BAYESIAN INFERENCE IN STATISTICAL ANALYSIS

George E.P. Box
University of Wisconsin

George C. Tiao
University of Chicago

Wiley Classics Library Edition Published 1992

A Wiley-Interscience Publication
JOHN WILEY AND SONS, INC.
New York / Chichester / Brisbane / Toronto / Singapore

This page intentionally left blank

BAYESIAN INFERENCE
IN STATISTICAL ANALYSIS

This page intentionally left blank

BAYESIAN INFERENCE IN STATISTICAL ANALYSIS

George E.P. Box
University of Wisconsin

George C. Tiao
University of Chicago

Wiley Classics Library Edition Published 1992

A Wiley-Interscience Publication
JOHN WILEY AND SONS, INC.
New York / Chichester / Brisbane / Toronto / Singapore

Wiley Classics Library Edition Published 1992

Library of Congress Cataloging-in-Publication Data

Box, George E. P.
Bayesian inference in statistical analysis / George E.P. Box,
George C. Tiao.—Wiley classics library ed.
p. cm.
Originally published: Reading, Mass. : Addison-Wesley Pub. Co.,
c1973.
"A Wiley-Interscience publication."
Includes bibliographical references and indexes.
ISBN 0-471-57428-7
1. Mathematical statistics. I. Tiao, George C., 1933-
II. Title.
[QA276.B677 1992]
519.5'4—dc20 92-3745
CIP

10 9 8 7

To BARBARA, HELEN, and HARRY

This page intentionally left blank

PREFACE

The object of this book is to explore the use and relevance of Bayes' theorem to problems such as arise in scientific investigation in which inferences must be made concerning parameter values about which little is known *a priori*.

In Chapter 1 we discuss some important general aspects of the Bayesian approach, including: the role of Bayesian inference in scientific investigation, the choice of prior distributions (and, in particular, of noninformative prior distributions), the problem of nuisance parameters, and the role and relevance of sufficient statistics.

In Chapter 2, as a preliminary to what follows, a number of standard problems concerned with the comparison of location and scale parameters are discussed. Bayesian methods, for the most part well known, are derived there which closely parallel the inferential techniques of sampling theory associated with t-tests, F-tests, Bartlett's test, the analysis of variance, and with regression analysis. These techniques have long proved of value to the practicing statistician and it stands to the credit of sampling theory that it has produced them. It is also encouraging to know that parallel procedures may, with at least equal facility, be derived using Bayes' theorem. Now, practical employment of such techniques has uncovered further inferential problems, and attempts to solve these, using sampling theory, have had only partial success. One of the main *objectives of this book, pursued from Chapter 3 onwards, is to study some of these* problems from a Bayesian viewpoint. In this we have in mind that the value of Bayesian analysis may perhaps be judged by considering to what extent it supplies insight and sensible solutions for what are known to be awkward problems.

The following are examples of the further problems considered:

1. How can inferences be made in small samples about parameters for which no parsimonious set of sufficient statistics exists?
2. To what extent are inferences about means and variances sensitive to departures from assumptions such as error Normality, and how can such sensitivity be reduced?
3. How should inferences be made about variance components?
4. How and in what circumstances should mean squares be pooled in the analysis of variance?

5. How can information be pooled from several sources when its precision is not exactly known, but can be estimated, as, for example, in the "recovery of interblock information" in the analysis of incomplete block designs?
6. How should data be transformed to produce parsimonious parametrization of the model as well as to increase sensitivity of the analysis?

The main body of the text is an investigation of these and similar questions with appropriate analysis of the mathematical results illustrated with numerical examples. We believe that this (1) provides evidence of the value of the Bayesian approach, (2) offers useful methods for dealing with the important problems specifically considered and (3) equips the reader with techniques which he can apply in the solution of new problems.

There is a continuing commentary throughout concerning the relation of the Bayes results to corresponding sampling theory results. We make no apology for this arrangement. In any scientific discussion alternative views ought to be given proper consideration and appropriate comparisons made. Furthermore, many readers will already be familiar with sampling theory results and perhaps with the resulting problems which have motivated our study.

This book is principally a bringing together of research conducted over the years at Wisconsin and elsewhere in cooperation with other colleagues, in particular David Cox, Norman Draper, David Lund, Wai-Yuan Tan, and Arnold Zellner. A list of the consequent source references employed in each chapter is given at the end of this volume.

An elementary knowledge of probability theory and of standard sampling theory analysis is assumed, and from a mathematical viewpoint, a knowledge of calculus and of matrix algebra. The material forms the basis of a two-semester graduate course in Bayesian inference; we have successfully used earlier drafts for this purpose. Except for perhaps Chapters 8 and 9, much of the material can be taught in an advanced undergraduate course.

We are particularly indebted to Fred Mosteller and James Dickey, who patiently read our manuscript and made many valuable suggestions for its improvement, and to Mukhtar Ali, Irwin Guttman, Bob Miller, and Steve Stigler for helpful comments. We also wish to record our thanks to Biyi Afonja, Yu-Chi Chang, William Cleveland, Larry Haugh, Hiro Kanemasu, David Pack, and John MacGregor for help in checking the final manuscript, to Mary Esser for her patience and care in typing it, and to Greta Ljung and Johannes Ledolter for careful proofreading.

The work has involved a great deal of research which has been supported by the Air Force Office of Scientific Research under Grants AF-AFOSR-1158-66, AF-49(638) 1608 and AF-AFOSR 69-1803, the Office of Naval Research under Contract ONR-N-00014-67-A-0128-0017, the Army Office of Ordnance Research under Contract DA-ARO-D-31-124-G917, the National Science Foundation under Grant GS-2602, and the British Science Research Council.

The manuscript was begun while the authors were visitors at the Graduate School of Business, Harvard University, and we gratefully acknowledge support from the Ford Foundation while we were at that institution. We must also express our gratitude for the hospitality extended to us by the University of Essex in England when the book was nearing completion.

We are grateful to Professor E. S. Pearson and the Biometrika Trustees, to the editors of *Journal of the American Statistical Association* and *Journal of the Royal Statistical Society Series B*, and to our coauthors David Cox, Norman Draper, David Lund, Wai-Yuan Tan, and Arnold Zellner for permission to reprint condensed and adapted forms of various tables and figures from articles listed in the principal source references and general references. We are also grateful to Professor O. L. Davies and to G. Wilkinson of the Imperial Chemical Industries, Ltd., for permission to reproduce adapted forms of Tables 4.2 and 6.3 in *Statistical Methods in Research and Production*, 3rd edition revised, edited by O. L. Davies.

We acknowledge especial indebtedness for support throughout the whole project by the Wisconsin Alumni Research Foundation, and particularly for their making available through the University Research Committee the resources of the Wisconsin Computer Center.

Madison, Wisconsin
August 1972

G.E.P.B.
G.C.T.

This page intentionally left blank

CONTENTS

Chapter 3 Bayesian Assessment of Assumptions

1. Effect of Non-Normality on Inferences about a Population Mean with Generalizations

Chapter 4 Bayesian Assessment of Assumptions

2. Comparison of Variances

Chapter 5 Random Effect Models

Chapter 10 **Transformation of Data**

CHAPTER 1

NATURE OF BAYESIAN INFERENCE

1.1 INTRODUCTION AND SUMMARY

Opinion as to the value of Bayes' theorem as a basis for statistical inference has swung between acceptance and rejection since its publication in 1763. During periods when it was thought that alternative arguments supplied a satisfactory foundation for statistical inference Bayesian results were viewed, sometimes condescendingly, as an interesting but mistaken attempt to solve an important problem. When subsequently it was found that initially unsuspected difficulties accompanied the alternatives, interest was rekindled. Bayes' mode of reasoning, finally buried on so many occasions, has recently risen again with astonishing vigor.

In addition to the present growing awareness of possible deficiencies in the alternatives, three further factors account for the revival. First, the work of a number of authors, notably Fisher, Jeffreys, Barna¨d, Ramsey, De Finetti, Savage, Lindley, Anscombe and Stein, has, although not always directed to that end, helped to clarify and overcome some of the philosophical and practical difficulties.

Second, while other inferential theories had yielded nice solutions in cases where rather special assumptions such as Normality and independence of errors could be made, in other cases, and particularly where no sufficient statistics existed, the solutions were often unsatisfactory and messy. Although it is true that these special assumptions covered a number of situations of scientific interest, it would be idle to pretend that the set of statistical problems whose solution has been or will be needed by the scientific investigator coincides with the set of problems thus amenable to convenient treatment. Data gathering is frequently expensive compared with data analysis. It is sensible then that hard-won data be inspected from many different viewpoints. In the selection of viewpoints, Bayesian methods allow greater emphasis to be given to scientific interest and less to mathematical convenience.

Third, the nice solutions based on the rather special assumptions have been popular for another reason—they were easy to compute. This consideration has much less force now that the desk calculator is no longer the most powerful instrument for executing statistical analysis. Suppose, using a desk calculator, it takes five hours to perform a data analysis appropriate to the assumption that errors are Normal and independent, then the five hundred hours it might take to explore less restrictive assumptions could be prohibitive. By contrast, the use of an electronic computer can so reduce the time base that, with general programs available, the wider analysis can be almost as immediate and economic as the more restricted one.

Scientific investigation uses statistical methods in an iteration in which controlled data gathering and data analysis alternate. Data analysis is a subiteration in which inference from a tentatively entertained model alternates with criticism of the conditional inference by inspection of residuals and other means. Statistical inference is thus only one of the responsibilities of the statistician. It is however an important one. Bayesian inference alone seems to offer the possibility of sufficient flexibility to allow reaction to scientific complexity free from impediment from purely technical limitation.

A prior distribution, which is supposed to represent what is known about unknown parameters before the data is available, plays an important role in Bayesian analysis. Such a distribution can be used to represent prior knowledge or relative ignorance. In problems of scientific inference we would usually, were it possible, like the data "to speak for themselves." Consequently, it is usually appropriate to conduct the analysis as if a state of relative ignorance existed *a priori*. In this book, therefore, extensive use is made of "noninformative" prior distributions and very little of informative priors. The aim is to obtain an inference which would be appropriate for an unprejudiced observer. The understandable uneasiness felt by some statisticians about the use of prior distributions is often associated with the fear that the prior may dominate and distort "what the data are trying to say." We hope to show by the examples in this book that, by careful choice of model structure and appropriate noninformative priors, Bayesian analysis can produce the reverse of what is feared. It can permit the data to comment on dubious aspects of a model in a manner not otherwise possible.

The usefulness of a theory is customarily assessed by tentatively adopting it, and then considering whether its consequences agree with common sense, and whether they provide insight where common sense fails. It was in this spirit that some years ago the authors with others began research in applications of the Bayesian theory of inference. A series of problems were selected in the solution of which difficulties or inconsistencies had been encountered with other approaches. Because Bayesian analysis of these problems has seemed consistently helpful and interesting, we believe it is now appropriate to bring this and other related work together, and to consider its wider aspects.

The objective of this book, therefore, is to explore Bayesian inference in statistical analysis. The book consists of ten chapters. Chapter 1 discusses the role of statistical inference in scientific investigation. In the light of that discussion the nature of Bayesian inference, including the choice of noninformative prior distributions, is considered. The chapter ends with an account of the role and relevance of sufficient statistics, and discusses the problem of nuisance parameters.

In Chapter 2 a number of standard Normal theory inference problems concerning location and scale parameters are considered. Bayes' solutions are given which closely parallel sampling theory techniques† associated with t-tests, F-tests, the analysis of variance and regression analysis. While these procedures have long proved valuable to practising statisticians, efforts to extend them in important directions using non-Bayesian theories have met serious difficulties. An advantage of the Bayes approach is that it can be used to explore the consequences of any type of probability model, without restriction to those having special mathematical forms. Thus, in Chapter 3 the problem of making inferences about location parameters is considered for a wider class of parent probability models of which the Normal distribution is a member. In this framework, we show how it is possible to assess to what extent inferences about location parameters are sensitive to departures from Normality. Further, it is shown how we can use the evidence from the data to make inferences about the form of the parent distributions of the observations. The analysis is extended in Chapter 4 to the problem of comparing variances.

Chapters 5 and 6 discuss various random effect and mixed models associated with hierarchical and cross classification designs. With sampling theory, one experiences a number of difficulties in estimating means and variance components in these models. Notably one encounters problems of negative variance estimates, of eliminating nuisance parameters, of constructing confidence intervals, and of pooling variance estimates. Analysis, from a Bayesian standpoint, is much more tractable, and in particular provides an interesting and sensible solution to the pooling dilemma.

Chapter 7 deals with two further important problems in the analysis of variance. The first concerns the estimation of means in the one-way classification. When it is sensible to regard such means as themselves a sample from a population, the appropriate Bayesian analysis shows that there are then *two* sources of information about the means and appropriately combines them. The chapter ends with a discussion of the recovery of interblock information in the balanced incomplete block design model. This is again a problem in which two sources of information need to be appropriately combined and for which the sampling theory solution is unsatisfactory.

In Chapters 8 and 9 a general treatment of linear and nonlinear Normal multivariate models is given. While Bayesian results associated with standard linear models are discussed, particular attention is given to the problem of estimating common location parameters from several equations. The latter problem is of considerable practical importance, but is difficult to tackle bysampling theory methods, and has not previously received much attention.

† We shall assume in this book that the reader has some familiarity with standard ideas of the sampling theory approach explained for example in Mood and Graybill (1963) and Hogg and Craig (1970).

Finally, in Chapter 10, we consider the important problem of data transformation from a Bayesian viewpoint. The problem is to select a transformation which, so far as possible, achieves Normality, homogeniety of variance, and simplicity of the expectation function in the transformed variate.

A bald statement of a mathematical expression, however correct, frequently fails to produce understanding. Many Bayesian results are of particular interest because they seem to provide a kind of higher intuition. Mathematical results which at first seemed puzzling have later been seen to provide a maturer kind of common sense. For this reason, throughout this book, individual mathematical formulae are carefully analyzed and illustrated with examples and diagrams. Also, appropriate approximations are developed when they provide deeper understanding of a situation, or where they simplify calculation. For the convenience of the reader a number of short summaries of formulas and calculations are given in appropriate places.

1.1.1 The Role of Statistical Methods in Scientific Investigation

Statistical methods are tools of scientific investigation. Scientific investigation is a controlled learning process in which various aspects of a problem are illuminated as the study proceeds. It can be thought of as a major iteration within which secondary iterations occur. The major iteration is that in which a tentative conjecture suggests an experiment, appropriate analysis of the data so generated leads to a modified conjecture, and this in turn leads to a new experiment, and so on. An idealization of this process is seen in Fig. 1.1.1, involving an alternation between *conjecture* and *experiment* carried out via experimental *design* and data *analysis*.† As indicated by the zig-zag line at the bottom of the figure, most investigations involve not one but a number of alternations of this kind.

An efficient investigation is one where convergence to the objective occurs as quickly and unambiguously as possible. A basic determinant of efficiency, which we must suppose is outside the control of the statistician, is the originality, imagination, and subject matter knowledge of the investigator. Apart from this vital determining factor, however, efficiency is decided by the appropriateness and force of the methods of design and analysis employed. In moving from conjecture to experimental data, (D), experiments must be designed which make best use of the experimenter's current state of knowledge and which best illuminate his conjecture. In moving from data to modified conjecture, (A), data must be analyzed so as to accurately present information in a manner which is readily understood by the experimenter.

† The words design and experiment are broadly interpreted here to refer to any data gathering process. In an economic study, a conjecture might lead the investigator to study the functional relationship between money supply and interest rate. The difficult decision as to what types of money supply and interest rate data to use, here constitutes the design. In social studies a particular sample survey might be the experiment.

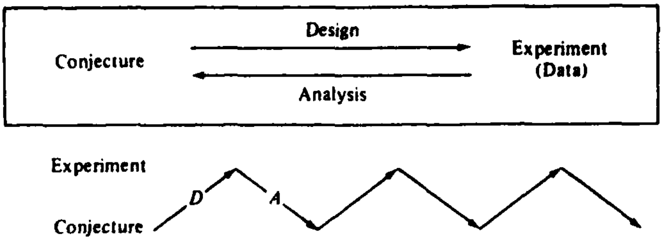


Fig. 1.1.1 Iterative process of scientific investigation (the alternation between conjecture and experiment).

A full treatise on the use of statistical methods in scientific investigation therefore would necessarily include consideration of statistical design as well as statistical analysis. The aims of this book are, however, much more limited. We shall not discuss experimental design, and will be concerned only with one aspect of statistical analysis, namely, *statistical inference.*

1.1.2 Statistical Inference as one Part of Statistical Analysis

For illustration, suppose we were studying the useful life of batteries produced by a particular machine. It might be appropriate to assume tentatively that the observed lives of batteries coming from the machine were distributed independently and Normally about some mean θ with variance σ^2. The probability distribution of a projected sample of n observations $\mathbf{y}' = (y_1, \ldots, y_n)$ would then be

$$p(\mathbf{y} \mid \theta, \sigma^2) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \theta)^2\right], \qquad -\infty < y_t < \infty. \quad (1.1.1)$$

Given the value of the parameters θ and σ^2, this expression permits the calculation of the probability density $p(\mathbf{y} \mid \theta, \sigma^2)$ associated with any *hypothetical* data set $\mathbf{y}$ *before* any data is taken. For statistical analysis this is, in most cases, the converse of what is needed. The analyst already has the data but he does not know θ and σ^2. He can, however, use $p(\mathbf{y} \mid \theta, \sigma^2)$ indirectly to make *inferences* about the values of θ and σ^2, given the n data values.

Two of the methods by which this may be attempted employ

a. Sampling Theory,

b. Bayes' Theorem.

We now give a brief description of each of these approaches using the Normal probability model (1.1.1) for illustration.

Sampling Theory Approach

In this approach inferences are made by directing attention to a reference set of hypothetical data vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_j, \ldots$ which could have been generated by the probability model $p(\mathbf{y} \mid \theta_0, \sigma_0^2)$ of (1.1.1), where θ_0 and σ_0^2 are the hypothetical true values of θ and σ^2. Estimators $\hat{\theta}(\mathrm{y})$ and $\hat{\sigma}^2(\mathrm{y})$, which are functions of the data vector y, are selected. By imagining values $\hat{\theta}(\mathrm{y}_j)$ and $\hat{\sigma}^2(\mathrm{y}_j)$ to be calculated for each hypothetical data vector y_j, reference sets are generated for $\hat{\theta}(\mathrm{y})$ and $\hat{\sigma}^2(\mathrm{y})$. Inferences are then made by comparing the values of $\hat{\theta}(\mathrm{y})$ and $\hat{\sigma}_2(\mathrm{y})$ actually observed with their "sampling distributions" generated by the reference sets.

The functions $\hat{\theta}(\mathrm{y})$ and $\hat{\sigma}^2(\mathrm{y})$ are usually chosen so that the sampling distributions of the extimators $\hat{\theta}(\mathrm{y}_j)$ and $\hat{\sigma}^2(\mathrm{y}_j)$ are, in some sense, concentrated as closely as possible about the true values θ_0 and σ_0. To provide some idea of how far away from the true values the calculated quantities $\hat{\theta}(\mathrm{y})$ and $\hat{\sigma}^2(\mathrm{y})$ might be, *confidence intervals* are calculated. For example, the $1-\alpha$ confidence interval for θ would be of the form

$$\bar{\theta}_1(\mathrm{y}) < \theta < \bar{\theta}_2(\mathrm{y}),$$

where $\bar{\theta}_1(\mathrm{y})$ and $\bar{\theta}_2(\mathrm{y})$ would be functions of y, chosen so that in repeated sampling the computed confidence intervals included the value θ_0, a proportion $1-\alpha$ of the time.

Bayesian Approach

In a Bayesian approach, a different line is taken. As part of the model a *prior* distribution $p(\theta, \sigma^2)$ is introduced. This is supposed to express a state of knowledge or ignorance about θ and σ^2 before the data are obtained. Given the prior distribution, the probability model $p(\mathrm{y} \mid \theta, \sigma^2)$ and the data y, it is now possible to calculate the probability distribution $p(\theta, \sigma^2 \mid \mathrm{y})$ of θ and σ^2, given the data y. This is called the *posterior* distribution of θ and σ^2. From this distribution inferences about the parameters are made.

1.1.3 The Question of Adequacy of Assumptions

Consider the battery-life example and suppose $n = 20$ observations are available. Then, whichever method of inference is used, *conditional on the assumptions* we can summarize all the information in the 20 data values in terms of inferences about just *two* parameters, θ and σ^2.

The inferences are, in particular, conditional on the adequacy of the probability model in (1.1.1). It is not difficult, however, to imagine situations in which this model, and therefore the associated inferences, could be inadequate. It might happen, for example, that during the period of observation, a quality characteristic x of a chemical additive, used in making the batteries, could vary and could cause, via an approximate linear relationship, a corresponding change in the mean life time of the batteries. In this case, a more appropriate model might be

$$p(\mathrm{y} \mid \mathrm{x}, \sigma^2, \theta_1, \theta_2) \propto \sigma^{-20} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{20}(y_t - \theta_1 - \theta_2 x_t)^2\right],$$

$$-\infty < y_t < \infty. \qquad (1.1.2)$$

Alternatively, it might be suspected that the first battery of a production run was always faulty, in which case a more adequate model could be

$$p(\mathbf{y} \mid \sigma_1^2, \sigma^2, \theta_1, \theta) \propto \sigma_1^{-1}\sigma^{-19} \exp\left[-\frac{1}{2\sigma_1^2}(y_1 - \theta_1)^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{20}(y_t - \theta)^2\right],$$
$$-\infty < y_t < \infty. \qquad (1.1.3)$$

Again it could happen that successive observations were not distributed independently but followed some time series, or it might be that their distribution was highly non-Normal. The reader will have no difficulty in inventing many other situations that might arise and the probability models that could describe them.

Clearly the inferences which can be made will depend upon which model is selected. Whence it is seen that a basic dilemma exists in all statistical analysis. Such analysis implies the summarizing of information contained in a body of data via a probability model containing a minimum of parameters. We need such a summary to see clearly, and so to make progress, but if the model were inappropriate the summary could distort and exclude relevant information.

1.1.4 An Iterative Process of Model Building in Statistical Analysis

Because we can *never* be sure that a postulated model is entirely appropriate, we must proceed in such a manner that inadequacies can be taken account of and their implications considered as we go along. To do this we must regard statistical analysis, which is a step in the major iteration of Fig. 1.1.1, as itself an iteration. To be on firm ground we must do more than merely postulate a model; we must build and test a tentative model at each stage of the investigation.

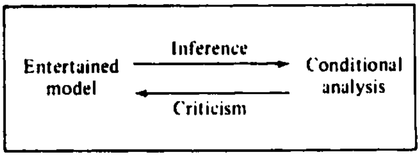


Fig. 1.1.2 Statistical analysis of data as an iterative process of model building.

Only when the analyst and the investigator are satisfied that no important fact has been overlooked and that the model is adequate to the purpose, should it be used to further the major iteration. The iterative model building process† taking place *within a statistical analysis* is depicted in Fig. 1.1.2.

The process usually begins by the postulating of a model worthy to be tentatively entertained. The data analyst will have arrived at this tentative model

† A fuller discussion is found, for example, in Box and Jenkins (1970), where in the context of time series analysis the steps in this iteration are discussed in terms of model identification, model fitting, and model diagnostic checking.

in cooperation with the scientific investigator. They will choose it so that, in the light of the then available knowledge, it best takes account of relevant phenomena in the simplest way possible. It will usually contain unknown parameters. Given the data the analyst can now make statistical inferences about the parameters conditional on the correctness of this first tentative model. These inferences form part of the conditional analysis. *If the model is correct*, they provide all there is to know about the problem under study, given the data.

Up to now the analyst has proceeded as if he believed the model absolutely. He now changes his role from tentative sponsor to tentative critic and broadens his analysis with computations throwing light on the question: "Is the model adequate?" Residual quantities are calculated which, while they would contain no information if the tentative model were true, could suggest appropriate modifications if it were false. Resulting speculation as to the appropriateness of the initially postulated model and possible need for modification, again conducted in cooperation with the investigator, may be called model *criticism*.†

For example, suppose the Normal probability model of (1.1.1) was initially postulated, and a sample of 20 successive observations were taken from a production run. These would provide the data from which, conditional on the model, inferences could be made about θ and σ^2.

An effective way of criticizing the adequacy of the assumed model (1.1.1) employs what is called "an analysis of residuals." Suppose for the moment that θ and σ^2 were known; then if the model (1.1.1) were adequate, the quantities $u_1 = (y_1 - \theta)/\sigma, \ldots, u_t = (y_t - \theta)/\sigma, \ldots$ would be a random sample from a Normal distribution with zero mean and unit variance. Such a sequence would by itself be informationless, and is sometimes referred to as white noise. Thus, a check on model adequacy would be provided by inspection of the quantities $y_t - \theta = u_t\sigma$, $t = 1, 2, \ldots$. Any suggestion that these quantities were nonrandom, or that they were related to some other known variable, could provide a hint that the entertained model (1.1.1) should be modified.

In practice, θ would be unknown but we could proceed by substituting the sample mean $\bar{y}$. The resulting quantities $r_t = y_t - \bar{y}$, $t = 1, 2, \ldots$ would for this example be the residuals. If, for example, they seemed to be correlated with the amount of additive x_t, this would suggest that a model like (1.1.2) might be more appropriate. This new model might then be entertained, and the iterative process of Fig. 1.1.2 repeated.

Useful devices for model criticism have been proposed, in particular by Anscombe (1961), Anscombe and Tukey (1963), and Daniel (1959). Many of these involve plotting residuals in various ways. However, these techniques are not part of statistical inference as we choose to consider it, but of model criticism which is an essential adjunct to inference in the adaptive process of data analysis depicted in Fig. 1.1.2.

† This apt term is due to Cuthbert Daniel.

1.1.5 The Role of Bayesian Analysis

The applications of Bayes' theorem which we discuss, therefore, are examples of statistical inference. While inference is only a part of statistical analysis, which is in turn only a part of design and analysis, used in the investigatory iteration, nevertheless it is an important part.

Among different systems of statistical inference, that derived from Bayes' theorem will, we believe, be seen to have properties which make it particularly appropriate to its role in scientific investigation. In particular:

1. Precise assumption introduced on the left in Fig. 1.1.2 leads, via a *leak proof* route, to consequent inference on the right.
2. It follows that, given the model, Bayesian analysis automatically makes use of all the information from the data.
3. It further follows that inferences that are unacceptable *must* come from inappropriate assumption and not from inadequacies of the inferential system. Thus all parts of the model, including the prior distribution, are exposed to appropriate criticism.
4. Because this system of inference may be readily applied to any probability model, much less attention need be given to the mathmetical convenience of the models considered and more to their scientific merit.
5. Awkward problems encountered in sampling theory, concerning choice of estimators and of confidence intervals, do not arise.
6. Bayesian inference provides a satisfactory way of explicitly introducing and keeping track of assumptions about prior knowledge or ignorance. It should be recognized that some prior knowledge is employed in all inferential systems. For example, a sampling theory analysis using (1.1.1) is made, as is a Bayesian analysis, as if it were believed *a priori* that the probability distribution of the data was *exactly* Normal, and that each observation had exactly the *same* variance, and was distributed *exactly* independently of every other observation. But after a study of residuals had suggested model inadequacy, it might be desirable to reanalyse the data in relation to a less restrictive model into which the initial model was embeded. If non-Normality was suspected, for example, it might be sensible to postulate that the sample came from a wider class of parent distributions of which the Normal was a member. The consequential analysis could be difficult via sampling theory but is readily accomplished in a Bayesian framework (see Chapters 3 and 4). Such an analysis allows evidence *from the data* to be taken into account about the form of the parent distribution besides making it possible to assess to what extent the prior assumption of exact Normality is justified.

The above introductory survey suggests that Bayes' theorem provides a system of statistical inference suited to iterative model building, which is in turn an essential part of scientific investigation. On the other hand, we have pointed out that statistical inference (Bayesian or otherwise) is only a part of statistical method. It is, we believe, equally unhelpful for enthusiasts to seem to claim that Bayesian analysis can do everything, as it is for its detractors to seem to assert that it can do nothing.

1.2 NATURE OF BAYESIAN INFERENCE

1.2.1 Bayes' Theorem

Suppose that $\mathbf{y}' = (y_1, \ldots, y_n)$ is a vector of n observations whose probability distribution $p(\mathbf{y} \mid \boldsymbol{\theta})$ depends on the values of k parameters $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_k)$. Suppose also that $\boldsymbol{\theta}$ itself has a probility distribution $p(\boldsymbol{\theta})$. Then,

$$p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{y}, \boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y})p(\mathbf{y}). \qquad (1.2.1)$$

Given the observed data $\mathbf{y}$, the conditional distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \qquad (1.2.2)$$

Also, we can write

$$p(\mathbf{y}) = E\,p(\mathbf{y} \mid \boldsymbol{\theta}) = c^{-1} = \begin{cases} \int p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta} & \boldsymbol{\theta} \text{ continuous} \\ \sum p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) & \boldsymbol{\theta} \text{ discrete} \end{cases} \qquad (1.2.3)$$

where the sum or the integral is taken over the admissible range of $\boldsymbol{\theta}$, and where $E[f(\boldsymbol{\theta})]$ is the mathematical expectation of $f(\boldsymbol{\theta})$ with respect to the distribution $p(\boldsymbol{\theta})$. Thus we may write (1.2.2) alternatively as

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = c\,p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}). \qquad (1.2.4)$$

The statement of (1.2.2), or its equivalent (1.2.4), is usually referred to as *Bayes' theorem.* In this expression, $p(\boldsymbol{\theta})$, which tells us what is known about $\boldsymbol{\theta}$ without knowledge of the data, is called the *prior* distribution of $\boldsymbol{\theta}$, or the distribution of $\boldsymbol{\theta}$ *a priori*. Correspondingly, $p(\boldsymbol{\theta} \mid \mathbf{y})$, which tells us what is known about $\boldsymbol{\theta}$ given knowledge of the data, is called the *posterior* distribution of $\boldsymbol{\theta}$ given $\mathbf{y}$, or the distribution of $\boldsymbol{\theta}$ *a posteriori*. The quantity c is merely a "normalizing" constant necessary to ensure that the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$ integrates or sums to one.

In what follows we sometimes refer to the prior distribution and the posterior distribution simply as the "prior" and the "posterior", respectively.

Bayes' Theorem and the Likelihood Function

Now given the data $\mathbf{y}$, $p(\mathbf{y} \mid \boldsymbol{\theta})$ in (1.2.4) may be regarded as a function not of $\mathbf{y}$ but of $\boldsymbol{\theta}$. When so regarded, following Fisher (1922), it is called the *likelihood function* of $\boldsymbol{\theta}$ for given $\mathbf{y}$ and can be written $l(\boldsymbol{\theta} \mid \mathbf{y})$. We can thus write Bayes' formula as

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = l(\boldsymbol{\theta} \mid \mathbf{y})p(\boldsymbol{\theta}). \qquad (1.2.5)$$

In other words, then, Bayes' theorem tells us that the probability distribution for $\boldsymbol{\theta}$ posterior to the data $\mathbf{y}$ is proportional to the product of the distribution for $\boldsymbol{\theta}$ prior to the data and the likelihood for $\boldsymbol{\theta}$ given $\mathbf{y}$. That is,

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

The likelihood function $l(\boldsymbol{\theta} \mid \mathbf{y})$ plays a very important role in Bayes' formula. It is *the* function through which the data $\mathbf{y}$ modifies prior knowledge of $\boldsymbol{\theta}$; it can therefore be regarded as representing the information about $\boldsymbol{\theta}$ coming from the data.

The likelihood function is defined up to a multiplicative constant, that is, multiplication by a constant leaves the likelihood unchanged. This is in accord with the role it plays in Bayes' formula, since multiplying the likelihood function by an arbitrary constant will have no effect on the posterior distribution of $\boldsymbol{\theta}$. The constant will cancel upon normalizing the product on the right hand side of (1.2.5). It is only the relative value of the likelihood which is of importance.

The Standardized Likelihood

When the integral $\int l(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta}$, taken over the admissible range of $\boldsymbol{\theta}$, is finite, then occasionally it will be convenient to refer to the quantity

$$\frac{l(\boldsymbol{\theta} \mid \mathbf{y})}{\int l(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta}}. \tag{1.2.6}$$

We shall call this the *standardized likelihood*, that is, the likelihood scaled so that the area, volume, or hypervolume under the curve, surface, or hypersurface, is one.

Sequential Nature of Bayes' Theorem

The theorem in (1.2.5) is appealing because it provides a mathematical formulation of how previous knowledge may be combined with new knowledge. Indeed, the theorem allows us to continually update information about a set of parameters $\boldsymbol{\theta}$ as more observations are taken.

Thus, suppose we have an initial sample of observations $\mathbf{y}_1$, then Bayes' formula gives

$$p(\boldsymbol{\theta} \mid \mathbf{y}_1) \propto p(\boldsymbol{\theta})\, l(\boldsymbol{\theta} \mid \mathbf{y}_1). \tag{1.2.7}$$

Now, suppose we have a second sample of observations $\mathbf{y}_2$ distributed independently of the first sample, then

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{y}_2, \mathbf{y}_1) &\propto p(\boldsymbol{\theta})\, l(\boldsymbol{\theta} \mid \mathbf{y}_1)\, l(\boldsymbol{\theta} \mid \mathbf{y}_2) \\ &\propto p(\boldsymbol{\theta} \mid \mathbf{y}_1)\, l(\boldsymbol{\theta} \mid \mathbf{y}_2). \end{aligned} \tag{1.2.8}$$

The expression (1.2.8) is precisely of the same form as (1.2.7) except that $p(\boldsymbol{\theta} \mid \mathbf{y}_1)$, the posterior distribution for $\boldsymbol{\theta}$ given $\mathbf{y}_1$, plays the role of the prior distribution for the second sample. Obviously this process can be repeated any number of times. In particular, if we have n independent observations, the posterior distribution can, if desired, be recalculated after each new observation, so that at the mth stage the likelihood associated with the mth observation is combined with the posterior distribution of θ after $m - 1$ observations to give the new posterior distribution

$$p(\theta \mid y_1, \ldots, y_m) \propto p(\theta \mid y_1, \ldots, y_{m-1}) l(\theta \mid y_m), \qquad m = 2, \ldots, n \qquad (1.2.9)$$

where

$$p(\theta \mid y_1) \propto p(\theta) l(\theta \mid y_1).$$

Thus, Bayes' theorem describes, in a fundamental way, the process of learning from experience, and shows how knowledge about the state of nature represented by θ is continually modified as new data becomes available.

1.2.2 Application of Bayes' Theorem with Probability Interpreted as Frequencies

Mathematically, Bayes' formula is merely a statement of conditional probability, and as such its validity is not in question. What *has* been questioned is its applicability to general problems of scientific inference. The difficulties concern

a. the meaning of probability, and

b. the choice of, and necessity for, the prior distribution.

Specific examples can be found of applications of Bayes' theorem where the probabilities involved may be directly interpreted in terms of frequencies and may therefore be said to be objective, and where the prior probabilities can be supposed exactly known. The validity of applications of this sort has not been in serious dispute. An example of this situation is described by Fisher (1959, p.19). In this example, there are mice of two colors, black and brown. The black mice are of two genetic kinds, homozygotes (*BB*) and heterozygotes (*Bb*), and the brown mice are of one kind (*bb*). It is known from established genetic theory that the probabilities associated with offspring from various matings are as follows:

Table 1.2.1

Probabilities for genetic character of mice offspring

Mice	*BB* (black)	*Bb* (black)	*bb* (brown)
BB mated with *bb*	0	1	0
Bb mated with *bb*	0	$\frac{1}{2}$	$\frac{1}{2}$
Bb mated with *Bb*	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Suppose we have a "test" mouse which is black and has been produced by a mating between two (*Bb*) mice. Using the information in the last line of the table, it is seen that, in this case, the prior probabilities of the test mouse being homozygous (BB) and heterozygous (Bb) are precisely known, and are $\frac{1}{3}$ and $\frac{2}{3}$ respectively. Given this prior information, Fisher supposed that the test mouse was now mated with a brown mouse and produced (by way of data) seven black offspring. One can then calculate, as Fishei did, the probabilities, posterior to the data, of the test mouse being homozygous (BB) and heterozygous (Bb) using Bayes' theorem.

Specifically, if we use θ to denote the test mouse being (BB) or (Bb),

$$\theta = \begin{cases} 0 & (BB) \\ 1 & (Bb) \end{cases}$$

then the prior knowledge is represented by the distribution

$$p(\theta = 0) = \Pr(BB) = \tfrac{1}{3}, \qquad p(\theta = 1) = \Pr(Bb) = \tfrac{2}{3}.$$

Further, letting y denote the offspring, we have the likelihood

$$l(\theta = 0 \mid \mathrm{y} = 7 \text{ black}) \propto \Pr(7 \text{ black} \mid BB) = 1,$$

$$l(\theta = 1 \mid \mathrm{y} = 7 \text{ black}) \propto \Pr(7 \text{ black} \mid Bb) = (\tfrac{1}{2})^7.$$

It follows from (1.2.5) that

$$p(\theta = 0 \mid \mathrm{y} = 7 \text{ black}) \propto \tfrac{1}{3}, \qquad p(\theta = 1 \mid y = 7 \text{ black}) \propto (\tfrac{2}{3})(\tfrac{1}{2})^7.$$

Upon normalizing the posterior probabilities are then

$$p(\theta = 0 \mid \mathrm{y} = 7 \text{ black}) = \Pr(BB \mid 7 \text{ black}) = \tfrac{64}{65},$$

$$p(\theta = 1 \mid \mathrm{y} = 7 \text{ black}) = 1 - \Pr(BB \mid 7 \text{ black}) = \tfrac{1}{64}.$$

which represent the posterior knowledge of the test mouse being (BB) or (Bb). We see that, given the genetic characteristics of the offspring, the mating results of 7 black offspring changes our knowledge considerably about the test mouse being (BB) or (Bb), from a prior probability ratio of 2:1 in favor of (Bb) to a posterior ratio of 64:1 against it.

As an illustration of the sequential nature of Bayes' theorem, suppose the 7 black offspring are viewed as a sequence of seven independent observations; then, if we let $\mathrm{y}' = (y_1, \ldots, y_7)$, the likelihood can be written

$$l(\theta \mid \mathrm{y} = 7 \text{ black}) = l(\theta \mid y_1 = \text{black}) \cdots l(\theta \mid y_7 = \text{black})$$

where

$$l(\theta \mid y_m = \text{black}) \propto \begin{cases} 1 & \theta = 0 \\ \frac{1}{2} & \theta = 1 \end{cases}, \qquad m = 1, \ldots, 7.$$

Applying (1.2.9), the changes in the probabilities of the test mouse being (BB) or (Bb) after the mth observation, $m = 1, \ldots, 7$, are given in Table 1.2.2.

Table 1.2.2

Probabilities for the test mouse being homozygous and heterozygous

Mice	Probabilities	
	$\theta = 0\,(BB)$	$\theta = 1\,(Bb)$
Initial	$\frac{1}{3}$	$\frac{2}{3}$
1st black	$\frac{1}{2}$	$\frac{1}{2}$
2nd black	$\frac{2}{3}$	$\frac{1}{3}$
3rd black	$\frac{4}{5}$	$\frac{1}{5}$
4th black	$\frac{8}{9}$	$\frac{1}{9}$
5th black	$\frac{16}{17}$	$\frac{1}{17}$
6th black	$\frac{32}{33}$	$\frac{1}{33}$
7th black	$\frac{64}{65}$	$\frac{1}{65}$

This shows the increasing certainty of the test mouse being (BB) as more and more black offspring are observed.

Other applications of this sort are to be found in the theory of design of sampling inspection schemes. See, for example, Barnard (1954). In these examples, all the probabilities, both prior and posterior, are *objective* in the sense that they may be given a direct limiting frequency interpretation and are, in principle, subject to experimental confirmation.

In most scientific applications, however, exactly known objective prior distributions are rarely available.

1.2.3 Application of Bayes' Theorem with Subjective Probabilities

Following Ramsay (1931), De Finetti (1937), and Savage (1954, 1961a, b, 1962), we shall in this book regard probability as a mathematical expression of our degree of belief with respect to a certain proposition. In this context the concept of verification of probabilities by repeated experimental trials is regarded merely as a means of calibrating a subjective attitude. Thus, to say that one feels the probability is one half that Miss A and Mr. B will get married means that we have the same belief in the proposition "Mr. B will marry Miss A" as we would in the proposition "a toss of a fair coin will produce a head." We do not need to imagine an infinite series of situations in half of which A and B are wedded, and in half of which they are not.

The actual elucidation of what is believed by a particular person can be attempted in terms of betting odds. If, for example, the value of a continuous parameter θ is in question, we may, in suitable circumstances, infer an experimenter's prior distribution by asking at what value θ_0 he would be prepared to bet at particular odds that $\theta > \theta_0$. Given that a subjective probability distribution of this kind represents *a priori* what a person believes, then the posterior distribution obtained by combining this prior with the likelihood function shows how the prior beliefs are modified by information coming from the data.

Estimation of a Physical Constant

To consolidate ideas, we consider the example illustrated in Fig. 1.2.1. Suppose two physicists, A and B, are concerned with obtaining more accurate estimates of some physical constant θ, previously known only approximately. Suppose physicist A, being very familiar with this area of study, can make a moderately good guess of what the answer will be, and that his prior opinion about θ can be approximately represented by a Normal distribution centered at 900, with a standard deviation of 20. Thus

$$p_A(\theta) = \frac{1}{\sqrt{2\pi}\,20} \exp\left[-\frac{1}{2}\left(\frac{\theta - 900}{20}\right)^2\right]. \tag{1.2.10a}$$

According to A, *a priori* $\theta \sim N(900, 20^2)$ where the notation means that θ is distributed Normally with "mean 900 and variance 20^2." This would imply, in particular, that to A the chance that the value of θ could differ from 900 by more than 40 was only about one in twenty. By contrast, we suppose that B has had little previous experience in this area ,and his rather vague prior beliefs are represented by the Normal distribution

$$p_B(\theta) = \frac{1}{\sqrt{2\pi}\,80} \exp\left[-\frac{1}{2}\left(\frac{\theta - 800}{80}\right)^2\right]. \tag{1.2.10b}$$

Thus, according to B, $\theta \sim N(800, 80^2)$. He centers his prior at 800 and is considerably less certain about θ than A is. To B, a value anywhere between 700 and 900 would certainly be plausible. The curves in Fig. 1.2.1(a) labelled $p_A(\theta)$ and $p_B(\theta)$ show these prior distributions for A and B.

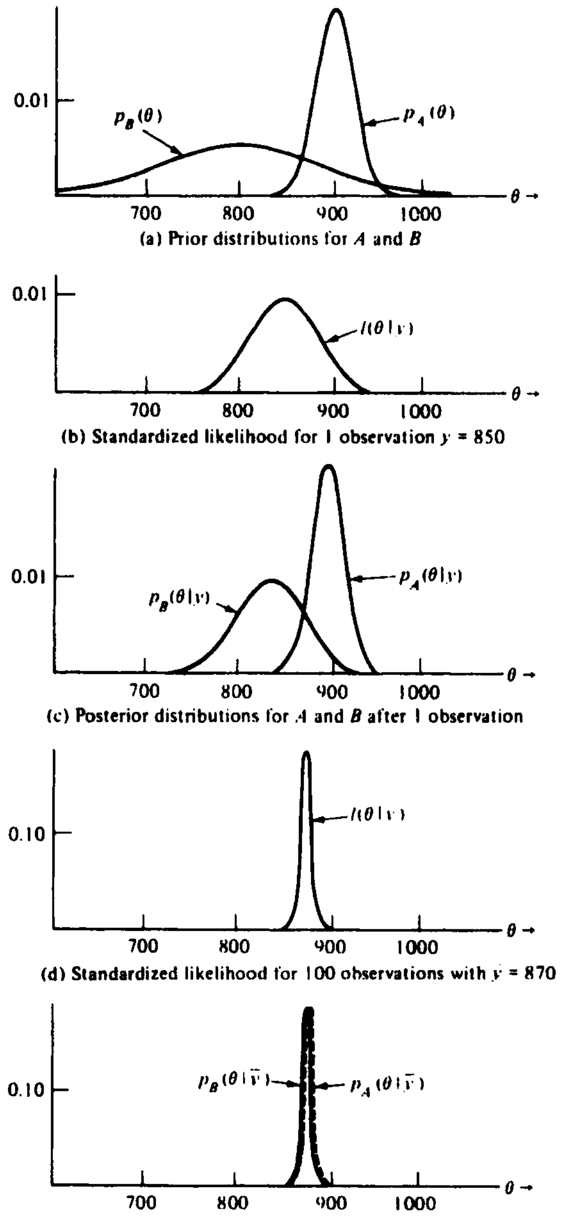
Suppose now that an unbiased method of experimental measurement is available and that an observation y made by this method, to a sufficient approximation, follows a Normal distribution with mean θ and standard deviation 40, that is $y \sim N(\theta, 40^2)$. If now a single observation y is made, the standardized likelihood function is represented by a Normal curve[†] centered at y with standard deviation 40. Then we can apply Bayes' theorem to show how each man's opinion regarding θ is modified by the information coming from that piece of data.

If *a priori* $\theta \sim N(\theta_0, \sigma_0^2)$, and the standardized likelihood function is represented by a Normal curve centered at y with standard deviation σ, then it is

[†] We refer to the function

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right]$$

as the Normal function, and the corresponding curve as the Normal curve. When the Normal function is employed to represent a probability distribution, it becomes the Normal distribution. The standardized likelihood function in this example is a Normal function, but it is not a probability distribution.



(a) Prior distributions for *A* and *B*

(b) Standardized likelihood for 1 observation $y = 850$

(c) Posterior distributions for *A* and *B* after 1 observation

(d) Standardized likelihood for 100 observations with $\bar{y} = 870$

(e) Posterior distributions for *A* and *B* after 100 observations

Fig. 1.2.1 Prior and posterior distributions for physicists *A* and *B*.

shown in Appendix A1.1 that the posterior distribution of θ given $y, p(\theta \mid y)$, is the Normal distribution $N(\bar{\theta}, \bar{\sigma}^2)$ where

$$\bar{\theta} = \frac{1}{w_0 + w_1}(w_0\theta_0 + w_1 y), \qquad \frac{1}{\bar{\sigma}^2} = w_0 + w_1$$

with

$$w_0 = \frac{1}{\sigma_0^2} \quad \text{and} \quad w_1 = \frac{1}{\sigma^2}. \tag{1.2.11}$$

The posterior mean $\bar{\theta}$ is a weighted average of the prior mean θ_0 and the observation y, the weights being proportional to w_0 and w_1 which are, respectively, the reciprocal of the variance of the prior distribution of θ and that of the observation. This is an appealing result, since the reciprocal of the variance is a measure of information which determines the weight to be attached to a given value. The variance of the posterior distribution is the reciprocal of the sum of the two measures of information w_0 and w_1, reflecting the fact that the two sources of information are pooled together.

Suppose the result of the single observation is $y = 850$; then the likelihood function is shown in Fig. 1.2.1(b). Physicist A's posterior opinion now is represented by the Normal distribution $p_A(\theta \mid y)$ with mean 890 and standard deviation 17.9, while that for B is represented by the Normal distribution $p_B(\theta \mid y)$ with mean 840 and standard deviation 35.78. These posterior distributions are shown in Fig. 1.2.1(c). The complete inferential process is sketched in Table 1.2.3.

Table 1.2.3

Prior and posterior distributions of θ for physicists A and B.

Prior distribution	Likelihood from data	Posterior distribution
A		A
$\theta \sim N(900, 20^2)$		$\theta \sim N(890, 17.9^2)$
	$N(850, 40^2)$	
B		B
$\theta \sim N(800, 80^2)$		$\theta \sim N(840, 35.70^2)$

We see that after this single observation the ideas of A and B about θ, as represented by the posterior distributions, are much closer than before, although they still differ considerably. We see that A, relatively speaking, did not learn much from the experiment, while B learned a great deal. The reason, of course, is that to A, the uncertainty in the measurement, as reflected by $\sigma = 40$, was larger than the uncertainty in his prior ($\sigma_0 = 20$). On the other hand, the uncertainty in the measurement was considerably smaller than that in B's prior ($\sigma_0 = 80$).

For A, the prior has a stronger influence on the posterior distribution than has the likelihood, while for B the likelihood has a stronger influence than the prior.

Suppose 99 further independent measurements are made and the sample mean $\bar{y} = \frac{1}{100}\Sigma y_i$ of the entire 100 observations is 870. In general, the likelihood function of θ *given* n independent observations from the Normal population $N(\theta, \sigma^2)$, is

$$l(\theta \mid \mathbf{y}) \propto \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\Sigma(y_i - \theta)^2\right]. \tag{1.2.12}$$

Also since

$$\Sigma(y_i - \theta)^2 = \Sigma(y_i - \bar{y})^2 + n(\theta - \bar{y})^2, \tag{1.2.13}$$

and, given the data, $\Sigma(y_i - \bar{y})^2$ is a fixed constant, the likelihood is

$$l(\theta \mid \mathbf{y}) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta - \bar{y}}{\sigma/\sqrt{n}}\right)^2\right], \tag{1.2.14}$$

which is a Normal function centred about $\bar{y}$ with standard deviation $\sigma/\sqrt{n}$.

In the present example, therefore, the likelihood is the Normal function centered at $\bar{y} = 870$ with standard deviation $\sigma/\sqrt{n} = \frac{40}{10} = 4$ shown in Fig. 1.2.1(d). We can thus apply the result in (1.2.11) as if $\bar{y}$ were a single observation with variance σ^2/n, that is, with weight n/σ^2. The posterior distribution of θ obtained by combining the likelihood function (1.2.14) with a Normal prior $N(\theta_0, \sigma_0^2)$ is the Normal distribution $N(\bar{\theta}_n, \bar{\sigma}_n^2)$, where

$$\bar{\theta}_n = \frac{1}{w_0 + w_n}(w_0\theta_0 + w_n\bar{y}), \qquad \frac{1}{\bar{\sigma}_n^2} = w_0 + w_n, \tag{1.2.15}$$

with

$$w_0 = \frac{1}{\sigma_0^2} \quad \text{and} \quad w_n = \frac{n}{\sigma^2}.$$

Thus the posterior distributions of A and B are $N(871.2, 3.9^2)$ and $N(869.8, 3.995^2)$, respectively. These two distributions, shown in Fig. 1.2.1e), are, for all practical purposes, the same, and are closely approximated by the Normal distribution $N(870, 4^2)$, which is the standardized form of the likelihood function in (1.2.14). Thus, after 100 observations, A and B would be in almost complete agreement. This is because the information coming from the data almost completely overrides prior differences.

Influence of the Prior Distribution on the Posterior Distribution

In the above example, we were concerned with the value of a *location* parameter θ, namely, the mean of a Normal distribution. In general, we shall say that a parameter η is a location parameter if addition of a constant c to all the observations changes η to $\eta + c$.

In this example, the contribution of the prior in helping to determine the posterior distribution of the location parameter θ was seen to depend on its sharpness or flatness *in relation* to the sharpness or flatness of the likelihood with which it was to be combined (see again Fig. 1.2.1). After a single observation, the likelihood was not sharply peaked relative to either of the prior distributions $p_A(\theta)$ or $p_B(\theta)$. These priors were therefore influential in deciding the posterior distribution. Because of this, the two different priors, when combined with the same likelihood, produced different posterior distributions. On the other hand, after 100 observations, both the priors $p_A(\theta)$ and $p_B(\theta)$ were rather flat *compared with* the likelihood function $l(\theta \mid \mathbf{y}) = l(\theta \mid \bar{y})$. These priors were therefore not very influential in deciding the corresponding posterior distributions of the location parameter θ. We can say that, after 100 observations, the priors were *dominated* by the likelihood.

1.2.4 Bayesian Decision Problems

The problems which we treat in this book are nearly all concerned with the situation common in scientific inference where the prior distribution is dominated by the likelihood. However, we must at least mention the important topic of Bayesian decision analysis [Schlaifer (1959), Raiffa and Schlaifer (1961), and DeGroot (1970)], where it is often not true that the prior is dominated by the likelihood. In Bayesian decision analysis, it is supposed that a choice has to be made from a set of available actions $(a_1, \ldots, a_r)$, where the payoff or utility of a given action depends on a state of nature, say θ, which is unknown. The decision maker's knowledge of θ is represented by a posterior distribution which combines prior knowledge of θ with the information provided by an experiment, and he is then supposed to choose that action which maximizes the *expected* payoff over the posterior distribution. An important application of such analysis is to business decision problems, such as whether or not to introduce a new industrial product. In such problems, a subjective prior distribution based, for example, on the opinion of an executive concerning the potential size θ of a market may be influential in determining the posterior distribution.

The fact that in such situations different decisions can result from different choices of prior distribution has worried some statisticians. We feel, however, that making explicit the dependence of the decision on the choice of what is believed to be true is an advantage of Bayesian analysis rather than the reverse. Suppose four different executives, after careful consideration, produce four different prior distributions for the size of a potential market and separate analyses are made for each. Then either (1) the decision (e.g. whether to market the product) will be the same in spite of differences in the priors, or (2) the decision will be different. In either case the Bayesian decision analysis will be valuable. In the first case, the ultimate arbiter would be reassured that such differences in opinion did not logically lead to differences on what the appropriate *action* should be. In the second case, it would be clear to him that *on present evidence* a real conflict existed. He would, in this case, either have to take the responsibility of ignoring the judgement of one or more of his executives, or of arranging that further data be obtained to resolve the conflict. Far from nullifying the value of Bayesian analysis, the fact that such analysis shows to what extent different decisions may or may not be appropriate when different prior opinions are held, seems to enhance it. For problems of this kind any procedure which took no account of such opinion would seem *necessarily* ill conceived.

1.2.5 Application of Bayesian Analysis to Scientific Inference

Important as the topic is, in this book, our concern will not be with statistical decision problems but with statistical inference problems such as occur in scientific investigation. By statistical inference we mean inference about the state of nature made in terms of probability, and a statistical inference problem is regarded as solved as soon as we can make an appropriate probability statement about the state of nature in question. Usually the state of nature is described by the value of one or more parameters. Such a parameter θ could, for example, be the velocity of light or the thermal conductivity of a certain alloy. Thus, a solution to the inference problem is supplied by a posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$ which shows what can be inferred about the parameters $\boldsymbol{\theta}$ from the data $\mathbf{y}$ given a relevant prior state of knowledge represented by $p(\boldsymbol{\theta})$.

Dominance of the Likelihood in the Normal Theory Example

Let us return again to the example of Section 1.2.3 concerning the estimation of the location parameter θ of a Normal distribution. In general, if the prior distribution is Normal $N(\theta_0, \sigma_0^2)$ and n independent observations with average $\bar{y}$ are taken from the distribution $N(\theta, \sigma^2)$, then from (1.2.15) the posterior distribution of θ is

$$\theta \sim N(\bar{\theta}_n, \bar{\sigma}_n^2), \quad \text{with} \quad \bar{\theta}_n = \frac{1}{w_0 + w_n}(w_0\theta_0 + w_n\bar{y}) \quad \text{and} \quad \bar{\sigma}_n^{-2} = w_0 + w_n,$$

where $w_0 = \sigma_0^{-2}$ is the weight associated with the prior distribution and $w_n = n/\sigma^2$ is the weight associated with the likelihood. In this expression, if w_0 is small *compared* with w_n, then *approximately* the posterior distribution is numerically equal to the standardized likelihood, and is

$$N\left(\bar{y}, \frac{\sigma^2}{n}\right). \tag{1.2.16}$$

Strictly speaking, this result is attained only when the prior variance σ_0^2 becomes infinite so that w_0 is zero. Such a limiting prior distribution would, however, by itself make little theoretical or practical sense. For, when $\sigma_0^2 \to \infty$, in the limit the prior density becomes uniform over the entire line from $-\infty$ to ∞, and is therefore not a proper density function. Furthermore, it represents a situation where all values of θ from $-\infty$ to ∞ are equally acceptable *a priori*. But it is difficult, if not impossible, to imagine a practical situation where sufficiently extreme values could not be virtually ruled out. The practical situation is represented *not* by the limiting case where $w_0 = 0$, but by the case where w_0 is small compared with w_n, that is, where the prior is locally flat so that the likelihood dominates the prior.

It is, therefore, important to note that the use of the limiting posterior in (1.2.16) corresponding to $w_0 = 0$ to supply a numerical approximation to the practical situation is not the same thing as *assuming* w_0 is actually zero. Limiting cases of this kind are frequently used in this book, but it must be remembered this is for the purpose of supplying a numerical approximation and for this purpose only.

"Proper" and "Improper" Prior Distributions

A basic property of a probability density function $f(x)$ is that it integrates or sums over its admissible range to 1, that is,

$$\left.\begin{matrix}\int f(x)\,dx \\ \Sigma f(x)\end{matrix}\right\} = 1 \qquad \left\{\begin{matrix}(x \text{ continuous}), \\ (x \text{ discrete}).\end{matrix}\right.$$

Now, if $f(x)$ is uniform over the entire line from $-\infty$ to ∞,

$$f(x) = \kappa, \qquad -\infty < x < \infty, \quad \kappa > 0, \tag{1.2.17}$$

then it is not a proper density since the integral

$$\int_{-\infty}^{\infty} f(x)\,dx = \kappa \int_{-\infty}^{\infty} dx$$

does not exist no matter how small κ is. Density functions of this kind are sometimes called *improper* distributions. As another example, the function

$$f(x) = \kappa x^{-1}, \qquad 0 < x < \infty, \quad \kappa > 0 \tag{1.2.18}$$

is also improper. In this book, density functions of the types in (1.2.17) and (1.2.18) are frequently employed to represent the *local* behavior of the prior distribution in the region where the likelihood is appreciable, but *not* over its entire admissible range. By supposing that to a sufficient approximation the prior follows the form (1.2.17) or (1.2.18) only over the range of appreciable likelihood and that it suitably tails to zero outside that range we ensure that the priors actually used are proper. Thus, by employing the distributions in a way that makes practical sense we are relieved of a theoretical difficulty.

The Role of the Dominant Likelihood in the Analysis of Scientific Experiments

It is often appropriate to analyze data from scientific investigations on the assumption that the likelihood dominates the prior. Two reasons for this are:

1. A scientific investigation is not usually undertaken unless information supplied by the investigation is likely to be considerably more precise than information already available. For instance, suppose a physical constant θ had been estimated at 0.85 ± 0.05; then usually there would be no justification for making

a new determination whose accuracy was ± 0.25,† but there might be considerable justification for making one whose accuracy was ± 0.01. In brief, a scientific investigation is not usually undertaken unless it is likely to increase knowledge by a substantial amount. Therefore, as is illustrated in Figs. 1.2.2 and 1.2.3, analysis with priors which are dominated by the likelihood often realistically represents the true inferential situation. Situations of this kind have been referred to by Savage (1962) and Edwards, Lindman, and Savage (1963) as those where the principle of "precise measurement" or "stable estimation" applies.

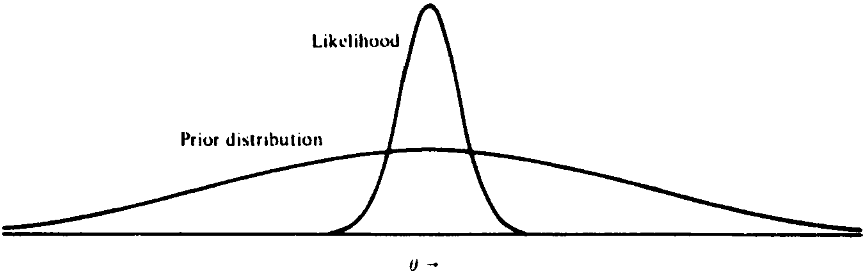


Fig. 1.2.2 Dominant likelihood (often appropriate to the analysis of scientific data).

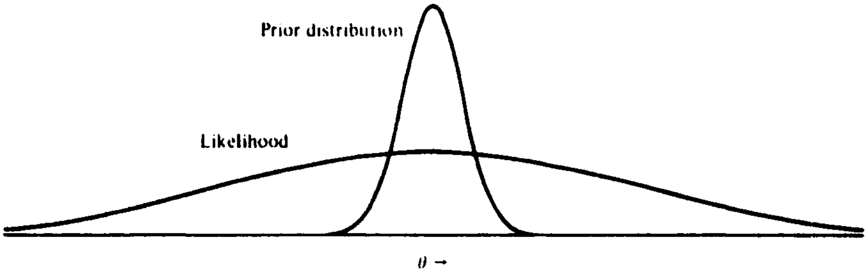


Fig. 1.2.3 Dominant prior (rarely appropriate to the analysis of scientific data).

2. Even when a scientist holds strong prior beliefs about the value of a parameter θ, nevertheless, in reporting his results it would usually be appropriate and most convincing to his colleagues if he analyzed the data against a *reference* prior which is dominated by the likelihood. He could then say that, irrespective of what he or anyone else believed to begin with, the posterior distribution represented what someone who *a priori* knew very little about θ should believe in the light of the data.‡

† Special circumstances could, of course, occur when the new determination *was* justified; for example, if it were suspected that the original method of determination might be subject to a major bias.

‡ As a *separate issue* his colleagues might also like to know what his prior opinion was and how this would affect the conclusions.

In judging the data in relation to a "neutral" reference prior, the scientist employs what may be called the "jury principle." Cases are tried in a law court before a jury which is carefully screened so that it has no possible connection with the principals and the events of the case. The intention is clearly to ensure that information gleaned from "data" or testimony may be assumed to dominate prior ideas that members of the jury may have concerning the possible guilt of the defendant.

The Reference Prior

In the above we have used the word *reference* prior. In general we mean by this a prior which it is convenient to use as a standard. In principle, a reference prior might or might not be dominated by the likelihood, but in this book reference priors which are dominated by the likelihood are often employed.

Dominant Likelihood and Locally Uniform Priors

The argument so far has been illustrated by the single example concerning the location parameter θ of a Normal distribution with a Normal prior. In particular, we have used this example to illustrate the important situation where the likelihood dominates the prior. We now consider the dominant likelihood idea more generally.

In general, a prior which is dominated by the likelihood is one which does not change *very much* over the region in which the likelihood is appreciable and does not assume large values outside that range (see Fig. 1.2.2). We shall refer to a prior distribution which has these properties as a *locally uniform* prior. For such a prior distribution we can approximate the result from Bayes' formula by substituting a constant for the prior distribution so that

$$p(\theta \mid \mathbf{y}) = \frac{l(\theta \mid \mathbf{y})\, p(\theta)}{\int l(\theta \mid \mathbf{y})\, p(\theta)\, d\theta} \doteq \frac{l(\theta \mid \mathbf{y})}{\int l(\theta \mid \mathbf{y})\, d\theta}. \qquad (1.2.19)$$

Thus, for a locally uniform prior, the posterior distribution is approximately numerically equal to the standardized likelihood as we have previously found in (1.2.16) for the very special case of a Normal prior dominated by a Normal likelihood.

Difficulties Associated with Locally Uniform Priors

Historically, the choice of a prior to characterize a situation where "nothing (or, more realistically, little) is known *a priori*" has long been, and still is, a matter of dispute. Bayes tentatively suggested that where such knowledge was lacking concerning the nature of the prior distribution, it might be regarded as uniform. This suggestion is usually referred to as Bayes' postulate. He seemed, however, to have been himself so doubtful as to the validity of this postulate that he did not publish it, and his work was presented (Bayes, 1763) to the Royal Society posthumously by his friend Richard Price. This was accompanied by Price's own commentary which might not have reflected Bayes' final view. Fisher (1959) pointed out that although Bayes considered this postulate in his essay, in his actual mathematics he avoided its use as open to dispute and showed by example how the prior distribution could be determined by an auxilliary experiment. The postulate was accepted without question by later writers such as Laplace, but its reckless application led unfortunately to the falling into disrepute of the theorem itself.

We now examine some objections which have been made to Bayes' postulate, and then discuss ways which have been proposed to overcome these objections and extend the concept. In refutation of Bayes' postulate, it has been argued that, if the distribution of a continuous parameter θ were taken locally uniform, then the distribution of $\log \theta, \theta^{-1}$, or some other transformation of θ (which might provide equally sensible bases for parametrizing the problem) would not be locally uniform. Thus, application of Bayes' postulate to different transformations of θ would lead to posterior distributions from the same data which were inconsistent.

This argument is of course correct, but the arbitrariness of the choice of parametrization does not by itself mean that we should not employ Bayes' postulate in practice. Arbitrariness exists to some extent in the specification of any statistical model. The only realistic expectation from a statistical analysis is that the conclusions will provide a good enough *approximation* to the truth. In applied (as opposed to pure) mathematics, arbitrariness is inadmissible only in so far as it produces results outside acceptable limits of approximation. In particular:

a) If, as would often be the case, the range of uncertainty for θ was not large compared with its mean value, then *over this range*, transformations such as the logarithmic and the reciprocal would be nearly linear, in which case approximate uniformity for θ would *imply* approximate uniformity for the transformed θ.

b) Although the argument (a) would fail for an extreme transformation such as θ^{10}, it is equally true that a rational experimenter would not agree to employ a uniform distribution after such a transformation. Thus, suppose that an investigator was concerned with measuring the specific gravity θ of a sample of ore; he expected that θ would be about 5 and felt happy with the idea that the probability that θ lay between 4 and 5 was about the same as the probability that θ lay between 5 and 6. A uniform distribution on θ^{10} would imply that the probability that it lay between 5 and 6 was almost six times as great as the probability that it lay between 4 and 5. Once he understood the implication of taking a constant prior distribution for this extreme transformation, he would be unwilling to accept it.

c) For large or even moderate-sized samples, fairly drastic modification of the prior distribution may only lead to minor modification of the posterior density. Thus, for independent observations $y_1, \ldots, y_n$, the posterior distribution can be written

$$p(\theta \mid y_1, \ldots, y_n) \propto p(\theta) \prod_{i=1}^{n} p(y_i \mid \theta). \tag{1.2.20}$$

and, for sufficiently large n, the n terms introduced by the likelihood will tend to overwhelm the single term contributed by the prior [see Savage, (1954)]. An illuminating illustration of the robustness of inference, under sensible modification of the prior, is provided by the study of Mosteller and Wallace (1964) on disputed authorship.

The above arguments indicate only that arbitrariness in the choice of the transformation in terms of which the prior is supposed locally uniform is often not catastrophic and that effects on the posterior distribution are likely to be of order n^{-1} and not of order 1 in relation to the data. For instance, we shall discuss in Chapter 2 a Bayesian derivation of Student's t distribution, and in so doing we must choose a prior distribution for the dispersion of the supposed Normal distribution of the observations. In various contexts, the dispersion of a Normal distribution can with some justification be measured in terms of $\sigma^2, \sigma, \log \sigma$, σ^{-1}, or σ^{-2}. Depending on which of these metrics are regarded as locally uniform, a t distribution is obtained having $n-3$, $n-2$, $n-1$, n, or $n+1$ degrees of freedom, respectively. What we have in this case is an uncertainty in the degrees of freedom (which in turn implies an uncertainty in the variance of the posterior distribution) of order n^{-1}. This degree of arbitrariness would not matter very much for large samples but it would have an appreciable effect for small samples. We are thus led to ask whether there is some way of eliminating, or at least reducing it so that the situation where "little is known *a priori*" can be more closely and meaningfully approximated.

1.3 NONINFORMATIVE PRIOR DISTRIBUTIONS

In this section we present an argument for choosing a particular metric in terms of which a locally uniform prior can be regarded as noninformative about the parameters. It is important to bear in mind that one can never be in a state of *complete* ignorance; further, the statement "knowing little *a priori*" can only have meaning *relative* to the information provided by an experiment. For instance, in Fig. 1.2.1, physicist A's prior knowledge is substantial compared with the information from a single observation but it is noninformative relative to that from a hundred observations. Now, a prior distribution is supposed to represent knowledge about parameters before the outcome of a projected experiment is known. Thus, the main issue is how to select a prior which provides little information relative to what is expected to be provided by the intended experiment. We consider first the case of a single parameter.

1.3.1 The Normal Mean θ (σ^2 Known)

Suppose $\mathbf{y}' = (y_1, \ldots, y_n)$ is a random sample from a Normal distribution $N(\theta, \sigma^2)$, where σ is a supposed known. Then, from (1.2.14), the likelihood function of θ is

$$l(\theta \mid \sigma, \mathbf{y}) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right] \qquad (1.3.1)$$

where, as before, $\bar{y}$ is the average of the observations. The standardized likelihood function of θ is graphically represented by a Normal curve located by $\bar{y}$, with standard deviation $\sigma/\sqrt{n}$. Figure 1.3.1(a) shows a set of standardized likelihood curves which could result from an experiment in which $n = 10$ and $\sigma = 1$. Three different situations are illustrated with data giving averages of $\bar{y} = 6$, $\bar{y} = 9$, and $\bar{y} = 12$. Now it could happen that the quantity of immediate scientific interest was not θ itself but the reciprocal $\kappa = \theta^{-1}$. In that case the likelihood is

$$l(\kappa \mid \sigma, \mathbf{y}) \propto \exp\left[-\frac{n}{2\sigma}(\kappa^{-1} - \bar{y})^2\right], \qquad (1.3.2)$$

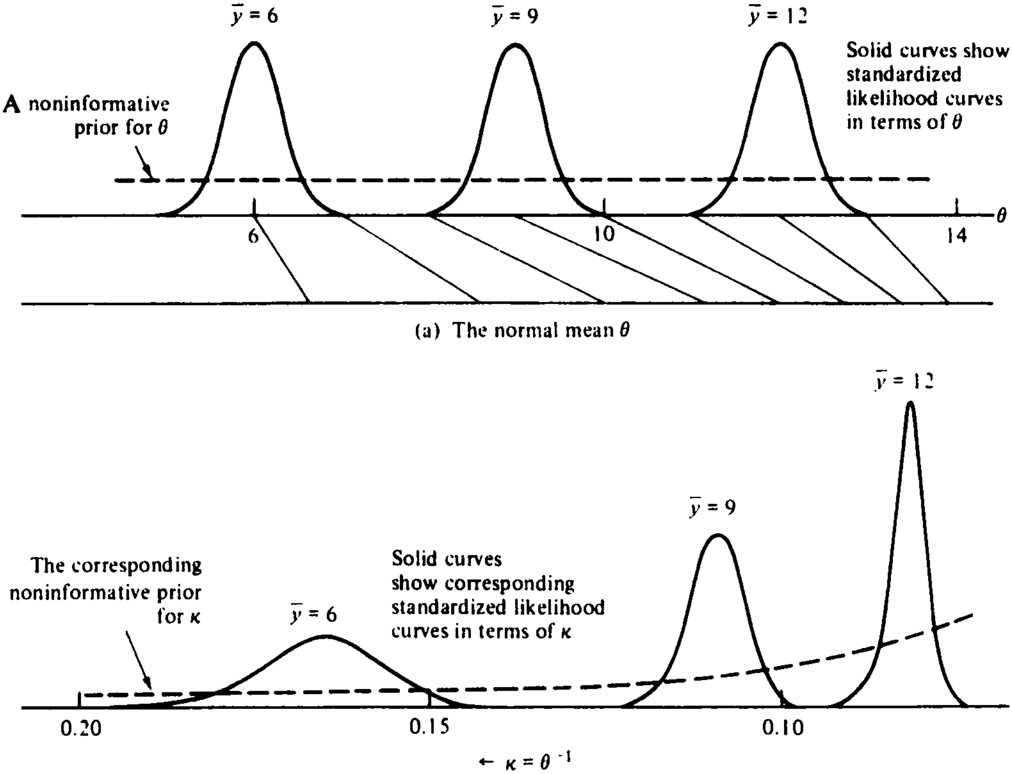
and the standardized likelihood curves would have the appearance shown in Fig. 1.3.1(b).

In our previous discussion of the Normal mean, the prior was taken to be locally uniform in θ, which implies of course that it is *not* uniform in κ. We now consider whether this choice can be justified, and whether the principle can be extended to a wider context.

Data Translated Likelihood and Non-informative Prior

Our problem is to express the idea that little is known *a priori* relative to what the data has to tell us about a parameter θ. What the data has to tell us about θ is expressed by the likelihood function, and in the case of the Normal mean with n and σ^2 known, the data enter the likelihood only via the sample average $\bar{y}$. Figure 1.3.1(a) illustrates how, when the likelihood is expressed in terms of θ, the sample average $\bar{y}$ affects only the *location* of the likelihood curve. Different sets of data *translate* the likelihood curve on the θ axis but leave it otherwise unchanged. On the other hand, Fig. 1.3.1(b) illustrates how, when the likelihood is expressed in terms of $\kappa = \theta^{-1}$, both the location and the spread of the likelihood curve are changed when the data (and hence $\bar{y}$) are changed.

Now, in general, suppose it is possible to express the unknown parameter θ in terms of a metric $\phi(\theta)$, so that the corresponding likelihood is *data translated.* This means that the likelihood curve for $\phi(\theta)$ is completely determined *a priori* *except for its location* which depends on the data yet to be observed. Then to say that we know little *a priori* relative to what the data is going to tell us, may be expressed by saying that we are almost equally willing to accept one value of $\phi(\theta)$ as another. This state of indifference may be expressed by taking $\phi(\theta)$ to be locally



(a) The normal mean θ

(b) Reciprocal of the normal mean $\kappa = \theta^{-1}$

Fig. 1.3.1 Noninformative prior distributions and standardized likelihood curves: (a) for the Normal mean θ, and (b) for $\kappa = \theta^{-1}$.

uniform, and the resulting prior distribution is called *noninformative* for $\phi(\theta)$ with respect to the data.

In the particular case of the Normal mean, the likelihood of θ is a Normal curve completely known *a priori* except for location which is determined by $\bar{y}$. That is, the likelihood is data translated in the original metric θ. Therefore, in this case, $\phi(\theta) = \theta$ and a noninformative prior is locally uniform in θ itself. That is, locally

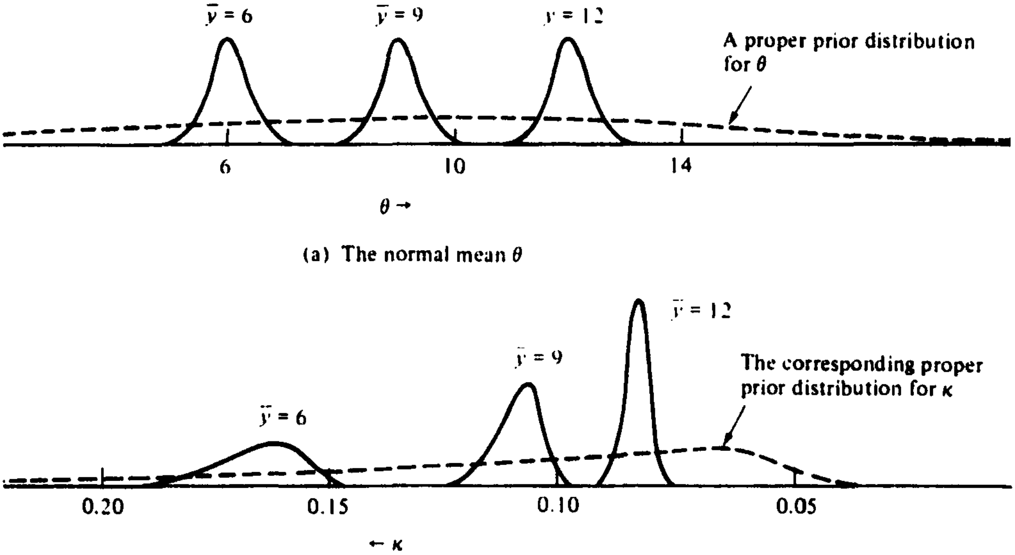
$$p(\theta \mid \sigma) \propto c. \tag{1.3.3}$$

This noninformative prior distribution is shown in Fig. 1.3.1(a) by the dotted line. Since

$$p(\kappa \mid \sigma) = p(\theta \mid \sigma)\left|\frac{d\theta}{d\kappa}\right| = p(\theta \mid \sigma)\theta^2 \propto \kappa^{-2}, \tag{1.3.4}$$

the corresponding noninformative prior for κ is not uniform but is locally proportional to θ^2, that is, to κ^{-2}. In general, if the noninformative prior is locally uniform in $\phi(\theta)$, then the corresponding noninformative prior for θ is locally proportional to $|d\phi/d\theta|$, assuming the transformation is one to one.

It is to be noted that we regard this argument only as indicating in what metric (transformation) the *local* behaviour of the prior should be uniform. Figure 1.3.2 illustrates what might be the situation over a wider range of the parameter. Here $p(\theta \mid \sigma)$ is a proper distribution which is merely flat over the region of interest. Similarly, $p(\kappa \mid \sigma)$ is a proper distribution obtained by transformation which is proportional to κ^{-2} over the region of interest. This point is important, because it would be inappropriate mathematically and meaningless practically to suppose, for example, that $p(\theta \mid \sigma)$ was uniform over an infinite range, or that $p(\kappa \mid \sigma)$ was proportional to κ^{-2} over an infinite range. We do not assume this nor do we need to.



(a) The normal mean θ

(b) Reciprocal of the normal mean $\kappa = \theta^{-1}$

Fig. 1.3.2 Noninformative prior distributions and standardized likelihood curves: (a) for the Normal mean θ, and (b) for $\kappa = \theta^{-1}$ seen over a wider range of parameter values.

Posterior Distribution of the Normal Mean θ

On multiplying the likelihood in (1.3.1) by the locally uniform noninformative prior in (1.3.3), and introducing the appropriate normalizing constant, we have

$$p(\theta \mid \sigma, \mathbf{y}) \doteq \left(\frac{2\pi\sigma^2}{n}\right)^{-1/2} \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right], \qquad -\infty < \theta < \infty. \qquad (1.3.5)$$

That is, when it is desired to assume little prior knowledge about θ relative to that which would be supplied from the data, and given a sample of n observations from a Normal distribution with known variance σ^2, then *a posteriori* θ is approximately Normally distributed with mean $\bar{y}$ and variance σ^2/n.

As an example, Fig. 1.3.3 shows the posterior distribution calculated from (1.3.5) when a sample of 16 observations has been taken whose average value is $\bar{y} = 10$, it being known that $\sigma = 8$. The figure shows θ distributed about $\bar{y} = 10$ with standard deviation $\sigma/\sqrt{n} = 2$. It is perhaps appropriate to emphasize the meaning which attaches to this distribution. To someone who, before the data was collected, was indifferent to the choice of θ in the relevant range, the posterior distribution represents what, given the data, his attitude should now be. He could, for example, state that the probability that θ was less than 8 was 15.9%, this being the size of the shaded area shown in the figure. Relative to the same state of prior indifference he could, moreover, employ the same posterior distribution of Fig. 1.3.3 to obtain, by transformation, the posterior distribution for any function $\kappa(\theta)$ which was of interest. For example he could state that the probability that κ was greater than 1/8 was 15.9%. Other probabilities are readily obtained by using a table of the Normal probability integral, such as Table I at the end of the book.

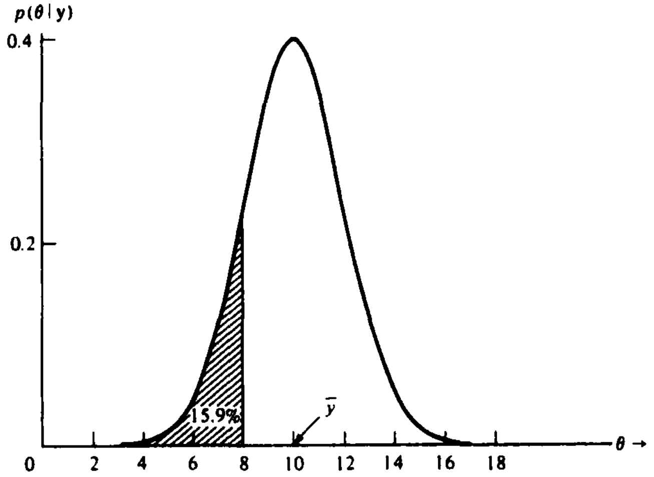


Fig. 1.3.3 Posterior distribution of the Normal mean θ (noninformative prior), when $\bar{y} = 10$, $\sigma = 8$, $n = 16$.

1.3.2 The Normal Standard Deviation σ (θ known)

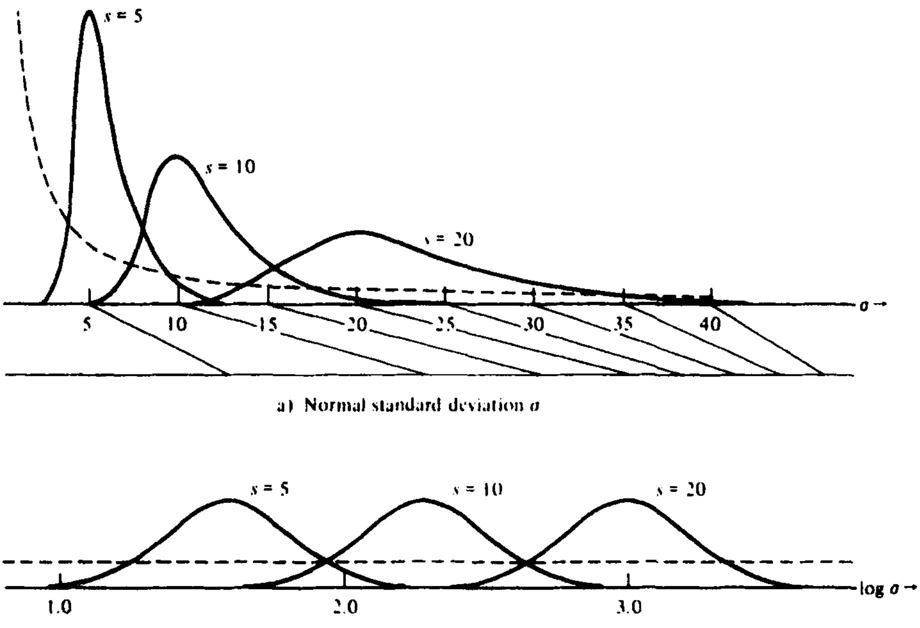
As a second example, consider the choice of a noninformative prior distribution for σ, the standard deviation of a Normal distribution for which the mean θ is supposed known. In this case, the likelihood is

$$l(\sigma \mid \theta, y) \propto \sigma^{-n} \exp\left(-\frac{ns^2}{2\sigma^2}\right), \tag{1.3.6}$$

where

$$s^2 = \Sigma\,(y_u - \theta)^2/n.$$

For illustration, suppose there are $n = 10$ observations, then Fig. 1.3.4(a) shows the standardized likelihood curves for σ with $s = 5$, $s = 10$, and $s = 20$. Clearly, *in the original metric σ, the likelihood curves are not data translated.* According to the principle stated in the preceding section therefore a noninformative prior should *not* be taken to be locally uniform in σ.



a) Normal standard deviation σ

b) Log of Normal standard deviation, log σ

Fig. 1.3.4 Noninformative prior distributions and standardized likelihood curves: (a) for the Normal standard deviation σ, and (b) for $\log \sigma$ (broken curves are noninformative priors and solid curves are the standard likelihoods).

Figure 1.3.4(b) shows, however, that the corresponding likelihood curves in terms of log σ are exactly data translated. To see this mathematically, note that multiplication by the constant s^n leaves the likelihood unchanged. Therefore we can express the likelihood of log σ as

$$l(\log\sigma \mid \theta, \mathbf{y}) \propto \exp\left\{-n(\log\sigma - \log s) - \frac{n}{2}\exp[-2(\log\sigma - \log s)]\right\}. \qquad (1.3.7)$$

Thus, in this logarithmic metric the data acting through s serve only to relocate the likelihood. A noninformative prior should therefore be locally uniform in $\log \sigma$. When expressed in the metric σ, the noninformative prior is thus locally proportional to σ^{-1},

$$p(\sigma \mid \theta) \propto \left| \frac{d \log \sigma}{d\sigma} \right| = \sigma^{-1}. \tag{1.3.8}$$

If we use this prior distribution, then the posterior distribution of σ is

$$p(\sigma \mid \theta, \mathbf{y}) \propto \sigma^{-(n+1)} \exp\left(-\frac{ns^2}{2\sigma^2}\right). \tag{1.3.9}$$

It will be seen in Section 2.3, where the implication of this distribution is discussed in greater detail, that the normalizing constant required to make the distribution integrate to unity is

$$k = \frac{(ns^2)^{n/2}}{2^{(n/2)-1}\,\Gamma(n/2)}. \tag{1.3.10}$$

Thus, given a sample $\mathbf{y}$ of n observations from a Normal distribution $N(\theta, \sigma^2)$, with θ known and little prior information about σ relative to that supplied by the data, the posterior distribution of σ is approximately

$$p(\sigma \mid \theta, \mathbf{y}) \doteq \frac{(ns^2)^{n/2}}{2^{(n/2)-1}\,\Gamma(n/2)} \sigma^{-(n+1)} \exp\left(-\frac{ns^2}{2\sigma^2}\right), \qquad \sigma > 0. \tag{1.3.11}$$

and the corresponding posterior distribution of any function of σ may be found by an appropriate transformation of (1.3.11).

Figure 1.3.5 illustrates the situation where the sample standard deviation calculated from $n = 10$ observations is

$$s = \left[\frac{\Sigma\,(y_u - \theta)^2}{10}\right]^{1/2} = 1.0.$$

The distribution shows what, given the assumptions and the data, can be said about σ. Tail area probabilities are readily found using the fact that (1.3.11) implies that ns^2/σ^2 has the "chi-square" (χ^2) distribution with n degrees of freedom,

$$p(\chi^2) = \frac{1}{\Gamma(n/2)2^{n/2}} (\chi^2)^{(n/2)-1} \exp\left(-\tfrac{1}{2}\chi^2\right), \qquad \chi^2 > 0. \tag{1.3.12}$$

For instance, suppose we wish to find the probability that σ is greater than $\sigma_0 = 1.5$. We have

$$\frac{ns^2}{\sigma_0^2} = \frac{10}{1.5^2} = 4.4,$$

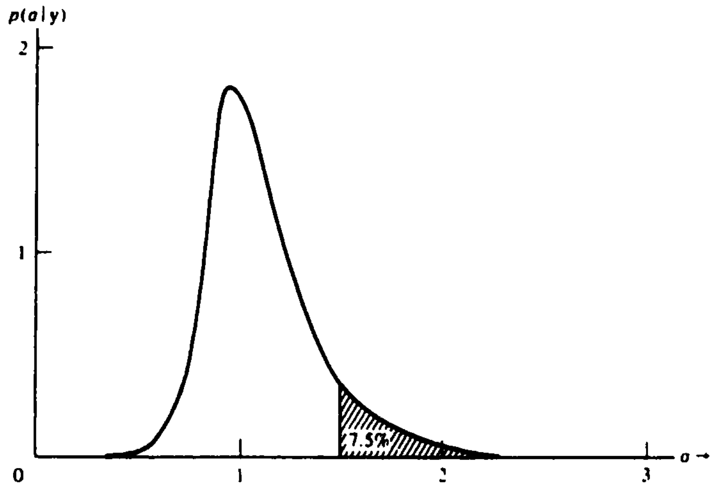


Fig. 1.3.5 Posterior distribution of the Normal standard deviation σ (noninformative prior), when $s = 1$ and $n = 10$.

so that, with χ^2_ν referring to a chi-square variate with ν degrees of freedom, the required probability corresponding to the shaded area in the diagram can be obtained from a table of χ^2 integral and is found to be

$$\Pr\{\chi^2_{10} < 4.4\} = 7.5\%.$$

1.3.3 Exact Data Translated Likelihoods and Noninformative Priors

We can summarize the above discussion of the choice of prior for a single parameter as follows.

If $\phi(\theta)$ is a one-to-one transformation of θ, we shall say that a prior distribution of θ which is locally proportional to $|d\phi/d\theta|$ is *noninformative* for the parameter θ if, in terms of ϕ, the likelihood curve is *data translated*, that is, the data only serve to change the location of the likelihood $l(\phi \mid \mathbf{y})$. Mathematically, a data translated likelihood must be expressible in the form

$$l(\theta \mid \mathbf{y}) = g\,[\phi(\theta) - f(\mathbf{y})], \qquad (1.3.13)$$

where $g(x)$ is a known function independent of the data $\mathbf{y}$ and $f(\mathbf{y})$ is a function of $\mathbf{y}$.

The examples we have so far considered are both special cases of the above principle. For the Normal mean, $\phi(\theta) = \theta, f(\mathbf{y}) = \bar{y}$, and for the Normal standard deviation, $\phi(\sigma) = \log \sigma$, $f(\mathbf{y}) = \log s$.

In particular, we see that any likelihood of the form

$$l(\sigma \mid \mathbf{y}) \propto l\left[\frac{s(\mathbf{y})}{\sigma}\right] \qquad (1.3.14)$$

can be bought into the form

$$l(\sigma \mid \mathbf{y}) = g[\log \sigma - \log s(\mathbf{y})] \tag{1.3.15}$$

so that it is data translated in terms of the logarithmic transformation $\phi(\sigma) = \log \sigma$.

The choice of a prior which is locally uniform in the metric ϕ for which the likelihood is data translated, can be viewed in another way. Let

$$l(\phi \mid \mathbf{y}) = g[\phi - f(\mathbf{y})], \tag{1.3.16}$$

and assume that the function g is continuous and has a unique maximum $\hat{g}$. Let α be an *arbitrary* positive constant such that $0 < \alpha < \hat{g}$. Then, for any given α, there exist two constants c_1 and c_2 $(c_1 < c_2)$, independent of $\mathbf{y}$ such that $g[\phi - f(\mathbf{y})]$ is greater than α for ϕ in the interval

$$f(\mathbf{y}) + c_1 < \phi < f(\mathbf{y}) + c_2. \tag{1.3.17}$$

This interval may be called the α highest likelihood interval. Now suppose the transformation from ϕ to λ is monotone. Then the corresponding α highest likelihood interval for λ is

$$\phi^{-1}[f(\mathbf{y}) + c_1] < \lambda < \phi^{-1}[f(\mathbf{y}) + c_2]. \tag{1.3.18}$$

We see that, in terms of ϕ, the length of the interval in (1.3.17) is $(c_2 - c_1)$ independent of the data $\mathbf{y}$, while for the metric λ the corresponding length

$$\phi^{-1}[f(\mathbf{y}) + c_2] - \phi^{-1}[f(\mathbf{y}) + c_1]$$

will in general depend upon $\mathbf{y}$ (except when the transformation is linear). For example, in the case of the Normal mean, $\phi(\theta) = \theta$, $f(\mathbf{y}) = \bar{y}$, and for $n = 10$, $\sigma = 1$,

$$g(x) = \exp\left(-\frac{n}{2\sigma^2}x^2\right) = \exp(-5x^2)$$

so that $\hat{g} = 1$. Suppose we take $\alpha = 0.05$; then

$$c_1 = -0.77, \qquad c_2 = 0.77.$$

For the three cases $\bar{y} = 6$, $\bar{y} = 9$, and $\bar{y} = 12$ considered earlier, the corresponding 0.05 highest likelihood intervals for θ are

$$\begin{array}{ccc} 6 \pm 0.77 & 9 \pm 0.77 & 12 \pm 0.77 \\ (5.23,\ 6.77), & (8.23,\ 9.77), & (11.23,\ 12.77), \end{array} \tag{1.3.19}$$

having the same length $c_2 - c_1 = 1.54$. However, in terms of the metric $\lambda = -\kappa = -1/\theta$, which is a monotone increasing function of θ, the 0.05 highest likelihood interval is

$$-(\bar{y} - 0.77)^{-1} < \lambda < -(\bar{y} + 0.77)^{-1},$$

so that for the three values of $\bar{y}$ considered we have

$\bar{y}$	6	9	12	
Interval	$(-0.191, -0.148)$	$(-0.122, -0.102)$	$(-0.089, -0.078)$	(1.3.20)
Length	0.043	0.020	0.011	

If we say we have little *a priori* knowledge about a parameter θ relative to the information expected to be supplied by the data, then we should be equally willing to accept the information from one experimental outcome as that from another. Since the information from the data is contained in the likelihood, this is saying that, over a relevant region of θ, we would have no *a priori* preference for one likelihood curve over another. This state of local indifference can then be represented by assigning approximately *equal* probabilities to all α-highest likelihood intervals. Now, in terms of ϕ for which the likelihood is data translated, the intervals all have the same length, so that the prior density must be locally uniform.

In the above example we would assign equal prior probabilities to the three intervals in (1.3.19), and the corresponding one in (1.3.20). It then follows that the noninformative prior distribution is locally uniform in θ but is locally proportional to $|d\theta/d\lambda| = \lambda^{-2}$ in terms of λ.

1.3.4 Approximate Data Translated Likelihood

As might be expected, a transformation which allows the likelihood to be expressed *exactly* in the form (1.3.13) is not generally available. However, for moderate sized samples, because of the insensitivity of the posterior distribution to minor changes in the prior, all that it would seem necessary to require is a transformation $\phi(\theta)$ in terms of which the likelihood is approximately data translated. That is to say, the likelihood for ϕ is nearly independent of the data $\mathbf{y}$ except for its location.

The Binomial Mean π

To illustrate the possibilities we consider the case of n independent trials, in each of which the probability of success is π. The probability of y successes in n trials is given by the binomial distribution

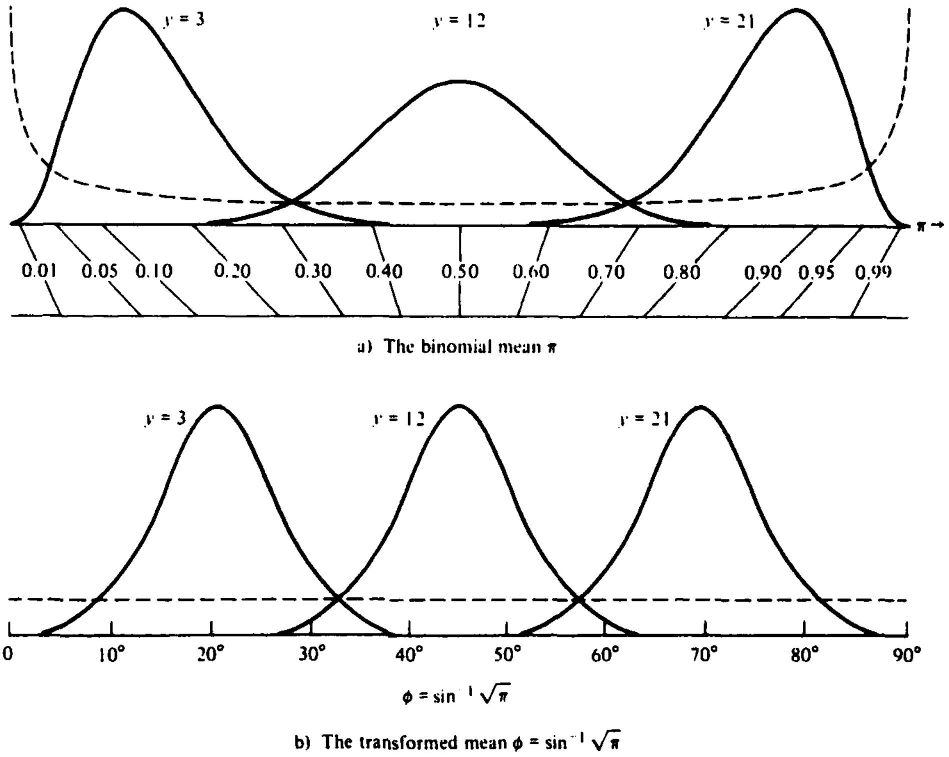
$$p(y \mid \pi) = \frac{n!}{y!\,(n-y)!}\,\pi^{y}(1-\pi)^{n-y}, \qquad y = 0, \ldots, n, \tag{1.3.21}$$

so that the likelihood is

$$l(\pi \mid y) \propto \pi^{y}(1-\pi)^{n-y}. \tag{1.3.22}$$

Suppose for illustration there are $n = 24$ trials. Then Fig. 1.3.6(a) shows the standardized likelihood for $y = 3$, $y = 12$, and $y = 21$ successes. Figure 1.3.6(b) is the corresponding diagram obtained by plotting in the transformed metric

$$\phi(\pi) = \sin^{-1}\sqrt{\pi}. \tag{1.3.23}$$



a) The binomial mean π

b) The transformed mean $\phi = \sin^{-1}\sqrt{\pi}$

Fig. 1.3.6 Noninformative prior distributions and standardized likelihood curves: (a) for the binomial mean π, and (b) for the transformed mean $\phi = \sin^{-1}\sqrt{\pi}$ (broken curves are the noninformative priors and solid curves are standardized likelihoods).

Although in terms of ϕ the likelihood curves are not exactly identical in shape and spread, they are nearly so. In this metric the likelihood curve is very nearly data translated and a locally uniform prior distribution is nearly noninformative. This in turn implies that the corresponding nearly noninformative prior for π is proportional to

$$p(\pi) \propto \left|\frac{d\phi}{d\pi}\right| = [\pi(1 - \pi)]^{-\frac{1}{2}}. \qquad (1.3.24)$$

If we employ this approximately noninformative prior, indicated in Fig. 1.3.6(a) and (b) by the dotted lines, then as was noted by Fisher (1922),

$$p(\pi \mid y) \propto \pi^{y-\frac{1}{2}}(1 - \pi)^{n-y-\frac{1}{2}}, \qquad 0 < \pi < 1. \qquad (1.3.25)$$

After substitution of the appropriate normalizing constant, we find that the corresponding posterior distribution for π is the beta distribution

$$p(\pi \mid y) = \frac{\Gamma(n+1)}{\Gamma(y+\frac{1}{2})\,\Gamma(n-y+\frac{1}{2})}\,\pi^{y-\frac{1}{2}}(1-\pi)^{n-y-\frac{1}{2}} \qquad 0 < \pi < 1. \qquad (1.3.26)$$

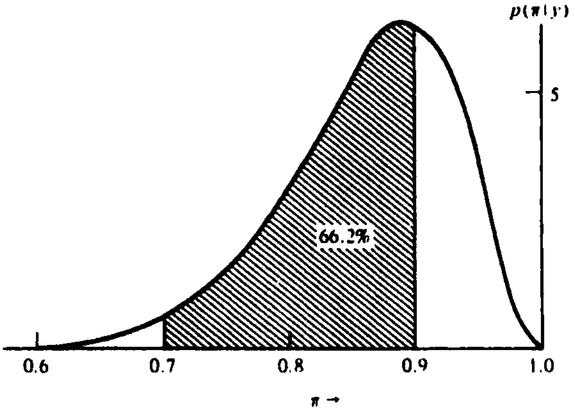


Fig. 1.3.7 Posterior distribution of the binomial mean π (noninformative prior) for 21 successes out of 24 trials.

Figure 1.3.7 shows the posterior distribution of π given that 21 out of 24 binomial trials (a proportion of 0.875) are successes. For illustration, tail area probabilities can be obtained by consulting the incomplete beta function tables.

The shaded area shown in the diagram is the probability that the parameter π lies between 0.7 and 0.9, and this is given by

$$\int_{0.7}^{0.9} \frac{\Gamma(25)}{\Gamma(21.5)\Gamma(3.5)}\,\pi^{20.5}(1-\pi)^{2.5}\,d\pi = 66.2\%.$$

We note in passing that, for this example where we have a moderately sized sample of $n = 24$ observations, the posterior density is not very sensitive to the precise choice of a prior. For instance, while for 21 successes the noninformative prior (1.3.24) yielded a posterior density proportional to $\pi^{20.5}(1-\pi)^{2.5}$, for a uniform prior in the original metric π the posterior density would have been proportional to $\pi^{21}(1-\pi)^{3}$. The use of the noninformative prior for π, rather than the uniform prior, is in general merely equivalent to reducing the number of successes and the number of "not successes" by 0.5.

Derivation of Transformations Yielding Approximate Data Translated Likelihoods

We now consider methods for obtaining parameter transformations in terms of which the likelihood is approximately data translated as in the binomial case. Again, let $\mathbf{y}' = (y_1, \ldots, y_n)$ be a random sample from a distribution $p(y \mid \theta)$. When the distribution obeys certain regularity conditions, Johnson (1967, 1970), then for sufficiently large n, the likelihood function of θ is approximately Normal, and remains approximately Normal under mild one-to-one transformations of θ. In such a case, the logarithm of the likelihood is approximately quadratic, so that

$$
\begin{aligned}
L(\theta \mid y) &= \log l(\theta \mid y) = \log \prod_{u=1}^{n} p(y_u \mid \theta) \\
&\doteq L(\hat{\theta} \mid y) - \frac{n}{2}(\theta - \hat{\theta})^2 \left(-\frac{1}{n} \frac{\partial^2 L}{\partial \theta^2} \right)_{\hat{\theta}}. \qquad (1.3.27)
\end{aligned}
$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ. In general, the quantity

$$\left(-\frac{1}{n} \frac{\partial^2 L}{\partial \theta^2} \right)_{\hat{\theta}}$$

is a positive function of y. For the moment we shall discuss the situation in which it can be expressed as *a function of $\hat{\theta}$ only*, and write

$$J(\hat{\theta}) = \left(-\frac{1}{n} \frac{\partial^2 L}{\partial \theta^2} \right)_{\hat{\theta}}. \qquad (1.3.28)$$

Now the logarithm of a Normal function $p(x)$ is of the form

$$\log p(x) = \text{const} - \tfrac{1}{2}(x - \mu)^2/\sigma^2 \qquad (1.3.29)$$

and, given the location parameter μ, is completely determined by its standard deviation σ. Comparison of (1.3.27) and (1.3.29) shows that the standard deviation of the likelihood curve is approximately equal to $n^{-\frac{1}{2}} J^{-\frac{1}{2}}(\hat{\theta})$. Now suppose $\phi(\theta)$ is a one-to-one transformation; then,

$$J(\hat{\phi}) = \left(-\frac{1}{n} \frac{\partial^2 L}{\partial \phi^2} \right)_{\hat{\phi}} = \left(-\frac{1}{n} \frac{\partial^2 L}{\partial \theta^2} \right)_{\hat{\theta}} \left(\frac{d\theta}{d\phi} \right)^2_{\hat{\theta}} = J(\hat{\theta}) \left(\frac{d\theta}{d\phi} \right)^2_{\hat{\theta}}. \qquad (1.3.30)$$

It follows that if $\phi(\hat{\theta})$ is chosen such that

$$\left| \frac{d\theta}{d\phi} \right|_{\hat{\theta}} \propto J^{-1/2}(\hat{\theta}), \qquad (1.3.31)$$

then $J(\hat{\phi})$ will be a constant independent of $\hat{\phi}$, and the likelihood will be approximately data translated in terms of ϕ. Thus, the metric for which a locally uniform prior is approximately noninformative can be obtained from the relationship

$$\frac{d\phi}{d\theta} \propto J^{1/2}(\theta) \qquad \text{or} \qquad \phi \propto \int^{\theta} J^{1/2}(t)\, dt. \qquad (1.3.32)$$

This, in turn, implies that the corresponding noninformative prior for θ is

$$p(\theta) \propto \left|\frac{d\phi}{d\theta}\right| \propto J^{1/2}(\theta). \tag{1.3.33}$$

As an example, consider again the binomial mean π. The log likelihood is

$$L(\pi \mid y) = \log l(\pi \mid y) = \text{const} + y \log \pi + (n - y) \log (1 - \pi). \tag{1.3.34}$$

Thus

$$\frac{\partial L}{\partial \pi} = \frac{y}{\pi} - \frac{n - y}{1 - \pi}, \qquad \frac{\partial^2 L}{\partial \pi^2} = -\frac{y}{\pi^2} - \frac{n - y}{(1 - \pi)^2}. \tag{1.3.35}$$

For $y \neq 0$ and $y \neq n$, by setting $\partial L/\partial \pi = 0$, one obtains the maximum likelihood estimates as $\hat{\pi} = y/n$, so that

$$J(\hat{\pi}) = \left(-\frac{1}{n}\frac{\partial^2 L}{\partial \pi^2}\right)_{\hat{\pi}} = \left(\frac{1}{\hat{\pi}} + \frac{1}{1 - \hat{\pi}}\right) = \frac{1}{\hat{\pi}(1 - \hat{\pi})}, \tag{1.3.36}$$

which is a function of $\hat{\pi}$ only, whence the noninformative prior for π is proportional to

$$J^{1/2}(\pi) \propto \pi^{-1/2}(1 - \pi)^{-1/2}, \tag{1.3.37a}$$

which is the prior used in (1.3.24). Also, the transformation

$$\phi = \int^{\pi} t^{-1/2}(1 - t)^{-1/2}\, dt \propto \sin^{-1}\sqrt{\pi} \tag{1.3.37b}$$

is precisely the metric employed in plotting the nearly data translated likelihood curves in Fig. 1.3.6. We recognize the $\sin^{-1}\sqrt{\pi}$ transformation as the well-known asymptotic variance stabilizing transformation for the binomial, originally proposed by Fisher. [See, for example, Bartlett (1937) and Anscombe (1948a)].

In the above we have specifically supposed that the quantity

$$\left(-\frac{1}{\pi}\frac{\partial^2 L}{\partial \theta^2}\right)_{\hat{\theta}}$$

is a function of $\hat{\theta}$ only. It can be shown that this will be true whenever the observations y are drawn from a distribution $p(y \mid \theta)$ of the form

$$p(y \mid \theta) = h(y)w(\theta) \exp [c(\theta)u(y)], \tag{1.3.38}$$

where the range of y does not depend upon θ. For the cases of the Normal mean θ with σ^2 known, the Normal standard deviation σ with θ known and the binomial mean π, the distributions are of this form. In fact, this is the form for which a single sufficient statistic for θ exists, a concept which will be discussed later in Section 1.4.

The Poisson Mean λ

As a further example, consider the Poisson distribution with mean λ,

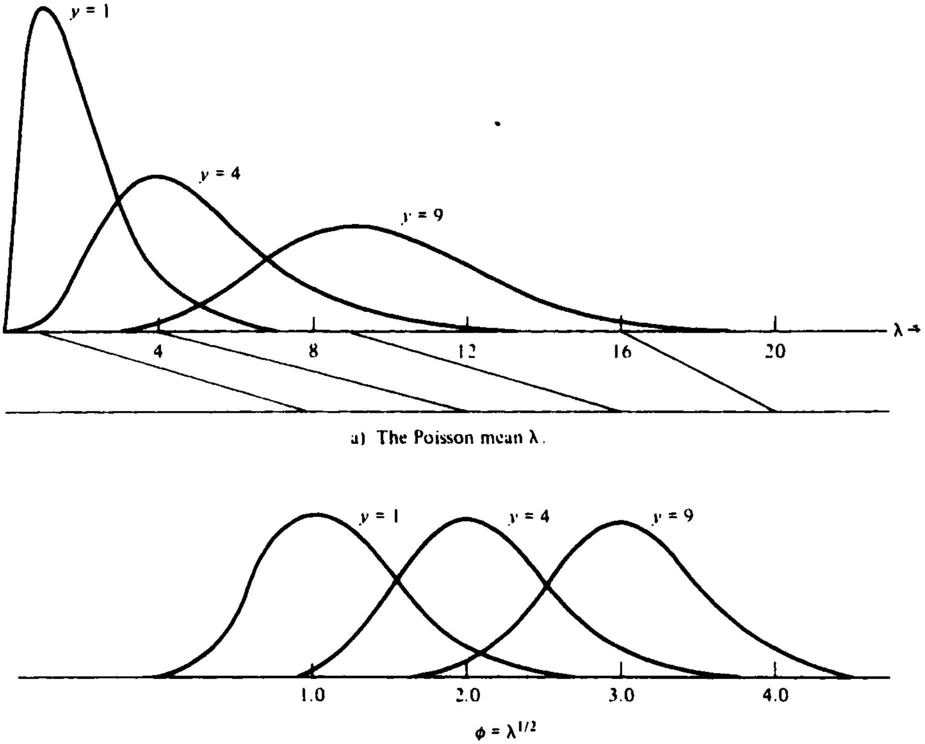
$$p(y \mid \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda), \qquad y = 0, \ldots, \infty, \tag{1.3.39}$$

which is of the form in (1.3.38). Suppose $\mathbf{y}' = (y_1, \ldots, y_n)$ is a set of n independent frequencies each distributed as (1.3.39). Then, given $\mathbf{y}$, the likelihood is

$$l(\lambda \mid \mathbf{y}) \propto \lambda^{n\bar{y}} \exp(-n\lambda), \qquad \bar{y} = \frac{1}{n} \Sigma y_u. \tag{1.3.40}$$

Thus,

$$L(\lambda \mid \mathbf{y}) = \text{const} + n\bar{y} \log \lambda - n\lambda \tag{1.3.41}$$



a) The Poisson mean λ.

b) The transformed mean $\phi = \lambda^{1/2}$

Fig. 1.3.8 Standardized likelihood curves: (a) for the Poisson mean λ, and (b) for the transformed mean $\phi = \lambda^{1/2}$.

and

$$\frac{\partial L}{\partial \lambda} = \frac{n\bar{y}}{\lambda} - n, \qquad \frac{\partial^2 L}{\partial \lambda^2} = \frac{-n\bar{y}}{\lambda^2}.$$

For $\bar{y} \neq 0$, the maximum likelihood estimate of λ obtained from $\partial L/\partial \lambda = 0$ is $\hat{\lambda} = \bar{y}$ so that

$$J(\hat{\lambda}) = \left(-\frac{1}{n}\frac{\partial^2 L}{\partial \lambda^2}\right)_{\hat{\lambda}} = \frac{1}{\hat{\lambda}}. \tag{1.3.42}$$

According to (1.3.33), a noninformative prior for λ is

$$p(\lambda) \propto J^{1/2}(\lambda) \propto \lambda^{-1/2}, \tag{1.3.43}$$

and $\phi = \lambda^{1/2}$ is the metric for which the approximate noninformative prior is locally uniform. The effectiveness of the transformation in achieving data translated curves is illustrated in Fig. 1.3.8(a) and (b), with $n = 1$ and $\bar{y} = y = 1$, $y = 4$, and $y = 9$.

Using the noninformative prior (1.3.43), the posterior distribution of λ is

$$p(\lambda \mid y) = c\lambda^{n\bar{y}-\frac{1}{2}} \exp(-n\lambda), \qquad \lambda > 0, \tag{1.3.44}$$

where, on integration, the Normalizing constant is found to be

$$c = n^{-(n\bar{y}+\frac{1}{2})}[\Gamma(n\bar{y}+\tfrac{1}{2})]^{-1}.$$

Equivalently, we have that $n\lambda$ is distributed as $\frac{1}{2}\chi^2$ with $2n\bar{y}+1$ degrees of freedom.

Figure 1.3.9 shows the posterior distribution of λ, given that $n = 1$ and a frequency of $y = 2$ has been observed, where little is known about λ *a priori*. The shaded area

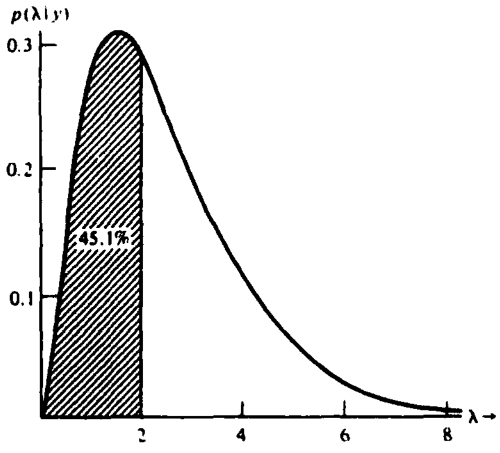


Fig. 1.3.9 Posterior distribution of the Poisson mean λ (noninformative prior) for an observed frequency $y = 2$.

corresponds to the probability that $\lambda < 2$ which is

$$\Pr\{\tfrac{1}{2}\chi_5^2 < 2\} = \Pr\{\chi_5^2 < 4\} = 45.1\%.$$

1.3.5 Jeffreys' Rule, Information Measure, and Noninformative Priors

In general, the distribution $p(y \mid \theta)$ need not belong to the family defined by (1.3.38), and the quantity

$$\left(-\frac{1}{n}\frac{\partial^2 L}{\partial \theta^2}\right)_{\hat\theta},$$

in (1.3.27) is a function of all the data $\mathbf{y}$. The argument leading to the approximate noninformative prior in (1.3.33) can then be modified as follows.

It is to be noted that, for given θ,

$$-\frac{1}{n}\frac{\partial^2 L}{\partial \theta^2} = -\frac{1}{n}\sum_{u=1}^{n}\frac{\partial^2 \log p(y_u \mid \theta)}{\partial \theta^2} \tag{1.3.45}$$

is the average of n identical functions of $(y_1, \ldots, y_n)$, respectively. Now suppose θ_0 is the true value of θ so that $\mathbf{y}$ are drawn from the distribution $p(\mathbf{y} \mid \theta_0)$. It then follows that, for large n, the average converges in probability to the expectation of the function, that is, to

$$\underset{y \mid \theta_0}{E}\left[-\frac{\partial^2 \log p(y \mid \theta)}{\partial \theta^2}\right] = -\int \frac{\partial^2 \log p(y \mid \theta)}{\partial \theta^2}\, p(y \mid \theta_0)\, dy = a(\theta, \theta_0),$$

assuming that the expectation exists. Also, for large n, the maximum likelihood estimate $\hat\theta$ converges in probability to θ_0. Thus, we can write, approximately,

$$\left(-\frac{1}{n}\frac{\partial^2 L}{\partial \theta^2}\right)_{\hat\theta} \doteq a(\hat\theta, \theta_0) \doteq a(\hat\theta, \hat\theta) = \mathscr{J}(\hat\theta), \tag{1.3.46}$$

where $\mathscr{J}(\theta) = a(\theta, \theta)$ is the function

$$\mathscr{J}(\theta) = -\underset{y \mid \theta}{E}\left[\frac{\partial^2 \log p(y \mid \theta)}{\partial \theta^2}\right] = \underset{y \mid \theta}{E}\left[\frac{\partial \log p(y \mid \theta)}{\partial \theta}\right]^2. \tag{1.3.47}$$

Consequently, if we use $\mathscr{J}(\hat\theta)$, which depends on $\hat\theta$ only, to approximate

$$\left(-\frac{1}{n}\frac{\partial^2 L}{\partial \theta^2}\right)_{\hat\theta}$$

in (1.3.27), then, arguing exactly as before, we find that the metric $\phi(\theta)$ for which a locally uniform prior is approximately noninformative is such that

$$\frac{d\phi}{d\theta} \propto \mathscr{J}^{1/2}(\theta) \qquad \text{or} \qquad \phi \propto \int^{\theta} \mathscr{J}^{1/2}(t)\, dt. \tag{1.3.48}$$