3–4	5–6	7–9	10–13	14–19	20–∞
Exp. freq.	5.0	8.55	6.93	5.61	6.48	6.00	5.36	6.08
Obs. freq.	6	13	8	4	5	4	5	5

$D_{\text{obs}} = 4.20$; d.f. $= (8 - 1) - 0 = 7$; SL ≈ 0.76. No evidence against the model.
(Other groupings are possible.)

12.5.16 $\tilde{p} = (0.1n + 0.05n + 0.12n)/3n = 0.09$
$$D_{\text{obs}} = 2n\left[0.1 \log \frac{0.1}{0.9} + 0.05 \log \frac{0.05}{0.09} + \cdots + 0.88 \log \frac{0.88}{0.91}\right] = 0.03418n;$$
d.f. $= 3 - 1 = 2$
$P(\chi^2_{(2)} \geq 5.991) = 0.05$ so $n \approx 5.991/0.03418 = 175$

12.5.18 $p_j = P(j \text{ siblings}) = (j - 1)a_{j-1}/\Sigma i a_i$; exp. freq. is np_j.
Exp. freq. 34.21 51.41 47.31 ... 3.91 (≥ 9)
$D_{\text{obs}} = 32.33$; d.f. $= (10 - 1) - 0 = 9$; SL < 0.001.

9.4.10 $L(c) = e^{nc}$ for $0 < c \le$ smallest sample value; $L(c)$ increases with c.
$R(c) = \exp\{n(c - x_{(1)})\}$ for $0 < c < x_{(1)}$;
$100p\%$ LI: $x_{(1)} + (1/n) \le \log p \le c \le x_{(1)}$.

9.4.12 $L(\theta) = \theta^{-n}$ for $\theta \le x_1, x_2, \ldots, x_n \le 2\theta$; i.e. for $\theta \le x_{(1)}$ and $x_{(n)} \le 2\theta$. Since $L(\theta)$ is decreasing, $\hat\theta = \frac{1}{2}x_{(n)}$; $R(\theta) = (\hat\theta/\theta)^n$ for $\hat\theta \le \theta \le x_{(1)}$.

9.5.2 $L(\theta) = p^k(1-p)^{n-k}$ where $p = 1 - e^{-T/\theta}$;

$$\hat\theta = T/\log\left(\frac{n}{n-k}\right);\ R(\theta) = L(\theta)/L(\hat\theta).$$

9.5.4 $f(x) = 4x\theta^{-2}e^{-2x/\theta}$; $L(\theta) = \theta^{-2m}e^{-2\Sigma x_i/\theta}(1 + 2T/\theta)^{n-m}e^{-2T(n-m)/\theta}$
$l'(\theta) = 0$ gives quadratic equation for $\hat\theta$.
$\hat\theta = 1813.42$; 10% LI (1382, 2449)

9.6.3 $P(T \le 100t) = 1 - e^{-100t/\theta} = 1 - \beta^t$
$L(\beta) = p_1^{29}p_2^{22} \ldots p_7^8 = \beta^{234}(1-\beta)^{92}(1+\beta)^{19}$ for $0 < \beta < 1$. $l'(\beta) = 0$ gives quadratic equation for $\hat\beta$. $\hat\beta = 0.7245$ and $\hat\theta = -100/\log\hat\beta = 310.3$ 10% LI $0.670 \le \beta \le 0.775 \leftrightarrow 249 \le \theta \le 392$
50% LI $0.695 \le \beta \le 0.753 \leftrightarrow 275 \le \theta \le 352$.

9.7.3 10% LI is $\bar x \pm \sigma((2/n)\log 10)^{1/2}$. Width is at most 2 for $n \ge 2\sigma^2 \log 10$.

9.8.1 $\hat\mu = 1.5936$; $l(\hat\mu) = -21.7233$

9.8.4 $1.3185 \le \mu \le 1.6184$

9R2 $L(\alpha) = (1-\alpha)^{50}[\alpha(1-\alpha)]^{23}[\alpha^2(1-\alpha)]^{14}[\alpha^3(1-\alpha)]^8[\alpha^4]^5$;
$\hat\alpha = \frac{1}{2}$; 10% LI is $0.4226 \le \alpha \le 0.5774$
Exp. freq. 50 25 12.5 6.25 6.25; a good fit.

9R6 $L(\theta) = (1/\theta)e^{-132/\theta}\cdot(1/\theta)e^{-768/\theta}\cdot[e^{-1000/\theta}]^3$
$l(\theta) = -2\log\theta - 3900/\theta$; $\hat\theta = 1950$
$\hat\phi = 1 - e^{-100/\theta} = 1 - e^{-1/19.5} = 0.0500$.

9R9 $l(\theta) = 72\log\theta + 160\log(1-\theta) + 20\log(3-2\theta)$;
$l'(\theta) = 0$ gives a quadratic equation. $\hat\theta = 0.2954$; exp. freq. 49.64 29.33 21.03.

10.1.3 $l(\mu, \sigma) = -n\log\sigma - \Sigma a_i(y_i - \mu)^2/2\sigma^2$

$$\hat\mu = \hat\mu(\sigma) = \Sigma a_i y_i/\Sigma a_i;\ \hat\sigma^2 = \frac{1}{n}\Sigma a_i(y_i - \hat\mu)^2$$

$\mathscr{I}(\hat\mu, \hat\sigma)$ is a diagonal matrix with positive diagonal elements $\Sigma a_i/\hat\sigma^2$ and $2n/\hat\sigma^2$.

10.1.5 If $f(x) = \theta^{x-1}(1-\theta)$ for $x = 1, 2, \ldots$ then $E(X) = (1-\theta)^{-1}$.
$E\{X_1 + X_2 + \cdots + X_m\} = m(1-\theta)^{-1}$. Since $\alpha > \beta$, $m(1-\alpha)^{-1} > m(1-\beta)^{-1}$.

$$\hat\alpha = \frac{a}{a+m},\ \hat\beta = \frac{b}{b+m}$$ where a, b are numbers of successes for treatments A and B.

10.1.8 $L(\lambda, c) = \lambda^n \exp\{-\lambda\Sigma t_i\}\cdot\exp\{n\lambda c\}$ for $0 < c \le t_{(1)}$ and $\lambda > 0$. $\hat c = t_{(1)}$;
$\hat\lambda = 1/(\bar t - t_{(1)})$.

10.1.10 $l(p, \lambda) = m\log p + \lambda\Sigma t_i - m\log\lambda + (n-m)\log(1 - p + pe^{-\lambda T})$. Solve $\partial l/\partial p$ for p as a function of λ. Substitute into $\partial l/\partial\lambda = 0$ and solve for λ.

10.2.2 $\hat\alpha = 3/8$, $\hat\beta = 11/24$; $\hat\phi = \hat\alpha + \hat\beta = 5/6$; $\hat p = \hat\alpha/\hat\phi = 9/20$.
Exp. freq. 15 11 9 29 (a perfect fit).
$r = 44\log\alpha + 15\log(1-\alpha) + 11\log\beta + 29\log(1 - \alpha - \alpha\beta) + 81.744$.
10% max LI $0.61 \le \phi \le 1$, $0.28 \le p \le 0.68$.

10.3.2 $$R_{\max}(\lambda) = \left(\frac{1}{1+\lambda}\right)^x\left(\frac{\lambda}{1+\lambda}\right)^y\Big/\left(\frac{x}{x+y}\right)^x\left(\frac{y}{x+y}\right)^y;$$
$\hat\lambda = 57/47$; $0.80 \le \lambda \le 1.86$. $R_{\max}(1) = 0.62$. $\lambda = 1$ (no change in rate) is very plausible.

10.3.6 $R_{\max}(\lambda) = (\lambda/\hat\lambda)^n \exp\{n - n\lambda/\hat\lambda\}$ for $0 < \lambda < \infty$;
$R_{\max}(c) = (\bar t - t_{(1)})^n(\bar t - c)^{-n}$ for $0 < c \le t_{(1)}$.

10.3.8 $L_{\max}(\lambda) = \lambda^m(1 - e^{-\lambda T})^{-m}\exp\{-\lambda\Sigma t_i\}$; $R_{\max}(\lambda) = L_{\max}(\lambda)/L_{\max}(\hat\lambda)$.

10.4.2 $\hat{\mathscr{I}}^{22} = 0.057097$; $0.7000 \le \lambda \le 1.7255$
$\hat{\mathscr{I}}_*^{22} = 0.038820$; $0.7946 \le \lambda \le 1.8510$
Exact 10% interval $0.7956 \le \lambda \le 1.8593$. Yes, the logarithmic transformation helps.

10.5.2 ML eqns. $\Sigma(n_i - a_i) = 0$, $\Sigma(n_i - a_i)d_i = 0$
2nd deriv. $-\Sigma b_i$, $-\Sigma b_i d_i$, $-\Sigma b_i d_i^2$, where $a_i = (n_i - x_i)/(1-p_i)$; $b_i = a_i p_i/(1-p_i)$.
$\hat\alpha = -0.0002126$, $\hat\beta = -0.03297$
$-0.0354 \le \beta \le -0.0306$; $0.99896 \le e^\alpha \le 0.999992$.

10.5.3 $l(\alpha, \beta) = \Sigma x_j\log\alpha + \Sigma jx_j\log\beta - \alpha\Sigma\beta^j$. ML equations $\Sigma x_j - \hat\alpha\Sigma\hat\beta^j = 0$; $\Sigma jx_j - \hat\alpha\Sigma j\hat\beta^j = 0$. Substitute $\hat\alpha = \Sigma x_j/\Sigma\hat\beta^j$ in 2nd equation; solve numerically for $\hat\beta$.

11.2.3 $c = \sqrt{-2\log 0.2} = 1.794$; $CP = P(-1.794 \le Z \le 1.794) = 0.927$
$P(-2.242 \le Z \le 2.242) = 0.975$, so $p = \exp\{-\frac{1}{2}(2.242)^2\} = 0.081$.

11.3.2 P.d.f. of $U \equiv 2\lambda X^2$ is

$$f\left(\sqrt{\frac{u}{2\lambda}}\right)\frac{d}{dy}\left(\sqrt{\frac{u}{2\lambda}}\right) = \frac{1}{2}e^{-u/2} \text{ for } u > 0.$$

Thus $U \sim \chi^2_{(2)}$ and $U_1 + U_2 + \cdots + U_n \sim \chi^2_{(2)}$.

$$P(D \le d) = P\left(Y - 1 - \log Y \le \frac{d}{2n}\right)$$ where $Y \sim \frac{1}{2n}\chi^2_{(2)}$ as in Example 11.3.3.

11.4.3 $r(p) = 300\log p + 200\log(1-p) + 336.506$; $\hat p = 0.6$
95% CI (14.7% LI) $0.5567 \le p \le 0.6423$
99% CI (3.6% LI) $0.5428 \le p \le 0.6554$
$\mathscr{I}(\hat p) = 2083.3$; 0.6 ± 0.0429; 0.6 ± 0.0564

11.4.4 $r(\theta) = -10\log\theta - 388/\theta + 46.584$; $\hat\theta = 38.8$
90% CI (25.8% LI) $24.03 \le \theta \le 68.60$; $p = e^{-50/\theta}$; $0.1248 \le p \le 0.4825$

11.4.7 $r(\sigma) = -15\log\sigma - 26.1026/2\sigma^2 + 11.655$; $\hat\sigma = 1.319$
95% CI (14.7% LI) $0.958 \le \sigma \le 1.979$

11.4.11 $L(\theta) = \theta^{-n}$ for $\theta \ge M$; $\hat\theta = M$. $P(M \le m) = [P(X_i \le m)]^n = (m/\theta)^n$;
$P(D \le d) = P(M \ge \theta e^{-d/2n}) = 1 - (e^{-d/2n})^n$;
$CP(\theta_0) = P(D \le -2\log p) = 1 - e^{-(-2\log p)/2}$;
95% CI (5% LI) $0.9537 \le \theta \le 1.2868$.

11.5.3 $a = 1$, $b = 2$; $\hat\gamma = 1.3201$
$\hat{\mathscr{I}}_*^{11} = \hat{\mathscr{I}}^{11} + 4\hat{\mathscr{I}}^{12} + 4\hat{\mathscr{I}}^{22} = 0.04891$
$\gamma \pm 2.576(\mathscr{I}_*^{11})^{1/2} = 1.3201 \pm 0.5697$
$p = e^\gamma/(1 + e^\gamma)$; $\hat p = 0.7892$; $0.6793 \le p \le 0.8687$.

11.5.6 $\gamma = \log m = \log\theta + \frac{1}{\beta}\log(\log 2)$; $a = 0.012213$, $b = 0.082944$; $\hat\gamma = 4.2309$
$\hat{\mathscr{I}}_*^{11} = 0.01366$
95% CI $4.0018 \le \gamma \le 4.4600$, or $54.70 \le m \le 86.49$.

11.6.1 $\mathscr{I}_E(\theta) = \dfrac{2n}{\theta(1-\theta)}$ in (a); $\dfrac{12n}{(1-\theta)(1+\theta)}$ in (b). Do (b) if it is thought that $\theta > 0.2$.

11.6.3 Expt 1: $l(p) = \Sigma Y_i\log p + \Sigma(Y_i - X_i)\log(1-p)$; $\mathscr{I}_E(p) = \dfrac{100n}{p(1-p)}$.

Strong evidence against hypothesis.
Adjusted frequencies f'_j are not multinomial.

12.5.19 $L(p_1 \dots p_4) = p_1^8(1-p_1)^{12}p_2^{70}(1-p_2)^{90} \dots p_4^{184}(1-p_4)^{200}$
$\hat{p}_1 = 8/20$, $\hat{p}_2 = 70/160$, etc., so $l(\hat{p}) = -586.791$.
Under H, $\tilde{p}_i = \dfrac{8+70+\cdots}{20+160+\cdots} = \dfrac{1}{2}$; $l(\tilde{p}) = -590.561$
$D_{obs} = 2[l(\hat{p}) - l(\tilde{p})] = 7.54$; d.f. $= 4-1 = 3$;
SL ≈ 0.057. Weak evidence against hypothesis.
Exp. freq. 10 10; 20 40 20; ...; 6 24 36 24 6
$D_{obs} = 14.27$; d.f. $= (1+2+3+4)-1 = 9$; SL ≈ 0.113.
In (c), estimate p separately for each litter;
$D_{obs} = 14.27 - 7.54 = 6.73$; d.f. $= 10-4 = 6$.

12.6.1
10(6.03)	117(120.97)
3(6.07)	144(140.03)

$D_{obs} = 5.31$; d.f. $= (2-1)(2-1) = 1$; SL ≈ 0.021. Fairly strong evidence against independence hypothesis.

12.6.5
74(85)	116(123.25)	68(63.75)	82(68)
126(115)	174(166.75)	82(86.25)	78(92)

$D_{obs} = 8.70$; d.f. $= (2-1)(4-1) = 3$; SL ≈ 0.034. Some evidence against independence hypothesis.

12.6.7 (a) $D_{obs} = 1.84$; d.f. $= (3-1)(3-1) = 4$; SL ≈ 0.77
(b) Obs. freq. 126 271 132; Exp. freq. 132.25 264.5 132.25
$D_{obs} = 0.46$; d.f. $= (3-1)-0 = 2$; SL ≈ 0.8
No evidence against hypothesis in (a) or (b).

12.6.11 (a)
Above	9(15)	6(8)	12(9)	23(18)
Below	21(15)	10(8)	6(9)	13(18)

$D_{obs} = 10.8$; d.f. $= 3$; SL ≈ 0.013. Strong evidence that birth weight is not independent of parental smoking habits.

(b)
	MF	$M\bar{F}$		$\bar{M}F$	$\bar{M}\bar{F}$
Above	9(9.8)	6(5.2)	Above	12(11.7)	23(23.3)
Below	21(20.2)	10(10.8)	Below	6(6.3)	13(12.7)

Given mother's smoking habits, there is no evidence that birth weight depends on father's smoking habits.

12.6.13 Test for independence in 2×10 table gives $D = 37.5$ (9 d.f.), which shows that insects tend to aggregate. This analysis is conditional on the total number of insects which land on area A or B in each trial.

12.7.3 Test for independence in 2×4 table gives $D = 66$ (3 d.f.); SL ≈ 0. Strong evidence against independence hypothesis. It may well be that only the best students chose to write the competition. There is no proof that writing the competition made them any better.

12.7.5 $D = 112$; d.f. $= 1$; SL ≈ 0. The admission rate is certainly lower for females. Only program A shows any evidence of bias, and here it appears to be against males. There are proportionately more female applicants to programs with low admission rates.

12.8.3 (a) Each of the 400 electors is counted twice in the table, so rows are not independent.
(b) $D = 256$; d.f. $= 1$; SL ≈ 0. Most electors have not changed their positions.
(c) Consider just those who changed their positions. Obs. freq. 17 33; exp. freq. 25 25 (assuming no change in overall support).
$D = 5.21$; d.f. $= 1$; SL ≈ 0.022. Evidence of a loss in support for the government.

13.3.3 $\bar{y} = 61.21$, $s^2 = 14.74$ (11 d.f.); $\mu \in 61.21 \pm 2.44$
14.7% LI $7.08 \le \sigma^2 \le 38.85$
$3.816 \le 11s^2/\sigma^2 \le 21.92$ gives $7.40 \le \sigma^2 \le 42.48$.

13.3.4 Assuming n large 95% CI for μ has width $2(1.96)(\sigma^2/n)^{1/2}$. For width 2, $n = (1.96)^2\sigma^2$. Variance estimate $s^2 = 80/9$, so $n \approx 34$. Advise about 25 additional measurements.

13.3.7 (a) $\bar{y} = 12.9$, $s^2 = 138.9$; $P\{Y<0\} \approx P\{Z < (0-\bar{y})/s\} = 0.14$ from Table B2.
(b) $\bar{y} = 2.082$, $s^2 = 27.67$. Negative counts are impossible.

13.4.2 $s^2 = 4.018$ (25 d.f.); $\mu_1 \in 6.23 \pm 2.06s/\sqrt{11}$

$$\mu_2 \in 12.74 \pm 2.06s/\sqrt{16};\ \mu_1 - \mu_2 \in -6.51 \pm 2.06s \Big/ \sqrt{\frac{1}{11}+\frac{1}{16}}$$

13.4.6 (a) $s_1^2 = 0.0914$, $\nu_1 = 13$; $s_2^2 = 0.0422$, $\nu_2 = 3$;
$s^2 = 0.082175$ (16 d.f.). $D_{obs} = \Sigma \nu_i \log(s^2/s_i^2) = 0.616$ (1 d.f.); SL ≈ 0.43. No evidence against equal variance hypothesis.

(b) $\mu_1 - \mu_2 \in 0.273 \pm 1.746\left[0.082175\left(\dfrac{1}{14}+\dfrac{1}{4}\right)\right]^{1/2}$.

13.4.9 (a) $s_1^2 = 0.25$, $s_2^2 = 0.268$, $s_3^2 = 0.183$, $s^2 = 0.267$; d.f. $= 4, 4, 4, 12$.
$D_{obs} = 4\Sigma \log(s^2/s_i^2) = 0.491$ (2 d.f.); SL ≈ 0.78.
No evidence against equal variance hypothesis.
(b) $s^2 = 0.267$ (12 d.f.); 14.7% LI $0.132 \le \sigma^2 \le 0.671$.
Or $4.404 \le 12s^2/\sigma^2 \le 23.34$ gives $0.137 \le \sigma^2 \le 0.728$.
(c) $\mu_2 - \mu_3 \in -0.3 \pm 2.179[0.267(\frac{1}{4}+\frac{1}{4})]^{1/2}$.

13.5.2 $\hat{\alpha} = 1.468$, $\hat{\beta} = 1.703$, $s^2 = 0.00502$ (28 d.f.). Plot shows curvature. Try a 2nd degree polynomial model.

13.5.6 $\hat{\alpha} = 47.864$, $\hat{\beta} = 48.247$, $s^2 = 0.1783$ (15 d.f.). Estimated boiling points 192.40 203.16 211.96. $\hat{\beta}$ would increase to 48.247log10; other results unchanged. One point (BP = 204.6) is seriously out of line. Redo analysis with this point omitted.

13.6.2 $\hat{\alpha} = 0.976$, $\hat{\beta} = 0.353$, $s^2 = 0.00978$ (6 d.f.). The last observation is somewhat larger than expected.
$\beta \in 0.353 \pm 2.447(s^2/1.229)^{1/2}$
$\alpha + 0.4\beta \in 1.117 \pm 2.447s(\frac{1}{8} + 0.2875^2/1.229)^{1/2}$

13.6.7 (a) $\hat{\alpha} = -0.228$, $\hat{\beta} = 0.9948$, $\Sigma\hat{\epsilon}_i^2 = 0.3419$, $s^2 = 0.04273$ (8 d.f.).
(b) $T_{obs} = (\hat{\beta}-1)/(s^2/1569)^{1/2} = -1.00$; SL $= 0.35$.
(c) $T_{obs} = \hat{\alpha}/s\left(\dfrac{1}{10} + 23.25^2/1569\right)^{1/2} = -1.65$; SL $= 0.14$.
No evidence against H: $\beta = 1$ or H: $\alpha = 0$.
(d) $\Sigma x^2 = 6974.25$, $\Sigma xy = 6884.65$, $\Sigma y^2 = 6796.66$; $\hat{\beta} = 0.9872$;
$s^2 = \frac{1}{9}(\Sigma y^2 - \hat{\beta}\Sigma xy) = 0.05099$ (9 d.f.); $T_{obs} = (\hat{\beta}-1)/(0.05099/6974)^{1/2} = -4.75$; SL < 0.001. There is very strong evidence that $\beta \neq 1$. If we insist that line goes through the origin, slope must be less than 1 to give a reasonable fit to the data. It is reasonable to take $\beta = 1$ or to take $\alpha = 0$, but it is not satisfactory to assume both $\beta = 1$ and $\alpha = 0$.

13.7.3 We assume that differences are independent $N(\mu, \sigma^2)$, and test H: $\mu = 0$. Here $\bar{y} = 2.5$, $s^2 = 5.10$ (5 d.f.).

$T_{obs} = (\bar{y} - 0)/(s^2/6)^{1/2} = 2.71$; SL = 0.042.
There is some evidence that brand A is superior. However, brand A tires were always tested first. It would be better to test A first on 3 randomly chosen cars, and B first on the other 3 cars.

13.7.8 (a) $\hat{\alpha} = -0.1605$, $\hat{\beta} = 35.348$, $s^2 = 0.01031$ (31 d.f.)

$$\alpha + 0.4\beta \in 3.3743 \pm 2.04s\left[\frac{1}{33} + 0.02764^2/0.07242\right]^{1/2} = 3.3743 \pm 0.0419$$

(b) Same $\hat{\alpha}$, $\hat{\beta}$; $s^2 = 0.00753$ (9 d.f.)

$$\alpha + 0.4\beta \in 3.3743 \pm 2.262s\left[\frac{1}{11} + 0.02764^2/0.02414\right]^{1/2} = 3.3743 \pm 0.0687$$

The interval in (a) is too narrow, owing to the treatment of repeat measurements as independent replicates.

13R2 $s_1^2 = 2.495$ (6 d.f.), $s_2^2 = 2.898$ (6 d.f.), $s^2 = 2.696$ (12 d.f.).
$D_{obs} = 6\Sigma \log(s^2/s_i^2) = 0.03$ (1 d.f.); SL ≈ 0.85.
No evidence against hypothesis of equal variances.

$$\mu_2 - \mu_1 \in 1.729 \pm 2.179s\left(\frac{1}{7} + \frac{1}{7}\right)^{1/2} = 1.729 \pm 1.912.$$

$4.404 \le 12s^2/\sigma^2 \le 23.34$ gives $1.386 \le \sigma^2 \le 7.346$. One possibility: divide 14 men into 7 pairs with nearly equal initial blood pressure. Choose one of pair at random to get drug 1, and the other gets drug 2. Analyze differences.

13R6 $s_1^2 = 0.1983$, $s_2^2 = 0.1820$, $s_3^2 = 0.1692$, $s_4^2 = 0.1722$, each with 8 d.f. $s^2 = 0.1804$ (32 d.f.).
$D_{obs} = 8\Sigma \log(s^2/s_i^2) = 0.06$ (3 d.f.); SL ≈ 0.996.
No evidence against hypothesis of equal variances. In fact, the variance estimates are so nearly equal that one might suspect some tampering with the data.
$18.3 \le 32s^2/\sigma^2 \le 49.5$ gives $0.1166 \le \sigma^2 \le 0.3155$.
Alternatively, 14.7% LI is $0.1147 \le \sigma^2 \le 0.3074$.

14.1.1
$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 0 & 4 \\ 1 & 0 & 5 \\ 1 & 0 & 6 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

14.2.2 $X^tX = \begin{bmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{bmatrix} = \begin{bmatrix} 12 & 628 \\ 628 & 34416 \end{bmatrix}$;
$X^ty = \begin{bmatrix} \Sigma y \\ \Sigma xy \end{bmatrix} = \begin{bmatrix} 1684 \\ 89894 \end{bmatrix}$
$(X^tX)^{-1} = \begin{bmatrix} 1.8495 & -0.0337 \\ -0.0337 & 0.000645 \end{bmatrix}$; $\hat{\beta} = (X^tX)^{-1}X^ty = \begin{bmatrix} 80.89 \\ 1.138 \end{bmatrix}$
$\hat{\beta}^t(X^ty) = 1684 \times 80.78 + 89894 \times 1.138 = 238330$
$\Sigma y^2 = 238822$; $\Sigma\hat{\varepsilon}_i^2 = 238822 - 238330 = 492$.

14.2.5 $X = \begin{bmatrix} 1 & -4 & 16 \\ 1 & -3 & 9 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 16 \end{bmatrix}$; $X^tX = \begin{bmatrix} 9 & 0 & 60 \\ 0 & 60 & 0 \\ 60 & 0 & 708 \end{bmatrix}$; $X^ty = \begin{bmatrix} 755.4 \\ -154.5 \\ 4924.3 \end{bmatrix}$

$\hat{\beta} = (X^tX)^{-1}X^ty$ gives $\hat{\beta}_1 = 86.35$, $\hat{\beta}_2 = -2.575$, $\hat{\beta}_3 = -0.3627$,
$\Sigma\hat{\varepsilon}_i^2 = 63845.72 - 63841.59 = 4.13$ (9 d.f.).
$\hat{\beta}_1 + (17.5 - 19)\hat{\beta}_2 + (17.5 - 19)^2\hat{\beta}_3 = 89.40$
$\frac{d}{dt}[\hat{\beta}_1 + (t - 19)\hat{\beta}_2 + (t - 19)^2\hat{\beta}_3] = 0$ for $t = 19 - \hat{\beta}_2/2\hat{\beta}_3 = 15.45$.

14.3.1 (a) $s^2 = 0.074317/25$ (25 d.f.)
H_1: $Q = 0.066343$ (3 d.f.); SL $= P\{F_{3,25} \ge 7.44\} = 0.001$.
H_2: $Q = 0.000043$ (2 d.f.); SL $= P\{F_{2,25} \ge 0.007\} = 0.993$.
(b) $s^2 = 0.07436/27$ (27 d.f.)
$Q = 0.0663$ (1 d.f.); SL $= P\{F_{1,27} \ge 24.07\} < 0.001$.

14.3.5 (a) $\Sigma\Sigma(y_{ij} - \bar{y}_i)^2 = 4.540$; $s^2 = 0.2389$ (19 d.f.)
$\Sigma\Sigma(y_{ij} - \bar{y})^2 = 6.838$ (22 d.f.); $Q = 2.298$ (3 d.f.)
$F_{obs} = (Q \div 3)/s^2 = 3.21$; SL $= P\{F_{3,19} \ge 3.21\} = 0.046$.
Some evidence that means are not equal.
(b) $s^2 = 0.1324 \quad 0.08 \quad 0.3329 \quad 0.377$; $\nu_i = 6 \quad 3 \quad 6 \quad 4$;
$s^2 = 0.2389$ (19 d.f.).
$D_{obs} = \Sigma\nu_i \log(s^2/s_i^2) = 3.01$ (3 d.f.);
SL $\approx P(\chi^2_{(3)} \ge 3.01) = 0.39$.
No evidence against hypothesis of equal variances.

14.3.8 Straight line model: $\Sigma\hat{\varepsilon}_i^2 = 118.91$ (18 d.f.).
5-sample model: $\Sigma\Sigma\hat{\varepsilon}_{ij}^2 = 117.27$, $s^2 = 7.82$ (15 d.f.).
$Q = 1.64$ (3 d.f.); $F_{obs} = 0.07$; SL $= P\{F_{3,15} \ge 0.07\} = 0.975$.
Straight line model fits very well.

14.4.1 $\hat{\beta}$, V, and s^2 (4 d.f.) are given in Problem 14.2.4.
(a) $\beta_3 \in \hat{\beta}_3 \pm 2.776(s^2v_{33})^{1/2} = 28.675 \pm 2.985$
(b) (i) $\theta = \beta_2 - 2\beta_1 = b^t\beta$ where $b^t = (-2, 1, 0)$.
$\hat{\theta} = -2.375$; $\text{var}(\hat{\theta}) = c\sigma^2$ where $c = b^tVb = 19/8$.
$T_{obs} = (\hat{\theta} - 0)/(s^2c)^{1/2} = -0.88$; SL $= P\{|t_{(4)}| \ge 0.88\} = 0.43$.
(ii) Under H, $\mu = X\beta$ where

$$X^t = \begin{bmatrix} 1 & 2 & 0 & 3 & 1 & 2 & 3 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}; \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_3 \end{bmatrix}$$

New residual SS is 14.71 (5 d.f.).
Old residual SS is 12.335 (4 d.f.); $s^2 = 3.08375$
$Q = 2.375$ (1 d.f.); $F_{obs} = 0.7702$; SL $= P\{F_{1,4} \ge 0.77\} = 0.43$
(c) For revised model, $\hat{\beta}_3 = 28.8$, $v_{33} = 0.3684$, $s^2 = 2.942$ (5 d.f.).
$\beta_3 \in \hat{\beta}_3 \pm 2.571(s^2v_{33})^{1/2} = 28.8 \pm 2.677$.

14.4.3 (a) $\theta = \beta_2 - \beta_1 = b^t\beta$ where $b^t = (0 \quad 1 \quad -1 \quad 0)$. $\hat{\theta} = 2.88$; $\text{var}(\hat{\theta}) = c\sigma^2$ where $c = b^tVb = 0.021$. $s^2 = 4.3$ (9 d.f.).
$\theta \in 2.88 \pm 2.262(cs^2)^{1/2} = 2.88 \pm 0.680$.
(b) $T_{obs} = (\hat{\beta}_3 - 0)/(v_{33}s^2)^{1/2} = -6.88$;
SL $= P\{|t_{(9)}| \ge 6.88\} < 0.001$.
(c) $Q = 167.6 - 38.7 = 128.9$ (2 d.f.). $F_{obs} = (Q \div 2)/s^2 = 14.99$;
SL $= P\{F_{2,9} \ge 14.99\} = 0.0014$.

14.5.6 (a) $m_{i1}^2 + m_{i2}^2 + \cdots + m_{in}^2 = m_{ii}$, so $m_{ii}(1 - m_{ii}) \ge 0$.
(c) If $m_{ii} = 1$, then $m_{ij} = 0$ for $j \ne i$ by result in (a).
$\hat{\mu}_i = m_{i1}y_1 + m_{i2}y_2 + \cdots + m_{in}y_n = y_i$; $\hat{\varepsilon}_i = y_i - \hat{\mu}_i = 0$.

15.1.2 Y is sufficient but not minimal because $L(\sigma; -y) = L(\sigma; y)$.

15.1.5 Yes, the pair (T_1, T_2) is minimally sufficient. $L(\theta; y')$ and $L(\theta; y)$ are proportional if and only if $T_1(y') = T_1(y)$ and $T_2(y') = T_2(y)$.

15.1.7 $f(x_1 \dots x_n) = (1/2\theta)^n$ for $-\theta < x_1, \dots, x_n < \theta$.
$L(\theta; x) = \theta^{-n}$ for $-\theta < x_{(1)}$ and $x_{(n)} < \theta$; i.e. for $\theta > \max\{x_{(n)}, -x_{(1)}\}$
$T \equiv \max\{X_{(n)}, -X_{(1)}\}$ is a sufficient statistic for θ.

15.1.9 $L(\mu) = \exp\{-n(\bar{X}-\mu)^2/2\sigma^2 - m(\bar{Y}-\mu)^2/2k\sigma^2\}$
$= c\exp\{-[(nk+m)\mu^2 - 2(kn\bar{X}+m\bar{Y})]/2k\sigma^2\}$
Thus $T \equiv kn\bar{X} + \bar{Y}$ is sufficient for μ. $T \sim N((kn+m)\mu, k(kn+m)\sigma^2)$ by (6.6.8) and (6.6.7).

15.1.15 $\log\dfrac{\theta}{1-\theta}$, $\log\mu$, $\dfrac{1}{\theta}$ (or any constant multiples of them).

15.2.3 $L(\lambda) = \lambda^n \exp(-\lambda t)$ for $\lambda > 0$ where $t = \Sigma x_i^2$. Thus $T \equiv \Sigma X_i^2$ is a sufficient statistic for λ. P.d.f. of $Y \equiv 2\lambda X^2$ is $\frac{1}{2}e^{-y/2}$, which is $\chi^2_{(2)}$. Now $Z \equiv 2\lambda T \equiv \Sigma 2\lambda X_i^2 \sim \chi^2_{(2n)}$ by (6.9.7). P.d.f. of Z is $g(z)\left|\dfrac{dz}{dt}\right| = k(2\lambda t)^{n-1}e^{-\lambda t}\cdot 2\lambda$.

15.2.5 Problem 15.1.6 only.

15.3.1 $T \equiv X + Y$ is ancillary; base inferences on conditional distribution of $\hat{\theta}$ given observed T, or equivalently, on distribution of X given observed T. This conditional distribution is binomial (t, θ).

15.6.2 See Example 15.6.1.

y_1, y_2	0, 4	1, 3	*2, 2	3, 1	4, 0
$D(y_1, y_2)$	2.02	0.07	2.41	8.04	20.02
$P(y\|t)$	0.376	0.462	0.149	0.013	0.000

SL $= P(D \geq 2.41 | T = 4) = 0.162$.

15.6.3 $P(x|t) = \binom{x_1+r-1}{r-1}\binom{x_2+r-1}{r-1}\Big/\binom{t+2r-1}{2r-1}$
$l(p_1, p_2) = r\log p_1 + x_1\log(1-p_1) + r\log p_2 + x_2\log(1-p_2)$
$D = 2[l(\hat{p}_1, \hat{p}_2) - l(\tilde{p}_1, \tilde{p}_2)]$ where $\hat{p}_i = r/(r+x_i)$ and $\tilde{p}_i = 2r/(2r+t)$.
$r = 2$, $t = 16$, $D_{\text{obs}} = 2.41$; $D \geq 2.41$ for $x_1 = 0, 1, 2, 14, 15, 16$;
SL $= P(D \geq 2.41 | T = 16) = 0.194$.

15.6.6 See Example 15.6.2 (test for independence in 2 × 2 table).
$g(x) = \binom{13}{x}\binom{17}{12-x}\Big/\binom{30}{12}$; $D_{\text{obs}} = 14.02$
SL $= g(0) + g(10) + g(11) + g(12) = 0.00054$.

15.6.10 Under H: $\mu_1 = \mu_2 = 2\mu_3$, $T \equiv X_1 + X_2 + X_3$ is sufficient.
$f(x|t) = \binom{t}{x_1\ x_2\ x_3}\left(\frac{2}{5}\right)^{x_1+x_2}\left(\frac{1}{5}\right)^{x_3}$ where $\Sigma x_i = t$
$D(x) = 2\Sigma x_i \log(x_i/\tilde{\mu}_i)$ where $\tilde{\mu}_3 = \dfrac{t}{5}$; $\tilde{\mu}_1 = \tilde{\mu}_2 = \dfrac{2t}{5}$.
SL $= P(D \geq D_{\text{obs}} | T = t)$.

15.6.12 (a) $T \equiv X_1 + X_4$; $T \sim$ binomial (n, p).
(b) $D \equiv 2T\log\dfrac{2T}{4} + 2(n-T)\log\dfrac{2(n-T)}{n}$
SL $= P(D \geq D_{\text{obs}}) =$ sum of binomial $(n, \frac{1}{2})$ probabilities.
(c) $P(x|t) = \binom{n}{x_1\ x_2\ x_3\ x_4}2^{-n}\Big/\binom{n}{t}$ where $\Sigma x_i = n$ and $x_1 + x_4 = t$.
$D \equiv 2\Sigma X_j \log(X_j/e_j)$ where $e_1 = e_4 = \dfrac{t}{2}$; $e_2 = e_3 = \dfrac{n-t}{2}$.
SL $= P(D \geq D_{\text{obs}} | T = t) \approx P(\chi^2_{(2)} \geq D_{\text{obs}})$.

APPENDIX B

Tables

Tables B1, B2

Standardized normal distribution

$$F(x) = P\{N(0,1) \leq x\} = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du.$$

Table B1 gives the value x whose cumulative probability $F(x)$ is the sum of the corresponding row and column headings. Example: the value x such that $F(x) = .64$ is 0.358 (from row .6 and column .04 of Table B1).

Table B2 gives the cumulative probability $F(x)$, where x is the sum of the corresponding row and column headings. Example: the cumulative probability at value 0.36 is $F(.36) = .6406$ (from row .3 and column .06 of Table B2).

Table B1. Percentiles of the Standardized Normal Distribution

F	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.5	.000	.025	.050	.075	.100	.126	.151	.176	.202	.228
.6	.253	.279	.305	.332	.358	.385	.412	.440	.468	.496
.7	.524	.553	.583	.613	.643	.674	.706	.739	.772	.806
.8	.842	.878	.915	.954	.994	1.036	1.080	1.126	1.175	1.227
.9	1.282	1.341	1.405	1.476	1.555	1.645	1.751	1.881	2.054	2.326

x	1.960	2.576	3.090	3.291	3.891	4.417	4.982
F	.975	.995	.999	.9995	.99995	.999995	.9999995
$2(1-F)$	.05	.01	.002	.001	.0001	.00001	.000001

Source: R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Table I; published by Longman Group Ltd., London (previously published by Oliver and Boyd, Edinburgh); reprinted by permission of the authors and publishers.

Table B2. Standardized Normal Cumulative Distribution Function

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189

Table B2. Standardized Normal Distribution (*continued*)

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.93319	.93448	.93574	.93669	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.990097	.990358	.990613	.990863	.991106	.991344	.991576
2.4	.991802	.992024	.992240	.992451	.992656	.992857	.993053	.993244	.993431	.993613
2.5	.993790	.993963	.994132	.994297	.994457	.994614	.994766	.994915	.995060	.995201
2.6	.995339	.995473	.995604	.995731	.995855	.995975	.996093	.996207	.996319	.996427
2.7	.996533	.996636	.996736	.996833	.996928	.997020	.997110	.997197	.997282	.997365
2.8	.997445	.997523	.997599	.997673	.997744	.997814	.997882	.997948	.998012	.998074
2.9	.998134	.998193	.998250	.998305	.998359	.998411	.998462	.998511	.998559	.998605
3.0	.998650	.998694	.998736	.998777	.998817	.998856	.998893	.998930	.998965	.998999

Source: A. Hald, *Statistical Tables and Formulas* (1952), Table II; reprinted by permission of John Wiley & Sons, Inc.

Table B3. Percentiles of Student's (t) Distribution

$$F(x) = P(t_{(\nu)} \le x) = \int_{-\infty}^{x} \left(1 + \frac{u^2}{\nu}\right)^{-(\nu+1)/2} du \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}.$$

The body of the table gives the values x corresponding to selected values of the cumulative probability (F) and degrees of freedom (ν).

ν \ F	.60	.70	.80	.90	.95	.975	.99	.995	.9995
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657	636.619
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925	31.598
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841	12.924
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604	8.610
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032	6.869
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707	5.959
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499	5.408
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355	5.041
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250	4.781
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169	4.587
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106	4.437
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055	4.318
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012	4.221
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977	4.140
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947	4.073
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921	4.015
17	.257	.534	.863	1.333	1.740	2.110	2.567	2.898	3.965
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878	3.922
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861	3.883
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845	3.850
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831	3.819
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819	3.792
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807	3.767
24	.256	.531	.857	1.318	1.711	2.064	2.492	2.797	3.745
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787	3.725
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779	3.707
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771	3.690
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763	3.674
29	.256	.530	.854	1.311	1.699	2.045	2.462	2.756	2.659
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750	3.646
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704	3.551
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660	3.460
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617	3.373
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576	3.291

Source: R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Table III; published by Longman Group Ltd., London (previously published by Oliver and Boyd, Edinburgh); reprinted by permission of the authors and publishers.

Table B4. Percentiles of the Chi-Square (χ^2) Distribution

$$F(x) = P(\chi^2_{(\nu)} \le x) = \int_0^x u^{\nu/2-1} e^{-u/2}\, du / 2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right).$$

The body of the table gives the values x corresponding to selected values of the cumulative probability (F) and degrees of freedom (ν).

ν \ F	.005	.01	.025	.05	.10	.25	.5	.75	.9	.95	.975	.99	.995	.999
1	$.0^4 3927$	$.0^3 1571$	$.0^3 9821$	$.0^2 3932$	.01579	.1015	.4549	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	.01003	.02010	.05064	.1026	.2107	.5754	1.386	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	.07172	.1148	.2158	.3518	.5844	1.213	2.366	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	.2070	.2971	.4844	.7107	1.604	1.923	3.357	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	.4117	.5543	.8312	1.145	1.160	2.675	4.351	6.626	9.236	11.07	12.83	15.09	16.75	20.52
6	.6757	.8721	1.237	1.635	2.204	3.455	5.348	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	.9893	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	1.344	1.646	2.180	2.733	3.490	5.071	7.344	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.603	3.053	3.816	4.575	5.578	7.584	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	3.074	3.571	4.404	5.226	6.304	8.438	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.565	4.107	5.009	5.892	7.042	9.299	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.075	4.660	5.629	6.571	7.790	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.601	5.229	6.262	7.261	8.547	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70

(continued on following page)

Table B4. Chi-Square Distribution (*continued*)

ν \ F	.005	.01	.025	.05	.10	.25	.5	.75	.9	.95	.975	.99	.995	.999
16	5.142	5.812	6.908	7.962	9.312	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.697	6.408	7.564	8.672	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.265	7.015	8.231	9.390	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.844	7.633	8.907	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.434	8.260	9.591	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	8.034	8.897	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.643	9.542	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.260	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.886	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70

For $\nu > 30$, $\sqrt{\frac{9\nu}{2}}\left\{\left[\frac{1}{\nu}\chi^2_{(\nu)}\right]^{1/3} - 1 + \frac{2}{9\nu}\right\}$ is approximately $N(0,1)$.

Source: E. S. Pearson and H. O. Hartley (editors), *Biometrika Tables for Statisticians, vol. I*, Table 8; Cambridge University Press (3rd edition, 1966); reprinted by permission of the Biometrika Trustees.

Table B5. Percentiles of the Variance Ratio (F) Distribution n Numerator and m Denominator Degrees of Freedom

$$F(x) = P(F_{n,m} \le x) = \int_0^x \left(\frac{n}{m}u\right)^{n/2-1}\left(1+\frac{n}{m}u\right)^{-(n+m)/2}\frac{n}{m}du \cdot \Gamma\left(\frac{n+m}{2}\right)\Big/\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)$$

90th Percentiles ($F = .9$)

m \ n	1	2	3	4	5	6	8	12	24	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	59.44	60.70	62.00	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.37	9.41	9.45	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.25	5.22	5.18	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.95	3.90	3.83	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.34	3.27	3.19	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	2.98	2.90	2.82	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.75	2.67	2.58	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.59	2.50	2.40	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.47	2.38	2.28	2.16
10	3.28	2.92	2.73	2.61	2.52	2.46	2.38	2.28	2.18	2.06
12	3.18	2.81	2.61	2.48	2.39	2.33	2.24	2.15	2.04	1.90
15	3.07	2.70	2.49	2.36	2.27	2.21	2.12	2.02	1.90	1.76
20	2.97	2.59	2.38	2.25	2.16	2.09	2.00	1.89	1.77	1.61
25	2.92	2.53	2.32	2.18	2.09	2.02	1.93	1.82	1.69	1.52
30	2.88	2.49	2.28	2.14	2.05	1.98	1.88	1.77	1.64	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.83	1.71	1.57	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.77	1.66	1.51	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.72	1.60	1.45	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.67	1.55	1.38	1.00

(*continued on following page*)

Appendix B: Tables

Table B5. Variance Ratio Distribution (*continued*)

95th Percentiles ($F = .95$)

m \ n	1	2	3	4	5	6	8	12	24	∞
1	161	200	216	225	230	234	239	244	249	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3 12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

Table B5. Variance Ratio Distribution (*continued*)

99th Percentiles ($F = .99$)

m \ n	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5982	6106	6234	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.5	27.1	26.6	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	14.8	14.4	13.9	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.3	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4 73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2 12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Table B5. Variance Ratio Distribution (*continued*)

99.9th Percentiles ($F = .999$)

m \ n	1	2	3	4	5	6	8	12	24	∞
1*	405*	500*	540*	563*	576*	586*	598*	611*	623*	637*
2	998	999	999	999	999	999	999	999	999	999
3	167	149	141	137	135	133	131	128	126	124
4	74.1	61.3	56.2	53.4	51.7	50.5	49.0	47.4	45.8	44.1
5	47.2	37.1	33.2	31.1	29.8	28.8	27.6	26.4	25.1	23.8
6	35.5	27.0	23.7	21.9	20.8	20.0	19.0	18.0	16.9	15.8
7	29.2	21.7	18.8	17.2	16.2	15.5	14.6	13.7	12.7	11.7
8	25.4	18.5	15.8	14.4	13.5	12.9	12.0	11.2	10.3	9.34
9	22.9	16.4	13.9	12.6	11.7	11.1	10.4	9.57	8.72	7.81
10	21.0	14.9	12.6	11.3	10.5	9.92	9.20	8.45	7.64	6.76
12	18.6	13.0	10.8	9.63	8.89	8.38	7.71	7.00	6.25	5.42
15	16.6	11.3	9.34	8.25	7.57	7.09	6.47	5.81	5.10	4.31
20	14.8	9.95	8.10	7.10	6.46	6.02	5.44	4.82	4.15	3.38
25	13.9	9.22	7.45	6.49	5.88	5.46	4.91	4.31	3.66	2.89
30	13.3	8.77	7.05	6.12	5.53	5.12	4.58	4.00	3.36	2.59
40	12.6	8.25	6.60	5.70	5.13	4.73	4.21	3.64	3.01	2.23
60	12.0	7.76	6.17	5.31	4.76	4.37	3.87	3.31	2.69	1.90
120	11.4	7.32	5.79	4.95	4.42	4.04	3.55	3.02	2.40	1.54
∞	10.8	6.91	5.42	4.62	4.10	3.74	3.27	2.74	2.13	1.00

*For $m = 1$, the 99.9th percentiles are 100 times the tabulated values.

Source: R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Table V; published by Longman Group Ltd., London (previously published by Oliver and Boyd, Edinburgh); reprinted by permission of the authors and publishers.

Index