Bayesian Inference in Statistical Analysis
by George E.P. Box and George C. Tiao

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, Barnard, 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 by sampling 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.

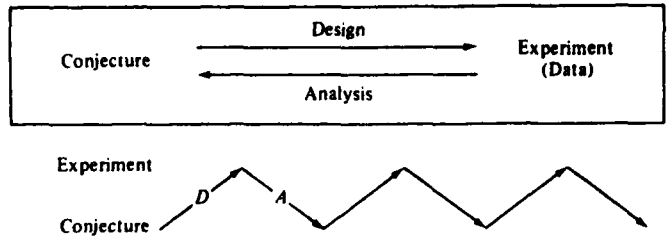


Fig. 1.1.1 Iterative process of scientific investigation (the alternation between conjecture and experiment).

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 $\hat{\theta}(\mathbf{y})$ and $\hat{\sigma}^2(\mathbf{y})$, which are functions of the data vector $\mathbf{y}$, are selected. By imagining values $\hat{\theta}(\mathbf{y}_j)$ and $\hat{\sigma}^2(\mathbf{y}_j)$ to be calculated for each hypothetical data vector $\mathbf{y}_j$, reference sets are generated for $\hat{\theta}(\mathbf{y})$ and $\hat{\sigma}^2(\mathbf{y})$. Inferences are then made by comparing the values of $\hat{\theta}(\mathbf{y})$ and $\hat{\sigma}_2(\mathbf{y})$ actually observed with their "sampling distributions" generated by the reference sets.

The functions $\hat{\theta}(\mathbf{y})$ and $\hat{\sigma}^2(\mathbf{y})$ are usually chosen so that the sampling distributions of the extimators $\hat{\theta}(\mathbf{y}_j)$ and $\hat{\sigma}^2(\mathbf{y}_j)$ 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 $\hat{\theta}(\mathbf{y})$ and $\hat{\sigma}^2(\mathbf{y})$ might be, *confidence intervals* are calculated. For example, the $1-\alpha$ confidence interval for θ would be of the form

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

where $\bar{\theta}_1(\mathbf{y})$ and $\bar{\theta}_2(\mathbf{y})$ would be functions of $\mathbf{y}$, 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 $p(\mathbf{y} \mid \theta, \sigma^2)$ and the data $\mathbf{y}$, it is now possible to calculate the probability distribution $p(\theta, \sigma^2 \mid \mathbf{y})$ of θ and σ^2, given the data $\mathbf{y}$. 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

$$p(\mathbf{y} \mid \mathbf{x}, \sigma^2, \theta_1, \theta_2) \propto \sigma^{-20} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{20}(y_i - \theta_1 - \theta_2 x_i)^2\right],$$

$$-\infty < y_i < \infty. \quad (1.1.2)$$

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

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

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.

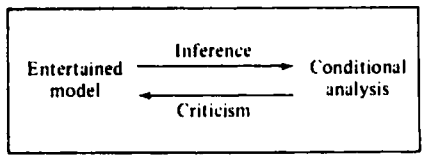


Fig. 1.1.2 Statistical analysis of data as an iterative process of model building.

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,

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

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

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

Also, we can write

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

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

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = cp(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1.2.4}$$

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

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

In other words, then, Bayes' theorem tells us that the probability distribution for θ posterior to the data y is proportional to the product of the distribution for θ prior to the data and the likelihood for θ given y. That is,

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

The likelihood function $l(\theta \mid y)$ plays a very important role in Bayes' formula. It is *the* function through which the data y modifies prior knowledge of θ; it can therefore be regarded as representing the information about θ 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 θ. 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 $\int l(\theta \mid y)\, d\theta$, taken over the admissible range of θ, is finite, then occasionally it will be convenient to refer to the quantity

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

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 θ as more observations are taken.

Thus, suppose we have an initial sample of observations y_1, then Bayes' formula gives

$$p(\theta \mid y_1) \propto p(\theta) l(\theta \mid y_1). \tag{1.2.7}$$

Now, suppose we have a second sample of observations y_2 distributed independently of the first sample, then

$$\begin{aligned} p(\theta \mid y_2, y_1) &\propto p(\theta) l(\theta \mid y_1) l(\theta \mid y_2) \\ &\propto p(\theta \mid y_1) l(\theta \mid y_2). \end{aligned} \tag{1.2.8}$$

The expression (1.2.8) is precisely of the same form as (1.2.7) except that $p(\theta \mid y_1)$, the posterior distribution for θ given 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

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

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 Fisher 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 $\mathbf{y}' = (y_1, \ldots, y_7)$, 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.

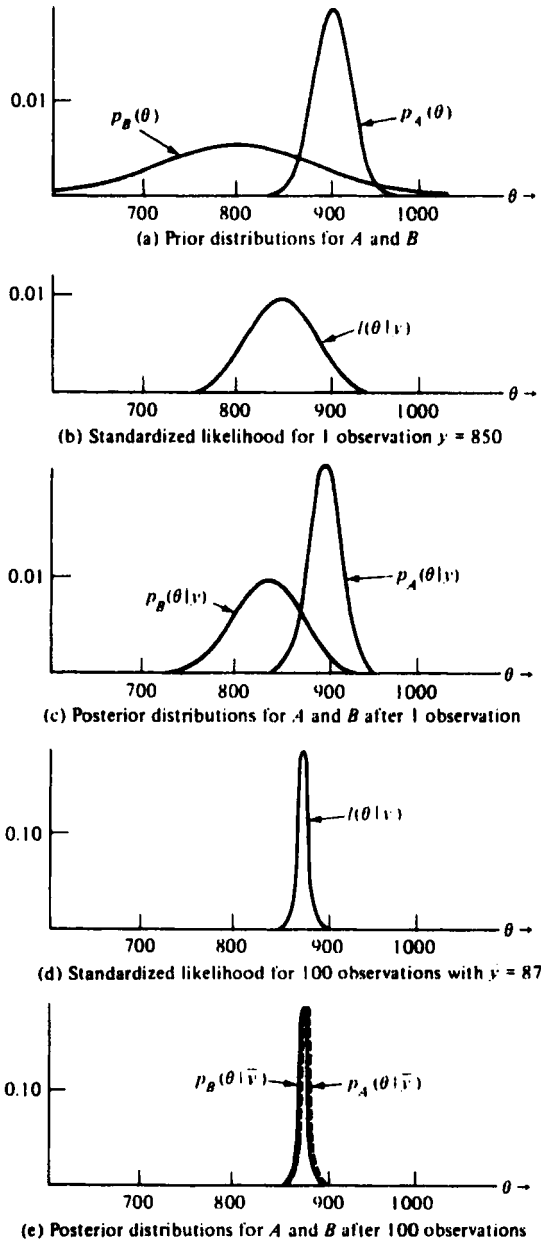


Fig. 1.2.1 Prior and posterior distributions for physicists A and B.

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: $\theta \sim N(900, 20^2)$	$N(850, 40^2)$	A: $\theta \sim N(890, 17.9^2)$
B: $\theta \sim N(800, 80^2)$		B: $\theta \sim N(840, 35.70^2)$

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 $\sigma/\sqrt{n} = \frac{400}{100} = 4$ 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 $l(\theta \mid y) = l(\theta \mid \bar{y})$. 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,

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

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 $\pm$ 0.25,† but there might be considerable justification for making one whose accuracy was $\pm$ 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.

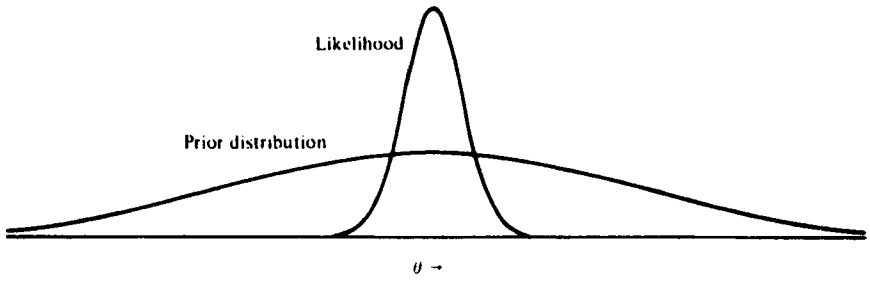


Fig. 1.2.2 Dominant likelihood (often appropriate to the analysis of scientific data).

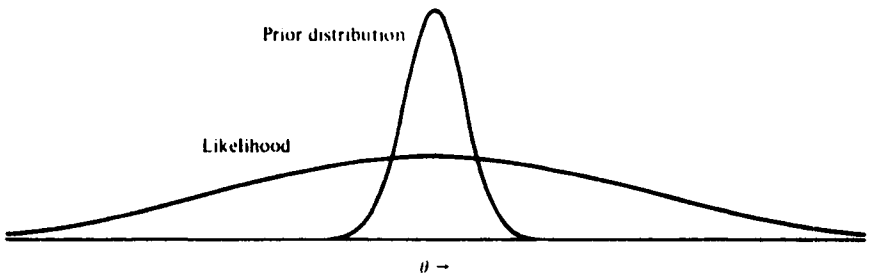


Fig. 1.2.3 Dominant prior (rarely appropriate to the analysis of scientific data).

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

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

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

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

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

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

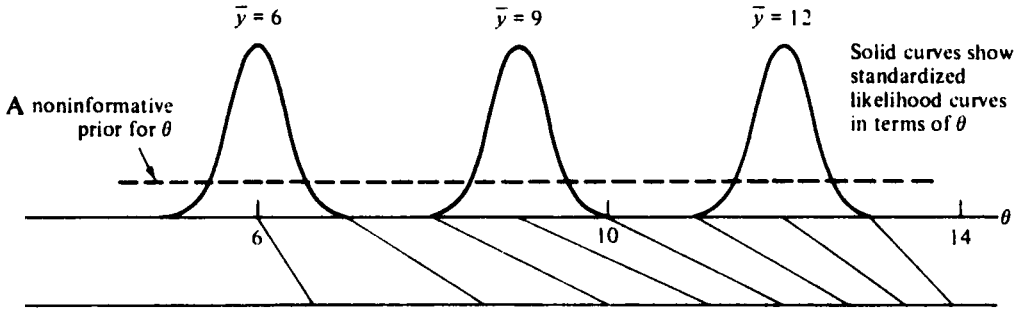
and the standardized likelihood curves would have the appearance shown in Fig. 1.3.1(b).

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.

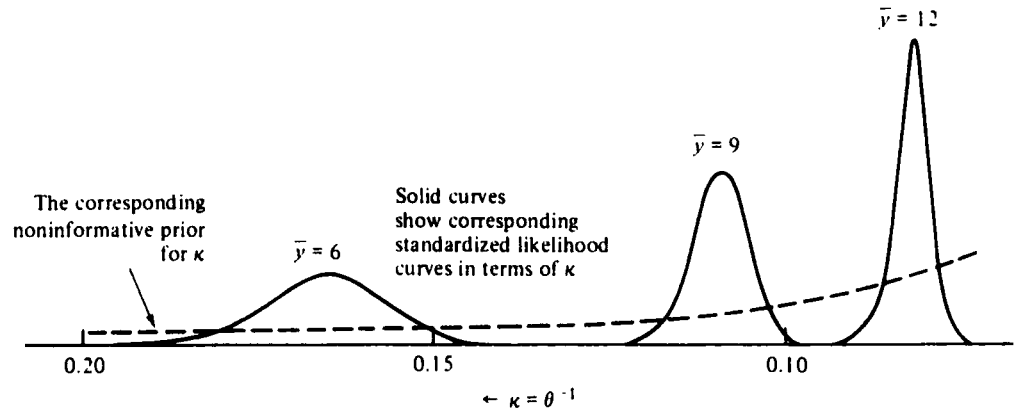
Data Translated Likelihood and Non-informative Prior

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.

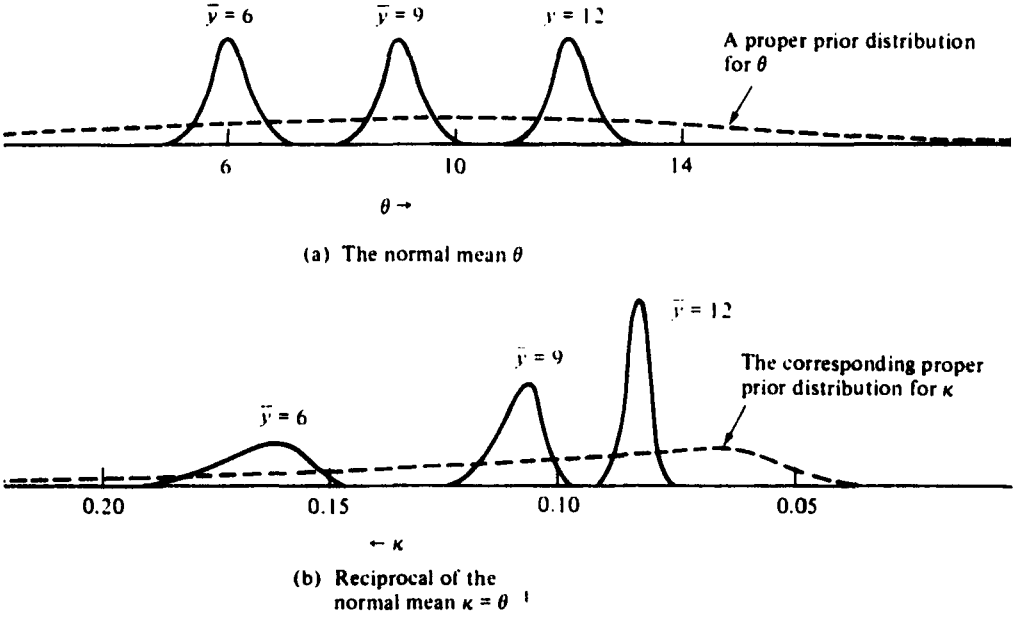


Fig. 1.3.2 Noninformative prior distributions and standardized likelihood curves: (a) for the Normal mean θ, and (b) for $\kappa = \theta^{-1}$ seen over a wider range of parameter values.

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

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

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.

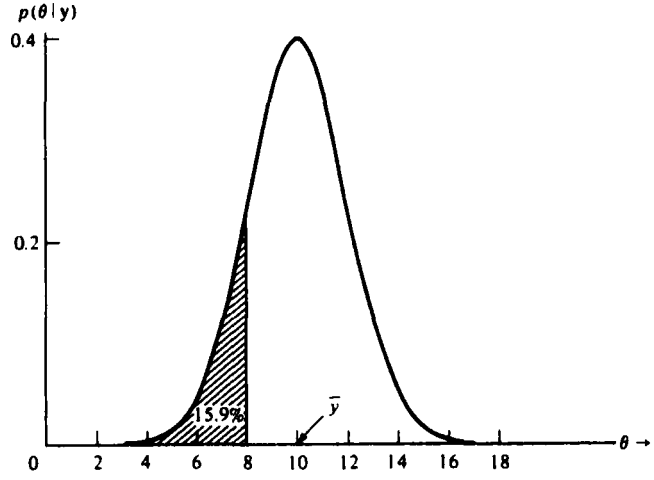


Fig. 1.3.3 Posterior distribution of the Normal mean θ (noninformative prior), when $\bar{y} = 10$, $\sigma = 8$, $n = 16$.

1.3.2 The Normal Standard Deviation σ (θ known)

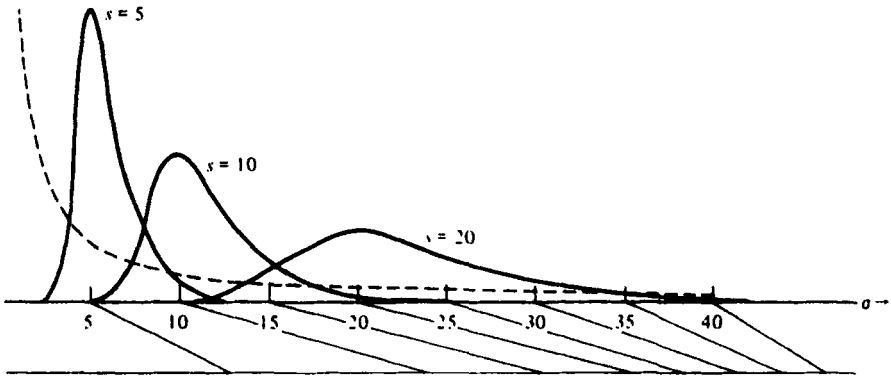
As a second example, consider the choice of a noninformative prior distribution for σ, the standard deviation of a Normal distribution for which the mean θ is supposed known. In this case, the likelihood is

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

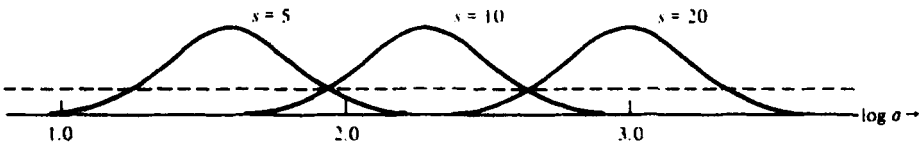
where

$$s^2 = \Sigma (y_u - 0)^2/n.$$

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 $\log \sigma$ 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 $\log \sigma$ as

$$l(\log \sigma \mid 0, \mathbf{y}) \propto \exp\left\{-n(\log \sigma - \log s) - \frac{n}{2}\exp[-2(\log \sigma - \log s)]\right\}. \qquad (1.3.7)$$

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

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

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 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

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

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,

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

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

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

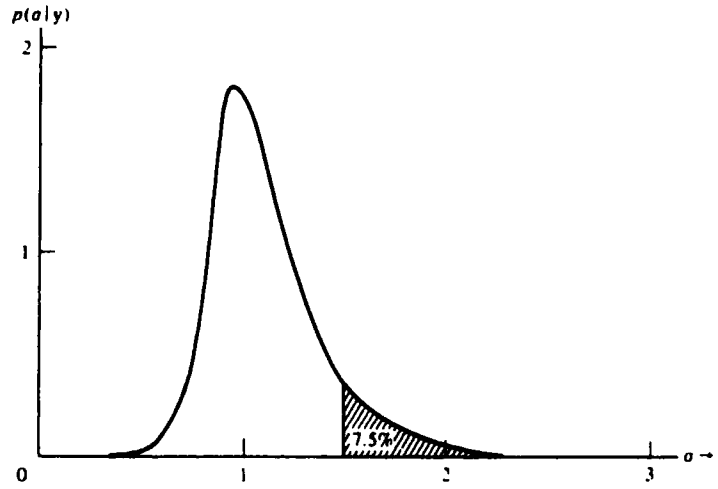


Fig. 1.3.5 Posterior distribution of the Normal standard deviation σ (noninformative prior), when $s = 1$ and $n = 10$.

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

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

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

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

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

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

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 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

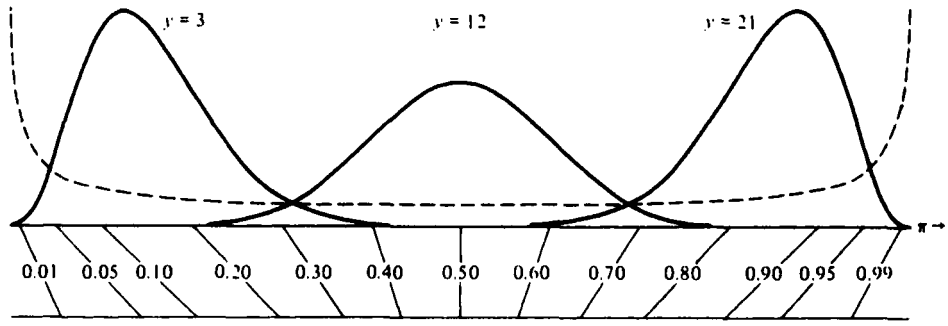
$$p(y \mid \pi) = \frac{n!}{y!\,(n-y)!}\,\pi^y(1-\pi)^{n-y}, \qquad y = 0, \ldots, n, \tag{1.3.21}$$

so that the likelihood is

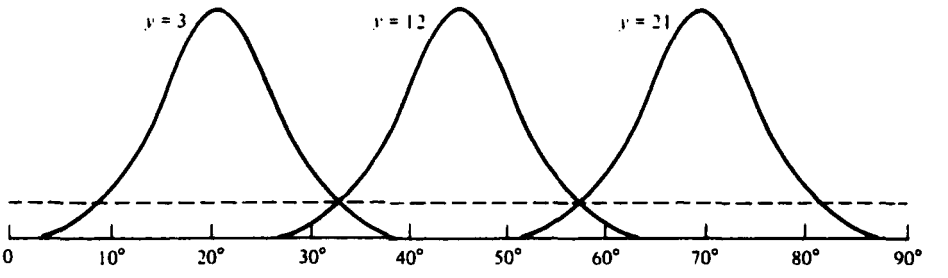
$$l(\pi \mid y) \propto \pi^y(1-\pi)^{n-y}. \tag{1.3.22}$$

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

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

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),

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

After substitution of the appropriate normalizing constant, we find that the

corresponding posterior distribution for π is the beta distribution

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

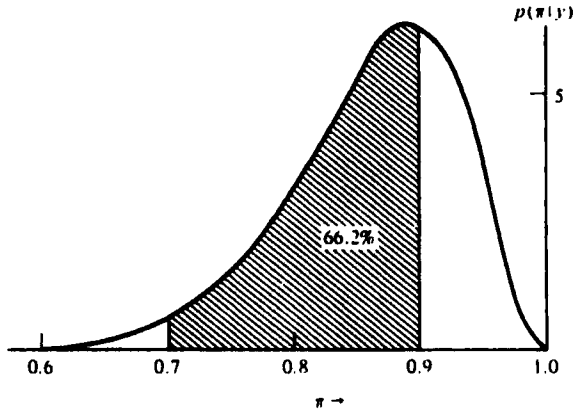


Fig. 1.3.7 Posterior distribution of the binomial mean π (noninformative prior) for 21 successes out of 24 trials.

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

$$L(\theta \mid \mathbf{y}) = \log l(\theta \mid \mathbf{y}) = \log \prod_{u=1}^{n} p(y_u \mid \theta)$$

$$\doteq L(\hat{\theta} \mid \mathbf{y}) - \frac{n}{2}(\theta - \hat{\theta})^2 \left(-\frac{1}{n} \frac{\partial^2 L}{\partial \theta^2} \right)_{\hat{\theta}}. \tag{1.3.27}$$

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

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

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

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

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,

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

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

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

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

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

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

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

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,

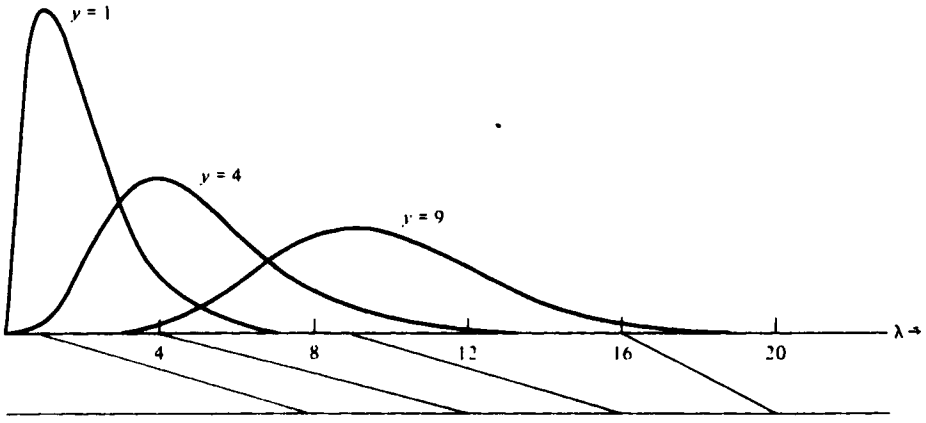
$$p(y \mid \lambda) = \frac{\lambda^y}{y!} \exp(-\lambda), \qquad y = 0, \ldots, \infty, \tag{1.3.39}$$

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

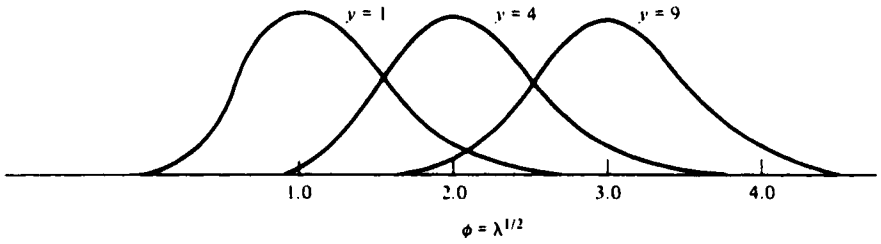
$$l(\lambda \mid \mathbf{y}) \propto \lambda^{n\bar{y}} \exp(-n\lambda), \qquad \bar{y} = \frac{1}{n} \Sigma y_u. \tag{1.3.40}$$

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

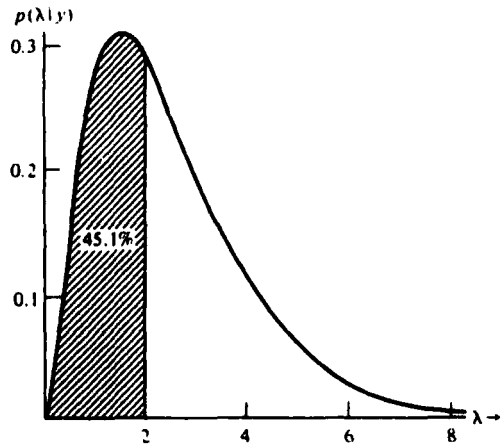


Fig. 1.3.9 Posterior distribution of the Poisson mean λ (noninformative prior) for an observed frequency $y = 2$.

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 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 y are drawn from the distribution $p(y \mid \theta_0)$. 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,

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

where $\mathcal{I}(\theta) = a(\theta, \theta)$ is the function

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

Consequently, if we use $\mathcal{I}(\hat{\theta})$, 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

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

Equivalently, the noninformative prior for θ should be chosen so that, locally,

$$p(\theta) \propto \mathcal{I}^{1/2}(\theta). \tag{1.3.49}$$

It is readily confirmed that, when the distribution $p(y \mid \theta)$ is of the form (1.3.38), $J(\hat{\theta}) \equiv \mathcal{I}(\hat{\theta})$. Thus, the prior in (1.3.33), when applicable, is in fact identical to the prior of (1.3.49) and the latter form can be used generally.

For illustration, consider again the binomial mean π. The likelihood is equivalent to the likelihood from a sample of n independent point binomials identically distributed as

$$p(x \mid \pi) = \pi^x (1 - \pi)^{1-x}, \qquad x = 0, 1. \tag{1.3.50}$$

Thus

$$\frac{\partial^2 \log p}{\partial \pi^2} = -\frac{x}{\pi^2} - \frac{1 - x}{(1 - \pi)^2}. \tag{1.3.51}$$

Since $\underset{x|\pi}{E}(x) = \pi$, it follows that

$$\mathcal{I}(\pi) = \underset{x|\pi}{E}\left[-\frac{\partial^2 \log p}{\partial \pi^2}\right] = \pi^{-1}(1 - \pi)^{-1}, \tag{1.3.52}$$

whence, according to (1.3.49), the noninformative prior of π is locally proportional to $\pi^{-1/2}(1 - \pi)^{-1/2}$ as obtained earlier in (1.3.37a). Also, we see from (1.3.36) and (1.3.52) that $J(\hat{\pi})$ and $\mathcal{I}(\hat{\pi})$ are identical.

Now, the quantity $\mathcal{I}(\theta)$ obtained in (1.3.47) above is Fisher's measure of information about θ in a single observation y. More generally, Fisher's measure (1922, 1925) of information about θ in a sample $\mathbf{y}' = (y_1, \ldots, y_n)$ is defined as

$$\mathcal{I}_n(\theta) = \underset{\mathbf{y}|\theta}{E}\left(-\frac{\partial^2 L}{\partial \theta^2}\right) = \underset{\mathbf{y}|\theta}{E}\left(\frac{\partial L}{\partial \theta}\right)^2. \tag{1.3.53}$$

where, as before, L is the log likelihood and the expectation is taken with respect to the distribution $p(\mathbf{y} \mid \theta)$. When $\mathbf{y}$ is a random sample, $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$. Thus, (1.3.49) can be expressed by the following rule.

Jeffreys' rule: The prior distribution for a single parameter θ is approximately noninformative if it is taken proportional to the square root of Fisher's information measure.

This rule for the choice of a noninformative prior distribution was first given by Sir Harold Jeffreys (1961), who justified it on the grounds of its invariance under parameter transformations. For, suppose $\phi = \phi(\theta)$ is a one-to-one transformation of θ; then it is readily seen that

$$\mathcal{I}(\phi) = \mathcal{I}(\theta)\left(\frac{d\theta}{d\phi}\right)^2. \tag{1.3.54}$$

Now, if some principle of choice led to $p(\theta)$ as a noninformative prior for θ, the same principle should lead to

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| \tag{1.3.55}$$

as a noninformative prior for ϕ. The principle of choice in (1.3.49) precisely satisfies this requirement: if we use it, then the prior of ϕ is

$$p(\phi) \propto \mathcal{I}^{1/2}(\phi) = \mathcal{I}^{1/2}(\theta)\left|\frac{d\theta}{d\phi}\right| \propto p(\theta)\left|\frac{d\theta}{d\phi}\right|. \tag{1.3.56}$$

The Location Scale Family

For illustration we show how results in (1.3.3) and (1.3.8) concerning the parameters (θ, σ) for the Normal distribution may, as an approximation, be extended to cover the general location-scale family of distributions

$$p(y \mid \theta, \sigma) \propto \sigma^{-1} h\left(\frac{y-\theta}{\sigma}\right). \tag{1.3.57}$$

where the range of y does not involve (θ, σ) and h satisfies certain regularity conditions. Suppose we have a sample of n independent observations $\mathbf{y}' = (y_1, \ldots, y_n)$ from this distribution.

a) *θ unknown, σ known.* The likelihood is

$$l(\theta \mid \sigma, \mathbf{y}) \propto \prod_{u=1}^{n} h\left(\frac{y_u - \theta}{\sigma}\right). \tag{1.3.58}$$

Now

$$\frac{\partial \log p(y \mid \theta, \sigma)}{\partial \theta} = -\frac{1}{\sigma}\frac{h'(x)}{h(x)}, \quad \text{where} \quad x = \left(\frac{y-\theta}{\sigma}\right). \tag{1.3.59}$$

Thus, from (1.3.47)

$$\mathcal{I}(\theta) = \underset{y|\theta}{E}\left[\frac{\partial \log p(y \mid \theta, \sigma)}{\partial \theta}\right]^2 = \frac{1}{\sigma^2}\underset{x}{E}\left[\frac{h'(x)}{h(x)}\right]^2, \tag{1.3.60}$$

where the expectation on the extreme right is taken over the distribution $p(x) = h(x)$. Since this expectation does not involve θ and σ^2 is known, $\mathcal{I}(\theta) =$ constant. So we are led to take θ locally uniform *a priori*,

$$p(\theta \mid \sigma) \propto \mathcal{I}^{1/2}(\theta) = \text{constant}, \tag{1.3.61}$$

and the corresponding posterior distribution of θ is approximately

$$p(\theta \mid \sigma, \mathbf{y}) \propto \prod_{u=1}^{n} h\left(\frac{y_u - \theta}{\sigma}\right), \qquad -\infty < \theta < \infty. \tag{1.3.62}$$

b) σ *unknown*, θ *known*. Here the likelihood is

$$l(\sigma \mid \theta, \mathbf{y}) \propto \sigma^{-n} \prod_{u=1}^{n} h\left(\frac{y_u - \theta}{\sigma}\right). \tag{1.3.63}$$

Since

$$\frac{\partial \log p(y \mid \theta, \sigma)}{\partial \sigma} = -\frac{1}{\sigma}\left[1 + \frac{xh'(x)}{h(x)}\right],$$

it follows that

$$\mathcal{I}(\sigma) = \frac{1}{\sigma^2} E_x\left[1 + \frac{xh'(x)}{h(x)}\right]^2 \propto \frac{1}{\sigma^2}. \tag{1.3.64}$$

Consequently, Jeffreys' rule leads to taking the prior

$$p(\sigma \mid \theta) \propto \frac{1}{\sigma} \qquad \text{or} \qquad p(\log \sigma) \propto \text{const.} \tag{1.3.65}$$

The corresponding posterior distribution of σ is then

$$p(\sigma \mid \theta, \mathbf{y}) \propto \sigma^{-(n+1)} \prod_{u=1}^{n} h\left(\frac{y_u - \theta}{\sigma}\right), \qquad \sigma > 0. \tag{1.3.66}$$

Caution in the Application of Jeffreys' Rule

Jeffreys' rule given by (1.3.49), like most rules, should not be mechanically applied. It is obviously inapplicable for example when $\mathcal{I}(\theta)$ does not exist. Furthermore, as Jeffreys pointed out, the rule has to be modified in certain circumstances which we will discuss later in Section 1.3.6. We believe that it is best, to treat as basic the idea of seeking a transformation for which the likelihood is approximately data translated, to treat each problem individually, and to regard equation (1.3.49) as a means whereby *in appropriate circumstances* this transformation can be determined.

Dependence of the Noninformative Prior Distribution on the Probability Model

In the development followed above the form of the noninformative prior distribution depends upon the probability model of the observations. It might be argued that this dependence is objectionable because the representation of *total* ignorance about a parameter ought to be the same, whatever the nature of a projected experiment. On the contrary, the position adopted above is that we seek to represent not total ignorance but an amount of prior information which is small *relative* to what the particular projected experiment can be expected to provide. The form of the prior *must* then depend on the expected likelihood.

As a specific example, suppose for example π is the proportion of successes in a Bernoulli population. Now π can be estimated (1) by counting the number of successes

y in a *fixed number of trials* n and using the fact that y has the binomial distribution in (1.3.21), or (2) by counting the number of trials z until a *fixed number of successes* r is obtained and supposing that z has the Pascal distribution

$$\binom{z-1}{r-1}\pi^r(1-\pi)^{z-r}, \qquad z = r, r+1, \ldots . \tag{1.3.67a}$$

These two experiments lead, on the argument given above, to slightly different noninformative priors. Specifically, for the binomial case, the information measure is, from (1.3.52),

$$E\left(-\frac{\partial^2 L}{\partial \pi^2}\right) = n\pi^{-1}(1-\pi)^{-1},$$

whence

$$p(\pi) \propto \pi^{-\frac{1}{2}}(1-\pi)^{-\frac{1}{2}}$$

and the corresponding noninformative prior is locally uniform in $\phi = \sin^{-1}\sqrt{\pi}$ as in (1.3.37a, b). On the other hand, it is easily shown that the information measure for the Pascal experiment is

$$E\left(-\frac{\partial^2 L}{\partial \pi^2}\right) = r\pi^{-2}(1-\pi)^{-1}, \tag{1.3.67b}$$

whence

$$p(\pi) \propto \pi^{-1}(1-\pi)^{-\frac{1}{2}} \tag{1.3.67c}$$

and the corresponding noninformative prior is locally uniform in

$$\phi = \log\frac{1-\sqrt{1-\pi}}{1+\sqrt{1-\pi}}.$$

Now, these two kinds of experiments would lead to exactly the same likelihood when $n = z$ and $y = r$; and when this is so it has been argued that inference about π from both experiments ought to be identical. The use of the above two noninformative priors however will not yield this result. For illustration let us suppose there were 24 trials with 21 successes. If to arrive at this result sampling had been continued till the number of *trials* was 24 the posterior distribution obtained with the appropriate noninformative prior would have been

$$p(\pi \mid y = 21) \propto \pi^{20.5}(1-\pi)^{2.5}.$$

However if sampling had been continued till the number of *successes* was 21 then the posterior distribution obtained with the appropriate noninformative prior in (1.3.67c) would have been

$$p(\pi \mid z = 24) \propto \pi^{20}(1-\pi)^{2.5}.$$

This says that when we sample till the number of successes reaches a certain value some downward adjustment of probability is needed relative to sampling with fixed n. We find this result much less surprising than the claim that they ought to agree.

In general we feel that it is sensible to choose a noninformative prior which expresses ignorance *relative* to information which can be supplied by a particular experiment. If the experiment is changed, then the expression of relative ignorance can be expected to change correspondingly.

1.3.6 Noninformative Priors for Multiple Parameters

We now extend previous ideas to include multiparameter models. We begin by considering the Normal linear model with σ assumed known.

The Parameters θ in a Normal Linear Model, σ Assumed Known

Suppose $\mathbf{y}' = (y_1, \ldots, y_n)$ is a set of Normally and independently distributed random variables having common variance σ^2, and the expected value of y_u is a linear function of k parameters $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_k)$ such that

$$E(y_u) = \theta_1 x_{u1} + \theta_2 x_{u2} + \cdots + \theta_k x_{uk}, \qquad u = 1, 2, \ldots, n, \tag{1.3.68}$$

where the x's are known constants.

This Normal linear model is of basic importance. In particular, for suitable choice of the x's, it provides the structure for general Analysis of Variance and for Regression (Least Squares) Analysis. Special cases include models already considered. For example, the model (1.1.1) for a Normal sample is obtained by setting $k = 1, \theta_1 = \theta$, and $x_{u1} = 1$ $(u = 1, 2, \ldots, n)$.

In general, if $\mathbf{X}$ is the $n \times k$ matrix $\{x_{uj}\}$ of known constants, then the n equations may be written concisely as

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\theta}.$$

If σ is known, then the likelihood is

$$l(\boldsymbol{\theta} \mid \sigma, \mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\right]. \tag{1.3.69}$$

We shall suppose that the rank of the matrix $\mathbf{X}$ is k. The quadratic form in (1.3.69) may then be written

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \tag{1.3.70}$$

where

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

is the vector of least squares estimate of $\boldsymbol{\theta}$, and $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$ is the vector of fitted values so that $(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$ is a function of data not involving $\boldsymbol{\theta}$. The

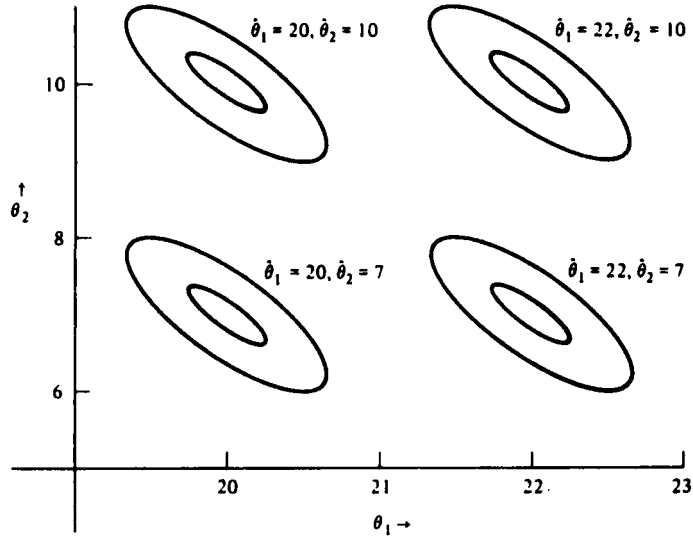


Fig. 1.3.10 Normal linear model: contours of likelihood function for different sets of data.

likelihood can therefore be expressed as

$$l(\boldsymbol{\theta} \mid \sigma, \mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right], \tag{1.3.71}$$

which is in the form of a multivariate Normal function† centered at $\boldsymbol{\theta}$ and having covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The likelihood contours in the parameter space of $\boldsymbol{\theta}$ are thus ellipses ($k = 2$), ellipsoids ($k = 3$), or hyperellipsoids ($k > 3$) defined by

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = \text{const.} \tag{1.3.72}$$

Figure 1.3.10 illustrates the case $k = 2$, where likelihood contours are shown for different sets of data yielding different values of $\hat{\theta}_1$ and $\hat{\theta}_2$. The likelihood is data translated. Specifically, from (1.3.71) it is seen that, as soon as an experimental

† We refer to the function

$$f(\mathbf{x}) = \frac{|\mathbf{V}|^{-1/2}}{(2\pi)^{p/2}} \exp\left[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\,\mathbf{V}^{-1}\,(\mathbf{x} - \boldsymbol{\mu})\right],$$

where $\mathbf{x}' = (x_1, \ldots, x_p)$ and $\boldsymbol{\mu}' = (\mu_1, \ldots, \mu_p)$ are $p \times 1$ vectors, and $\mathbf{V}$ is a $p \times p$ positive definite symmetric matrix, as the multivariate Normal function. When $\mathbf{x}$ are random variables, the function becomes the multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$.

design has been decided and hence X is known, all features of the likelihood except its location are known prior to the data being taken. By contrast, location of the likelihood is determined through $\hat{\theta}$ *solely* by the data. The idea that little is known *a priori* relative to what the data will tell us is therefore expressed by a prior distribution such that, locally,

$$p(\theta \mid \sigma) \propto c. \tag{1.3.73}$$

On multiplying the likelihood in (1.3.71) by a nearly constant prior, and introducing the appropriate normalizing constant so that the posterior distribution integrates to one, we have approximately

$$p(\theta \mid \sigma, \mathbf{y}) \doteq \frac{|\mathbf{X}'\mathbf{X}|^{1/2}}{(2\pi\sigma^2)^{k/2}} \exp\left[-\frac{1}{2\sigma^2}(\theta - \hat{\theta})' \mathbf{X}'\mathbf{X}(\theta - \hat{\theta})\right],$$

$$-\infty < \theta_i < \infty, \quad i = 1, \ldots, k, \tag{1.3.74}$$

a multivariate Normal distribution which we denote by $N_k[\hat{\theta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$. This distribution will be discussed in detail in Section 2.7.

For certain common situations discussed more fully in Chapter 7 this formulation of the multivariate prior distribution may be inappropriate. For example, it may be known that the θ's themselves are a sample from some distribution. When such information is available and particularly when k is large, the locally uniform prior of (1.3.73) may supply an inadequate approximation.

Multiparameter Data Translated Likelihoods

In general, suppose the distribution of the data y involves k parameters $\theta' = (\theta_1, \ldots, \theta_k)$. A data translated likelihood must be of the form

$$l(\theta \mid \mathbf{y}) = g[\phi - \mathbf{f}(\mathbf{y})], \tag{1.3.75}$$

where $g(\mathbf{x})$ is a known function independent of y, $\phi' = (\phi_1, \ldots, \phi_k)$ is a one-to-one transformation of θ, and the elements $f_1(\mathbf{y}), \ldots, f_k(\mathbf{y})$ of the $k \times 1$ vector $\mathbf{f}(\mathbf{y})$ are k functions of y. Extending the single parameter case, a noninformative prior is supposed locally uniform in ϕ. The corresponding noninformative prior in θ is then

$$p(\theta) \propto |J|, \tag{1.3.76}$$

where

$$|J| = \left|\frac{\partial(\phi_1, \ldots, \phi_k)}{\partial(\theta_1, \ldots, \theta_k)}\right|_+ = \begin{vmatrix} \frac{\partial \phi_1}{\partial \theta_1} & \cdots & \frac{\partial \phi_1}{\partial \theta_k} \\ \vdots & & \vdots \\ \frac{\partial \phi_k}{\partial \theta_1} & \cdots & \frac{\partial \phi_k}{\partial \theta_k} \end{vmatrix}_+$$

is the absolute value of the Jacobian of the transformation. For the Normal linear model (1.3.68), $\boldsymbol{\theta} = \boldsymbol{\phi}$ so that a noninformative prior is locally uniform in $\boldsymbol{\theta}$ as given earlier in (1.3.73).

Multiparameter Problems Involving Location and Scale Parameters

Special care must be exercised in choosing noninformative priors for location and scale parameters simultaneously. As mentioned earlier, we regard a parameter η as a *location* parameter of a distribution $p(y)$ if addition of a constant c to y changes η to $\eta + c$. Thus, the Normal mean and, more generally, the parameter θ in the family of distributions (1.3.57) are location parameters. The elements of $\boldsymbol{\theta}$ in the Normal linear model (1.3.68) are also location parameters in a general sense, since it can be shown that they are location parameters of the distribution of the least squares estimates $\hat{\boldsymbol{\theta}}$. On the other hand, a *scale* parameter λ of a distribution $p(y)$ is such that multiplication of y by a constant c changes λ to $|c|\lambda$. Examples of scale parameters are the Normal standard deviation and, more generally, the parameter σ in (1.3.57) and in the linear model (1.3.68).

Normal Mean θ and Standard Deviation σ

We first consider the choice of prior in relation to a sample from a Normal distribution $N(\theta, \sigma^2)$, where θ and σ are both unknown. The likelihood of (θ, σ) is

$$l(\theta, \sigma \mid \mathbf{y}) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\Sigma(y_u - \theta)^2\right]. \tag{1.3.77a}$$

Now

$$\begin{aligned}\Sigma(y_u - \theta)^2 &= \Sigma(y_u - \bar{y})^2 + n(\bar{y} - \theta)^2 \\ &= (n-1)s^2 + n(\bar{y} - \theta)^2,\end{aligned}$$

where $s^2 = \Sigma(y_u - \bar{y})^2/(n-1)$. Thus,

$$l(\theta, \sigma \mid \mathbf{y}) \propto \sigma^{-n} \exp\left[-\frac{n(\theta - \bar{y})^2}{2\sigma^2} - \frac{(n-1)s^2}{2\sigma^2}\right]. \tag{1.3.77b}$$

Also, since multiplication by the constant s^n leaves the likelihood unchanged,

$$l(\theta, \sigma \mid \mathbf{y}) \propto \left(\frac{s}{\sigma}\right)^n \exp\left[-\frac{n(\theta - \bar{y})^2}{2s^2}\left(\frac{s^2}{\sigma^2}\right) - \frac{(n-1)s^2}{2\sigma^2}\right], \tag{1.3.77c}$$

which can be written,

$$\begin{aligned}l(\theta, \sigma \mid \mathbf{y}) \propto \exp&\left\{-\frac{n}{2}\left(\frac{\theta - \bar{y}}{s}\right)^2 \exp\left[-2(\log\sigma - \log s)\right]\right\} \\ &\times \exp\left\{-n(\log\sigma - \log s) - \left(\frac{n-1}{2}\right)\exp\left[-2(\log\sigma - \log s)\right]\right\}\end{aligned} \tag{1.3.77d}$$

and is therefore of the form

$$l(\theta, \sigma \mid \mathbf{y}) \propto F\left(\frac{\theta - \bar{y}}{s}, \log \sigma - \log s\right) G(\log \sigma - \log s). \tag{1.3.77e}$$

To aid understanding of this expression, Fig. 1.3.11 shows likelihood contours for θ and $\log \sigma$ given by four different samples of $n = 10$ observations. For fixed s a change in $\bar{y}$ relocates the likelihood surface on the θ axis. For fixed $\bar{y}$ a magnification in s appropriately relocates the likelihood surface on the $\log \sigma$ axis, while at the same time its spread along the θ axis is correspondingly magnified. This magnification reflects the greater uncertainty in the likelihood about θ which occurs when a larger σ is implied by an increase in s. It would usually be the case, however, that prior opinions about θ bear little relationship to those about σ, so that such magnifications would be irrelevant to the choice of transformations of θ. Thus, we are led to seek a transformation which, *apart from this inherent magnification along the θ axis*, is such that the data serves only to relocate the likelihood function. In this case the appropriate transformation is clearly obtained in terms of θ and $\log \sigma$. A noninformative prior is therefore taken to be one for which approximately $\log \sigma$ and θ are locally uniform.

$$p(\theta, \log \sigma) \propto c. \tag{1.3.78}$$

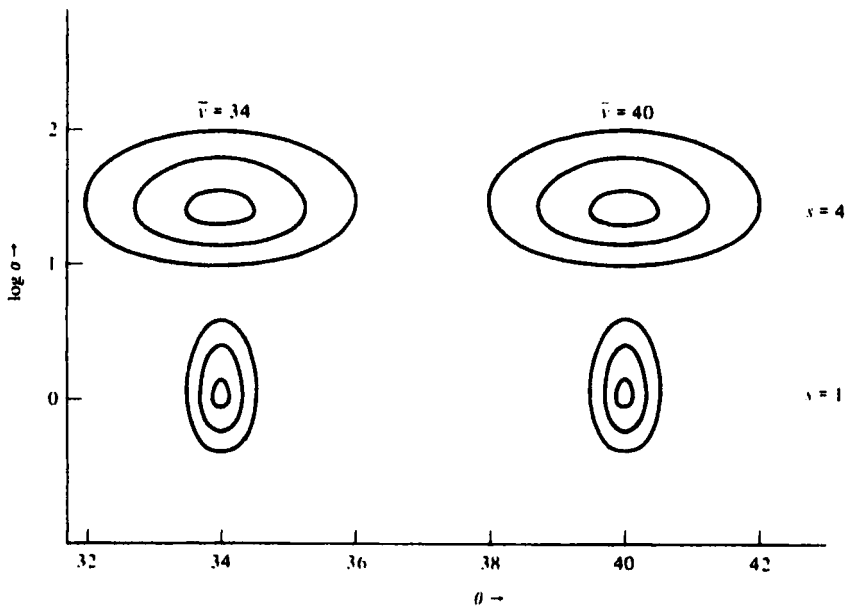


Fig. 1.3.11 The Normal mean θ and standard deviation σ: contours of likelihood function of $(\theta, \log \sigma)$ for different sets of data.

Equivalently

$$p(\theta, \sigma) \propto \sigma^{-1}. \tag{1.3.79}$$

Employing the prior in (1.3.79) with the likelihood (1.3.77b), we found that the posterior distribution of (θ, σ) is

$$p(\theta, \sigma \mid \mathbf{y}) \propto \sigma^{-(n+1)} \exp\left[-\frac{n(\theta - \bar{y})^2}{2\sigma^2} - \frac{(n-1)s^2}{2\sigma^2} \right], \quad -\infty < \theta < \infty, \quad \sigma > 0, \tag{1.3.80}$$

which will be discussed in detail in Section 2.4.

Normal Linear Model, σ Unknown

We now turn to the case of the general Normal linear model in (1.3.68), and suppose that the standard deviation σ as well as the parameters $\boldsymbol{\theta}$ are unknown. The likelihood can be written

$$l(\boldsymbol{\theta}, \sigma \mid \mathbf{y}) \propto \left(\frac{s}{\sigma}\right)^n \exp\left[-\frac{(n-k)s^2}{2\sigma^2} - \frac{s^2}{2\sigma^2} \frac{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \mathbf{X}'\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})}{s^2} \right], \tag{1.3.81}$$

where

$$s^2 = \frac{1}{(n-k)} (\mathbf{y} - \hat{\mathbf{y}})' (\mathbf{y} - \hat{\mathbf{y}}).$$

By comparing (1.3.81) with (1.3.77b–e), it is evident that in terms of $\boldsymbol{\theta}$ and $\log \sigma$, for any fixed s, a change in $\hat{\boldsymbol{\theta}}$ merely relocates the likelihood. On the other hand, for fixed $\hat{\boldsymbol{\theta}}$, the volume enclosed in a given contour is proportional to s^k. Arguing as in the case of (θ, σ), we seek a transformation which, apart from this inherent magnification in the space of $\boldsymbol{\theta}$, is such that the data serves only to relocate the likelihood surface. Clearly, the appropriate transformation is obtained in terms of $\boldsymbol{\theta}$ and $\log \sigma$. A noninformative prior in this case is then one for which approximately $\log \sigma$ and $(\theta_1, \ldots, \theta_k)$ are locally uniform. Specifically, it is assumed that, locally,

$$p(\boldsymbol{\theta}, \log \sigma) \propto c \qquad \text{or} \qquad p(\boldsymbol{\theta}, \sigma) \propto 1/\sigma. \tag{1.3.82}$$

The corresponding posterior distribution is then

$$p(\boldsymbol{\theta}, \sigma \mid \mathbf{y}) \propto \sigma^{-(n+1)} \exp\left[-\frac{(n-k)s^2}{2\sigma^2} - \frac{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \mathbf{X}'\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})}{2\sigma^2} \right]$$

$$-\infty < \theta_i < \infty, \quad i = 1, \ldots, k; \quad \sigma > 0, \tag{1.3.83}$$

which will be further discussed in Section 2.7.

Prior Independence Between Parameters

In some examples, certain parameters or sets of parameters may be judged *a priori* to be distributed independently of certain other parameters or sets of

parameters. When this is so, the choice of prior distribution is sometimes simplified because the independent sets of parameters may be separately considered.

In particular, it is usually appropriate to take location parameters to be distributed independently of scale parameters. This is because any prior idea one might have about the location θ of a distribution would usually not be much influenced by one's idea about the value of its scale parameter σ. Thus $p(\theta \mid \sigma) \doteq p(\theta)$. Consider the case of the Normal distribution. We have seen that for the case where σ is known, a noninformative prior for θ is obtained by taking $p(\theta \mid \sigma)$ locally uniform. With the additional independence assumption this implies that $p(\theta)$ should be uniform. A similar argument leads to our taking $p(\sigma) \propto 1/\sigma$. Thus

$$p(\theta, \sigma) \doteq p(\theta)p(\sigma) \propto 1/\sigma \tag{1.3.84}$$

as in (1.3.79).

In certain circumstances, such an assumption of independence between θ and σ could appear inappropriate. If, for example, we know that grains of sand, with mean weight one milligram, were to be weighed, we should expect that σ would be less than a milligram, and so it has been argued that if θ is small, then σ is likely to be small also, while if θ is large, σ is likely to be large.

To this we reply, (1) that dependence of this kind is most often associated with a natural constraint on the data (for example, that all observations must be non-negative); (2) that this kind of dependence is usually removed when a more appropriate metric is adopted in terms of which the data are not so constrained; (3) that usually when there is no such constraint one does not expect this dependence. We illustrate with the following examples.

An example of prior dependence removed by appropriate transformation. Suppose we knew that the mean income θ of a certain community was \$5000 and the standard deviation was \$1000. Then given another community where the mean income was known to be \$50,000, we might guess that the standard deviation was closer to \$10,000 than to \$1000. In other words, prior beliefs about θ and σ would be dependent. However, this guess is clearly based on the supposition that, in examples of this kind, it is the coefficient of variation σ/θ and not σ itself which is more likely to be approximately constant over different values of θ. But if this supposition is correct, then log income is the appropriate quantity to consider. For, if we denote income by y and suppose that ϕ and λ are the mean and standard deviation of $\log y$, then

$$\phi = E(\log y) \doteq \log \theta, \qquad \lambda = \sqrt{\operatorname{var}(\log y)} \doteq \sigma/\theta.$$

Measured in the logarithmic metric, the standard deviation λ can realistically be assumed independent of the mean ϕ *a priori*. We notice that, in this example, if y is an observed income, $0 < y < \infty$ but $-\infty < \log y < \infty$.

An example where measurements may be negative. Except in examples like the above, where the measurement scale has a natural constraint such as a truncation at zero, values of θ which are small in magnitude need not be associated with small values of σ. For example, suppose we were checking the declination of a compass from magnetic north, using an instrument which could detect a declination from $-180°$ to $180°$. For a properly constructed compass, we would expect the declination θ to be close to zero but this would not ordinarily affect our ideas about σ.

In this book, we shall usually assume that location and scale parameters are approximately independent *a priori*. In particular, arguing as before, for the parameters (θ, σ) in the linear Normal model, we suppose that $p(\theta, \sigma) \doteq p(\theta)p(\sigma)$ so that $p(\theta) \doteq p(\theta \mid \sigma)$ and $p(\sigma) \doteq p(\sigma \mid \theta)$. It then follows that $p(\theta, \sigma) \propto \sigma^{-1}$ as given in (1.3.82).

Extension of Jeffreys' Rule to Multiparameter Models

In the multiparameter examples discussed above, transformations were available which had the property that, apart from the inherent magnifying effect of the scale factor—in the location parameter space, the likelihood was data translated. Although transformations of this kind are available for many of the applications discussed in later chapters, they are not available in general. In some cases, then, to obtain noninformative prior distributions, we must rely on a somewhat less satisfactory argument leading to the multiparameter version of Jeffreys' rule.

If the distribution of $\mathbf{y}$, depending on k parameters $\boldsymbol{\theta}$, obeys certain regularity conditions, then, for sufficiently large samples, the likelihood function for $\boldsymbol{\theta}$ and for mild transformations of $\boldsymbol{\theta}$ approaches a multivariate Normal distribution. The log likelihood is thus approximately quadratic,

$$L(\boldsymbol{\theta} \mid \mathbf{y}) = \log l(\boldsymbol{\theta} \mid \mathbf{y}) \doteq L(\hat{\boldsymbol{\theta}} \mid \mathbf{y}) - \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \mathbf{D}_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \tag{1.3.85}$$

where $\hat{\boldsymbol{\theta}}$ is the vector of maximum likelihood estimates of $\boldsymbol{\theta}$ and $-n\mathbf{D}_{\hat{\boldsymbol{\theta}}}$ is the $k \times k$ matrix of second derivatives evaluated at $\hat{\boldsymbol{\theta}}$, that is,

$$\mathbf{D}_{\hat{\boldsymbol{\theta}}} = \left\{ -\frac{1}{n} \frac{\partial^2 L}{\partial \theta_i \partial \theta_j} \right\}_{\hat{\boldsymbol{\theta}}}, \qquad i, j = 1, \ldots, k.$$

In general, $\mathbf{D}_{\hat{\boldsymbol{\theta}}}$ will depend upon all the data $\mathbf{y}$. But for large n, it can be closely approximated by

$$\mathbf{D}_{\hat{\boldsymbol{\theta}}} \doteq \frac{1}{n} \mathscr{I}(\hat{\boldsymbol{\theta}}), \tag{1.3.86}$$

which is a function of $\hat{\boldsymbol{\theta}}$ only. Specifically, $\mathscr{I}_n(\boldsymbol{\theta})$ is the matrix function

$$\mathscr{I}_n(\boldsymbol{\theta}) = E\left\{ -\frac{\partial^2 L}{\partial \theta_i \partial \theta_j} \right\}, \tag{1.3.87}$$

where the expectation is taken with respect to the data distribution $p(\mathbf{y} \mid \boldsymbol{\theta})$. In other words, $\mathcal{I}_n(\boldsymbol{\theta})$ is the information matrix associated with the sample y.

Now, ideally one would seek a transformation $\boldsymbol{\phi}(\boldsymbol{\theta})$ such that $\mathcal{I}_n(\boldsymbol{\phi})$ would be a constant matrix independent of $\hat{\boldsymbol{\phi}}$ so that the likelihood would be approximately data translated. Because this is not possible in general, we may seek a transformation $\boldsymbol{\phi}$ which ensures that the *content* of the approximate likelihood region of $\boldsymbol{\phi}$,

$$(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})' \mathcal{I}_n(\hat{\boldsymbol{\phi}})(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) < \text{const.}, \tag{1.3.88}$$

remains constant for different $\hat{\boldsymbol{\phi}}$. Since the square root of the determinant, $|\mathcal{I}_n(\hat{\boldsymbol{\phi}})|^{1/2}$, measures the volume of the likelihood region, the above requirement is equivalent to asking for a transformation for which the $|\mathcal{I}_n(\hat{\boldsymbol{\phi}})|$ is independent of $\hat{\boldsymbol{\phi}}$. To find such a transformation, note that

$$\mathcal{I}_n(\boldsymbol{\phi}) = \mathbf{A} \mathcal{I}_n(\boldsymbol{\theta}) \mathbf{A}', \tag{1.3.89}$$

where $\mathbf{A}$ is the $k \times k$ matrix of partial derivatives,

$$\mathbf{A} = \left[\frac{\partial(\theta_1, \ldots, \theta_k)}{\partial(\phi_1, \ldots, \phi_k)} \right].$$

Thus,

$$|\mathcal{I}_n(\boldsymbol{\phi})| = |\mathbf{A}|^2 |\mathcal{I}_n(\boldsymbol{\theta})|, \tag{1.3.90}$$

whence the above requirement will be satisfied if $\boldsymbol{\phi}$ is such that

$$|\mathbf{A}| = \left| \frac{\partial(\theta_1, \ldots, \theta_k)}{\partial(\phi_1, \ldots, \phi_k)} \right| \propto |\mathcal{I}_n(\boldsymbol{\theta})|^{-1/2}, \tag{1.3.91}$$

and an approximate noninformative prior is one which is locally uniform in $\boldsymbol{\phi}$. The corresponding noninformative in $\boldsymbol{\theta}$ is then

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\phi}) \left| \frac{\partial(\phi_1, \ldots, \phi_k)}{\partial(\theta_1, \ldots, \theta_k)} \right|_+;$$

that is,

$$p(\boldsymbol{\theta}) \propto |\mathcal{I}_n(\boldsymbol{\theta})|^{1/2}. \tag{1.3.92}$$

From the above we obtain the following rule:

Jeffreys' rule for multiparameter problems: The prior distribution for a set of parameters is taken to be proportional to the square root of the determinant of the information matrix.

As in the case of a single parameter, Jeffreys derived this general rule by *requiring* invariance under parameter transformation. He himself pointed out that this multiparameter rule must be applied with caution, especially where scale and location parameters occur simultaneously. We first consider an application where no difficulty occurs.

Multinomial Distribution

Consider the derivation of an appropriate prior for the parameters of the multinomial distribution. Suppose the result of a trial must be to produce one of m different outcomes, the probabilities for which are $\pi_1, \pi_2, \ldots, \pi_m$. Thus, the trial might consist of the random drawing with replacement of a ball from a bag. The probabilities could then refer to the proportions $\pi' = (\pi_1, \ldots, \pi_m)$ of balls of m different colors where $\pi_m = 1 - \sum_{i=1}^{m-1} \pi_i$. Suppose n independent trials were made, resulting in a sample of $\mathbf{y}' = (y_1, \ldots, y_m)$ balls of the various types, where $y_m = n - \sum_{i=1}^{m-1} y_i$. Then

$$p(\mathbf{y} \mid \boldsymbol{\pi}) = \frac{n!}{(y_1!)(y_2!)\cdots(y_m!)} \cdot (\pi_1^{y_1})(\pi_2^{y_2}) \cdots (\pi_m^{y_m}) \tag{1.3.93a}$$

so that

$$L = \log l(\boldsymbol{\pi} \mid \mathbf{y}) = \sum_{j=1}^{m} y_j \log \pi_j. \tag{1.3.93b}$$

On differentiating, we obtain

$$b_{jj} = \frac{\partial^2 L}{(\partial \pi_j)^2} = -\frac{y_j}{\pi^2} + \frac{y_m}{\pi_m^2} \tag{1.3.94}$$

and

$$b_{ij} = \frac{\partial^2 L}{\partial \pi_i \, \partial \pi_j} = -\frac{y_m}{\pi_m^2}, \qquad i, j = 1, \ldots, m-1.$$

Taking expectations over the distribution $p(\mathbf{y} \mid \boldsymbol{\pi})$, we have

$$-E(b_{jj}) = \frac{n}{\pi_j} + \frac{n}{\pi_m}, \qquad -E(b_{ij}) = \frac{n}{\pi_m}. \tag{1.3.95}$$

After some algebraic reduction we find that

$$|\mathcal{I}_n(\boldsymbol{\pi})| = -|E\{b_{ij}\}| = n(\pi_1 \cdot \pi_2 \cdots \pi_m)^{-1}. \tag{1.3.96}$$

Thus, Jeffreys' rule says that we should take for a noninformative prior

$$p(\boldsymbol{\pi}) \propto (\pi_1 \cdot \pi_2 \cdots \pi_m)^{-1/2}. \tag{1.3.97}$$

For the case where little is known *a priori* about the probabilities, this leads to the posterior density

$$p(\boldsymbol{\pi} \mid \mathbf{y}) \propto \pi_1^{y_1 - \frac{1}{2}} \pi_2^{y_2 - \frac{1}{2}} \cdots \pi_m^{y_m - \frac{1}{2}}. \tag{1.3.98}$$

which is proportional to the likelihood for $\boldsymbol{\pi}$, with each cell frequency reduced by one half. In the particular case $k = 2$, we obtain the binomial result (1.3.26) considered earlier.

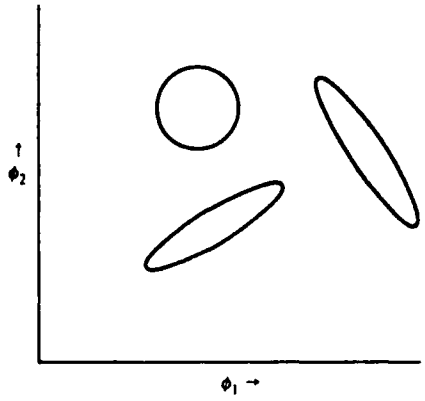


Fig. 1.3.12 Likelihood regions of different shape having the same size.

Some Comments on the Application of the Multiparameter Jeffreys' Rule

The multiparameter version of Jeffreys' rule leading to $p(\boldsymbol{\theta}) \propto |\mathcal{I}_n(\boldsymbol{\theta})|^{\frac{1}{2}}$ corresponds with a less stringent and less convincing transformation requirement on the likelihood than data translation. Specifically, if approximate Normality of the likelihood function is assumed, the rule seeks a transformation which ensures, irrespective of the data, that corresponding likelihood regions for $\boldsymbol{\phi}$ are of the same size. As illustrated in Fig. 1.3.12 for the case of $k = 2$ parameters, regions can of course be of the same size and yet be very different.

A further difficulty associated with the blanket application of Jeffreys' rule arises when parameters of different kinds are considered simultaneously. We have already seen, for example, that when the mean θ and standard deviation σ of a Normal distribution are being considered simultaneously—see Fig. 1.3.11, it is not usually appropriate to seek a transformation which produces likelihood regions of the same size. To further appreciate the difficulty, we discuss again the choice of prior for the location and scale parameters of the Normal distribution $N(\theta, \sigma^2)$.

Location and Scale Parameters

Jeffreys argued that in cases where θ and σ are known to be independent *a priori*, priors for the two parameters should be considered separately, leading to $p(\theta, \sigma) \doteq p(\theta)\, p(\sigma) \propto \sigma^{-1}$ as in (1.3.84).

Now, the information matrix of (θ, σ) is

$$\mathcal{I}_n(\theta, \sigma) = n \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{bmatrix} = n \begin{bmatrix} \mathcal{I}(\theta) & 0 \\ 0 & \mathcal{I}(\sigma) \end{bmatrix}. \tag{1.3.99}$$

If the prior independence of (θ, σ) were ignored, then application of the generalized rule would lead to a prior locally proportional to

$$|\mathcal{I}_n(\theta, \sigma)|^{1/2} \propto \sigma^{-2}, \tag{1.3.100}$$

which differs from the prior in (1.3.84) by an additional factor σ^{-1}. Some light is thrown on this factor if we consider the problem from a transformation point of view.

Suppose $\kappa(\theta)$ and $\phi(\sigma)$ are, respectively, one-to-one transformations of θ and σ. Then the information matrix of (κ, ϕ) is

$$\mathcal{I}_n[\kappa(\theta), \phi(\sigma)] = n \begin{bmatrix} \sigma^{-2}\left|\frac{d\theta}{d\kappa}\right|^2 & 0 \\ 0 & 2\sigma^{-2}\left|\frac{d\sigma}{d\phi}\right|^2 \end{bmatrix} = n \begin{bmatrix} \mathcal{I}(\kappa) & 0 \\ 0 & \mathcal{I}(\phi) \end{bmatrix}, \quad (1.3.101)$$

and by setting $\phi(\sigma) = \log \sigma$, the lower right hand element in the matrix, representing information about $\phi(\sigma)$, is made independent of σ so that

$$\mathcal{I}_n[\kappa(\theta), \log \sigma] = n \begin{bmatrix} \sigma^{-2}\left|\frac{d\theta}{d\kappa}\right|^2 & 0 \\ 0 & 2 \end{bmatrix}. \quad (1.3.102)$$

This reflects the fact that whatever transformation $\kappa(\theta)$ is made on θ, information about κ will be inversely proportional to σ^{-2}—data having a small σ lead to a more accurate determination of the location parameter θ or of any transformation $\kappa(\theta)$ of it. On the other hand, the metric for which the information is independent of θ is clearly θ itself.

The situation can be related to the likelihood regions illustrated in Fig. 1.3.11. First, the zero off-diagonal elements of the matrix (1.3.102) reflects the fact that the "axes" of the likelihood contours lie parallel to the θ and $\log \sigma$ axes. Second, independence of $\mathcal{I}(\log \sigma)$ and $\mathcal{I}(\theta)$ with respect to θ corresponds to the fact that, for a fixed s, the likelihood is merely relocated when $\bar{y}$ is changed. Third, constancy of $\mathcal{I}(\log \sigma)$ has led to a relocation of the likelihood along the $\log \sigma$ axis when s is changed. Finally, the same change in s magnifies the spread of the likelihood along the θ axis because $\mathcal{I}(\theta) \propto \sigma^{-2}$.

Thus, when the assumption of prior independence of θ and σ is incorporated, the fact that the information of θ, or any function of it, is proportional to σ^{-2} is clearly irrelevant to the choice of a noninformative prior for θ. The additional factor σ^{-1} in (1.3.100) arises only from a misapplication which ignores prior independence.

A similar situation occurs for the linear Normal model in (1.3.68). The information matrix for $\boldsymbol{\theta}$ and $\log \sigma$ is

$$\mathcal{I}_n(\boldsymbol{\theta}, \log \sigma) = \begin{bmatrix} \mathcal{I}_n(\boldsymbol{\theta}) & 0 \\ 0 & 2n \end{bmatrix}, \quad \text{where} \quad \mathcal{I}_n(\boldsymbol{\theta}) = \sigma^{-2}(\mathbf{X}'\mathbf{X})^{-1}. \quad (1.3.103)$$

Thus,

$$|\mathcal{I}_n(\boldsymbol{\theta}, \log \sigma)|^{1/2} \propto \sigma^{-k}.$$

Arguing exactly as above, the factor σ^{-k} is irrelevant and the appropriate prior distribution is $p(\boldsymbol{\theta}, \log \sigma) \propto c$ or $p(\boldsymbol{\theta}, \sigma) \propto \sigma^{-1}$ as in (1.3.82).

Finally, for the general location-scale family of (1.3.57),

$$p(y \mid \theta, \sigma) \propto \sigma^{-1} h\left(\frac{y-\theta}{\sigma}\right),$$

the information matrix of $(\theta, \log \sigma)$ is

$$\mathcal{I}_n(\theta, \log \sigma) = n \begin{bmatrix} b_1 \sigma^{-2} & 0 \\ 0 & b_2 \end{bmatrix}, \tag{1.3.104}$$

where b_1 and b_2 are two positive constants independent of (θ, σ). If θ and σ are considered independent *a priori*, then Jeffreys' rule applied separately yields

$$p(\theta, \log \sigma) \propto c \qquad \text{or} \qquad p(\theta, \sigma) \propto \frac{1}{\sigma}, \tag{1.3.105}$$

whence the corresponding posterior distribution of (θ, σ) is

$$p(\theta, \sigma \mid \mathbf{y}) \propto \sigma^{-(n+1)} \prod_{u=1}^{n} h\left(\frac{y_u - \theta}{\sigma}\right), \qquad -\infty < \theta < \infty, \quad \sigma > 0. \tag{1.3.106}$$

Necessity for Individual Consideration of Multiparameter Prior Distributions

Choice of noninformative prior distributions in multiparameter problems requires careful consideration in each particular instance. Although devices such as Jeffreys' rule can be suggestive, it is necessary to investigate transformation implications for each example in the light of any appropriate knowledge of prior independence.

1.3.7 Noninformative Prior Distributions: A Summary

In the previous sections, methods have been developed for selecting prior distributions to represent the situation where little is known *a priori* in relation to the information from the data. The concept of a "data translated likelihood" leads to a class of what we call "noninformative" priors. In the case of a single parameter, the resulting prior distributions correspond exactly to those proposed by Jeffreys on grounds of invariance.

A noninformative prior does not necessarily represent the investigator's prior state of mind about the parameters in question. It ought, however, to represent an "unprejudiced" state of mind. In this book, noninformative priors are frequently employed as a point of *reference* against which to judge the kind of unprejudiced inference that can be drawn from the data.

The phrase "knowing little" can only have meaning relative to a specific experiment. The form of a noninformative prior thus depends upon the experiment to be performed, and for two different experiments, each of which can throw light on the same parameter, the choice of "noninformative" prior can be different.

Table 1.3.1

A summary of noninformative prior and corresponding posterior distributions

Parameter(s)	Noninformative Prior	Posterior
Binomial mean π	$\pi^{-\frac{1}{2}}(1-\pi)^{-\frac{1}{2}}$	$\pi^{y-\frac{1}{2}}(1-\pi)^{n-y-\frac{1}{2}}$
Multinomial means $\pi_1, \ldots, \pi_m$	$\pi_1^{-\frac{1}{2}} \ldots \pi_m^{-\frac{1}{2}}$	$\pi_1^{y_1-\frac{1}{2}} \ldots \pi_m^{y_m-\frac{1}{2}}$
Poisson mean λ	$\lambda^{-\frac{1}{2}}$	$\lambda^{n\bar{y}-\frac{1}{2}} \exp(-n\lambda)$
Normal mean θ (σ known)	c	$\exp\left[-\dfrac{n(\theta-\bar{y})^2}{2\sigma^2}\right]$
Normal standard deviation σ (θ known)	σ^{-1}	$\sigma^{-(n+1)}\exp\left[-\dfrac{\Sigma(y_u-\theta)^2}{2\sigma^2}\right]$
Normal θ and σ	σ^{-1}	$\sigma^{-(n+1)}\exp\left[-\dfrac{n(\theta-\bar{y})^2}{2\sigma^2}-\dfrac{\Sigma(y_u-\bar{y})^2}{2\sigma^2}\right]$
Normal linear model $\boldsymbol{\theta}$ (σ known)	c	$\exp\left[-\dfrac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{2\sigma^2}\right]$
Normal linear model $\boldsymbol{\theta}$ and σ	σ^{-1}	$\sigma^{-(n+1)}\exp\left[-\dfrac{(\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}})+(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{2\sigma^2}\right]$
Location-scale family θ (σ known)	c	$\prod_{u=1}^{n} h\left(\dfrac{y_u-\theta}{\sigma}\right)$
Location-scale family σ (θ known)	σ^{-1}	$\sigma^{-(n+1)}\prod_{u=1}^{n} h\left(\dfrac{y_u-\theta}{\sigma}\right)$
Location-scale family θ and σ	σ^{-1}	$\sigma^{-(n+1)}\prod_{u=1}^{n} h\left(\dfrac{y_u-\theta}{\sigma}\right)$

When more than one parameter is involved, the problem of choosing noninformative priors can be complex. Each problem has to be considered on its merits. Careful consideration must be given to transformation implications and to knowledge of prior independence.

Considerable literature exists concerning the choice of prior distribution to characterize a state of "ignorance". Jeffreys himself has discussed a number of criteria for choosing prior distributions, invariance being the most important among them. Other contributions in this area include those of Huzurbazar (1955), Perks (1947), Welch and Peers (1963), Novick and Hall (1965), Novick (1969), Hartigan (1964, 1965) and Jaynes (1968).

Table 1.3.1 summarizes the results obtained in Sections 1.3.1–6 of the noninformative priors, and the corresponding posteriors for the parameters of the various models considered. The distributions are given in unnormalized form and we suppose that n observations are available.

1.4 SUFFICIENT STATISTICS

When discussing the example in which the mean θ of a Normal distribution was supposed unknown but σ was known, we found in (1.3.1) that the only function of the observations appearing in the likelihood, apart from the sample size n, was the sample average $\bar{y}$. Thus, for example, if σ was known to be equal to unity, the likelihood would be

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

Since the data enter Bayes' formula only through the likelihood, it follows that all other aspects of the data, with the exception of $\bar{y}$, are irrelevant in deciding the posterior distribution of θ and hence in making inferences about θ. In these circumstances, following Fisher (1922, 1925), $\bar{y}$ is said to be *sufficient* for θ, and is called a *sufficient statistic* for θ. Similarly when, for a Normal sample, θ is assumed known but σ is unknown, the likelihood corresponds to (1.3.6) and the posterior distribution employs the data only through n and $s^2 = \Sigma(y - \theta)^2/n$. In this case, s^2 is said to be sufficient for σ (or for σ^2). Further, when both θ and σ are unknown, the joint likelihood function for (θ, σ), given a random sample from the Normal distribution $N(\theta, \sigma^2)$, is

$$l(\theta, \sigma \mid \mathbf{y}) = \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{u=1}^{n}(y_u - \theta)^2\right], \tag{1.4.1}$$

which, as in (1.3.77a), by setting $s^2 = \Sigma(y_u - \bar{y})^2/(n-1)$ can be written

$$l(\theta, \sigma \mid \mathbf{y}) = \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y} - \theta)^2\right]\right\}. \tag{1.4.2}$$

Apart from n, the only functions of the observations involved are $\bar{y}$ and s^2, which are said to be *jointly sufficient* for θ and σ. If n is given, $\bar{y}$ and s^2 can be constructed from knowledge of Σy_u and $\Sigma(y_u - \bar{y})^2$ or from Σy_u and Σy_u^2, so that any one of these pairs of quantities are jointly sufficient for θ and σ.

Of course, it would also be true that the *three* quantities Σy_u, $\Sigma(y_u - \bar{y})^2$, Σy_u^2 were sufficient for θ and σ. However, since

$$\Sigma(y_u - \bar{y})^2 = \Sigma y_u^2 - n^{-1}(\Sigma y_u)^2$$

there is an obvious redundancy. The notion is therefore used of a "minimally sufficient set" of statistics, which contains the smallest number of independent functions of the data needed to write down the likelihood. In this example, assuming n given, a minimally sufficient set contains two functions which could be chosen in any way which allows the likelihood to be written down. In particular, they could be any one of the three choices $(\bar{y}, s^2)$, $(\Sigma y_u, \Sigma(y_u - \bar{y})^2)$, or $(\Sigma y_u, \Sigma y_u^2)$. Since n is also needed to write down the likelihood, this quantity is sometimes treated as a statistic and added to the sufficient set.

We see that for Normal samples sufficient statistics are available which conveniently "match" the parameters. For example, $\bar{y}$ and s^2 are sample quantities which one would expect to supply information about θ and σ.

Convenient parsimonious sets of sufficient statistics do exist for some other distributions. In general, however, if we have k parameters in the likelihood, the minimal sufficient set of $q \geqslant k$ functions of the data which appear in the likelihood function will not be such that $q = k$. Consider, for example, the distribution

$$p(y \mid \theta, \sigma) \propto \sigma^{-1} \exp\left[-\left(\frac{y-\theta}{\sigma}\right)^4\right], \quad -\infty < y < \infty. \tag{1.4.3}$$

The likelihood function based upon n independent observations is

$$l(\theta, \sigma \mid \mathbf{y}) \propto \sigma^{-n} \exp\left[-\sum_{u=1}^{n}\left(\frac{y_u - \theta}{\sigma}\right)^4\right] \tag{1.4.4}$$

$$= \sigma^{-n} \exp\left[\frac{1}{\sigma^4}\left(-S_4 + 4\theta S_3 - 6\theta^2 S_2 + 4\theta^3 S_1 - \theta^4\right)\right] \tag{1.4.5}$$

where $S_p = \sum_{u=1}^{n} y_u^p$. In this case, then, a minimal sufficient set of statistics for the $k = 2$ parameters θ and σ consists of the $q = 4$ sums S_1, S_2, S_3, and S_4. Note, however, that if the power in the exponent in (1.4.3) had not been an integer (suppose, for example, it had been 4.1 instead of 4.0) then no finite expansion of the form of (1.4.5) would have been possible, and the $q = n$ observations themselves would have been a minimally sufficient set of statistics.

In general, we may define sufficient statistics as follows.

Definition (1.4.1) Let $\mathbf{y}' = (y_1, \ldots, y_n)$ be a vector of observations whose distribution depends upon the k parameters $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_k)$. Let $\mathbf{t}' = (t_1, \ldots, t_q)$

be q functions of $\mathbf{y}$. Then the set of statistics $\mathbf{t}$ is said to be jointly sufficient for $\boldsymbol{\theta}$ if the likelihood function $l(\boldsymbol{\theta} \mid \mathbf{y})$ can be expressed in the form

$$l(\boldsymbol{\theta} \mid \mathbf{y}) \propto g(\boldsymbol{\theta} \mid \mathbf{t}), \tag{1.4.6}$$

and provided the ranges of $\boldsymbol{\theta}$, if dependent on the observations, can also be expressed as functions of $\mathbf{t}$.

Thus, considering distributions which have been used as examples in this chapter, sufficient statistics exist for the parameters (θ, σ) in the Normal distribution, for $(\boldsymbol{\theta}, \sigma)$ of the Normal linear model in (1.3.68), for the Poisson mean λ in (1.3.39), for the binomial mean π in (1.3.21), for the multinomial means $(\pi_1, \ldots, \pi_m)$ in (1.3.93a), and for (θ, σ) in the "fourth power" distribution in (1.4.3). In the case of a single parameter θ, if y is a random sample from the distribution $p(y \mid \theta)$ and the range of y does not depend on θ, then it was shown by Pitman (1936) that a *single* sufficient statistic exists for θ if and only if $p(y \mid \theta)$ is a member of the exponential family previously referred to in (1.3.38).

When the ranges of the observations $\mathbf{y}$ are independent of $\boldsymbol{\theta}$, a useful property of sufficient statistics is given by the following lemma.

Lemma 1.4.1 Let $\mathbf{t}$ be jointly sufficient for $\boldsymbol{\theta}$, having joint distribution $p(\mathbf{t} \mid \boldsymbol{\theta})$. Then,

$$l(\boldsymbol{\theta} \mid \mathbf{y}) \propto l_1(\boldsymbol{\theta} \mid \mathbf{t}) \qquad \text{where} \qquad l_1(\boldsymbol{\theta} \mid \mathbf{t}) \propto p(\mathbf{t} \mid \boldsymbol{\theta}). \tag{1.4.7}$$

In other words, the likelihood function obtained from the distribution of $\mathbf{t}$ is the same as that obtained from the distribution of $\mathbf{y}$.

Proofs of the lemma can be found in standard text books such as Kendall and Stuart (1961) and Wilks (1962). Two examples follow.

1. *Normal mean, variance known.* We have seen in (1.3.1) that if $\mathbf{y}' = (y_1, \ldots, y_n)$ is a random sample from $N(\theta, \sigma^2)$, where σ^2 is assumed known, then the likelihood function is

$$l(\theta \mid \sigma, \mathbf{y}) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right]. \tag{1.4.8}$$

Now, the sample mean $\bar{y}$ is distributed as

$$p(\bar{y} \mid \theta, \sigma^2) = \frac{\sqrt{n}}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{n}{2\sigma^2}(\bar{y} - \theta)^2\right], \qquad -\infty < \bar{y} < \infty, \tag{1.4.9}$$

so that, given $\bar{y}$,

$$l_1(\theta \mid \sigma, \bar{y}) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right], \tag{1.4.10}$$

which is the same as in (1.4.8).

2. *Normal distribution, both mean and variance unknown.* In this case, the likelihood function $l(\theta, \sigma \mid \mathbf{y})$ based upon the n independent observations $\mathbf{y}$ is that given in (1.4.2). Now, it is well known that

a) $\bar{y}$ is distributed as $N(\theta, \sigma^2/n)$,

b) $(n-1)s^2 = \Sigma_u (y_u - \bar{y})^2$ is distributed as $\sigma^2 \chi^2_{n-1}$, and

c) $\bar{y}$ and s^2 are statistically independent.

Thus,

$$p(\bar{y}, s^2 \mid \theta, \sigma) = p(\bar{y} \mid \theta, \sigma^2)\, p(s^2 \mid \sigma^2), \tag{1.4.11}$$

where $p(\bar{y} \mid \theta, \sigma^2)$ is that in (1.4.9) and

$$p(s^2 \mid \sigma^2) = \frac{1}{\Gamma[\frac{1}{2}(n-1)]} \left(\frac{n-1}{2\sigma^2}\right)^{\frac{1}{2}(n-1)} (s^2)^{\frac{1}{2}(n-1)-1} \times \exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right], \qquad s^2 > 0. \tag{1.4.12}$$

It follows that, given $(\bar{y}, s^2)$,

$$l(\theta, \sigma \mid \bar{y}, s^2) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\theta - \bar{y})^2\right]\right\}, \tag{1.4.13}$$

which is identical with (1.4.2).

1.4.1 Relevance of Sufficient Statistics in Bayesian Inference

Sufficient statistics play a vital role in sampling theory. For, if inferences about fixed parameters are to be made using the distributional properties of statistics which are functions of the data, then, to avoid inefficiency due to leakage of information, it is essential that a small minimally sufficient set of statistics be available containing all the information about the parameters.

By happy mathematical accident such sets of sufficient statistics do exist for a number of important distributions and, in particular, for the Normal distribution. However, serious difficulties can accompany the exploration of less restricted models which may be motivated by scientific interest, but for which no convenient set of sufficient statistics happens to be available.

Because Bayesian analysis is concerned with the distribution of *parameters*, given known (fixed) data, it does not suffer from this artificial constraint. It does not matter whether or not the distribution of interest happens to have the special form which yields sufficient statistics. For example, the likelihood, and hence the posterior density, can be calculated with almost the same ease from (1.4.4), which expresses the likelihood in terms of the data, as it can from (1.4.5), which expresses it in terms of the sufficient statistics alone. Furthermore, very little more effort would be needed to compute (1.4.4) if the power in the exponent were 4.1 or some other noninteger value.

1.4.2 An Example Using the Cauchy Distribution

To illustrate these points further, we consider the problem of making inferences about the location parameter θ of the Cauchy distribution

$$p(y \mid \theta) = \pi^{-1}[1 + (y - \theta)^2]^{-1}, \qquad -\infty < y < \infty, \tag{1.4.14}$$

where from the form of the distribution it is apparent that no summarizing statistics exist and that the observations themselves are a minimum sufficient set. This fact does not embarrass the Bayesian approach. The Cauchy distribution is a special case of the location scale family (1.3.57) and, using the argument leading to (1.3.61), a noninformative prior is locally uniform in θ, whence the posterior density function is immediate. To provide numerical illustration, a sample of $n = 5$ observations (11.4, 7.3, 9.8, 13.7, 10.6) were randomly drawn from the Cauchy distribution shown in Fig. 1.4.1. Thus, assuming little were known about θ *a priori*, the posterior distribution is approximately

$$p(\theta \mid \mathbf{y}) \doteq cH(\theta), \qquad -\infty < \theta < \infty, \tag{1.4.15}$$

where

$$H(\theta) = 10^5[1 + (7.3 - \theta)^2]^{-1}[1 + (9.8 - \theta)^2]^{-1} \ldots [1 + (13.7 - \theta)^2]^{-1},$$

the factor 10^5 is merely a convenient multiplier, and c is the normalizing constant. The posterior distribution obtained by evaluating this expression for a suitable series of values of θ is shown in Fig. 1.4.2. Thus, in spite of the fact that we do not have a sufficient statistic for θ, the posterior distribution, from which inferences can be made, is easily determined.

Calculation of $p(\theta \mid \mathbf{y})$

To obtain the density explicitly, we need to determine the normalizing factor c. That is, we have to evaluate the integral

$$c^{-1} = \int_{-\infty}^{+\infty} H(\theta)\, d\theta. \tag{1.4.16}$$

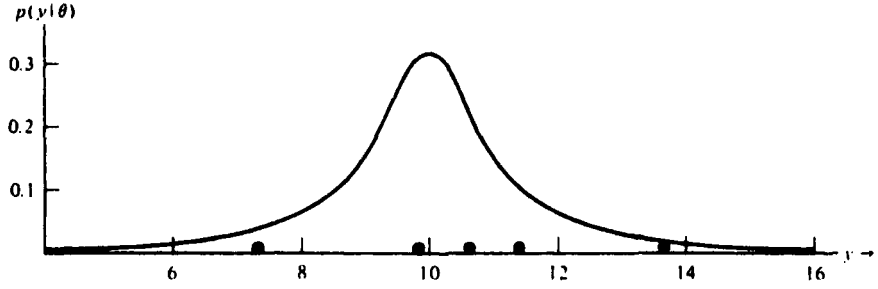


Fig. 1.4.1 Density curve of a Cauchy distribution (dots show 5 observations drawn from the distribution).

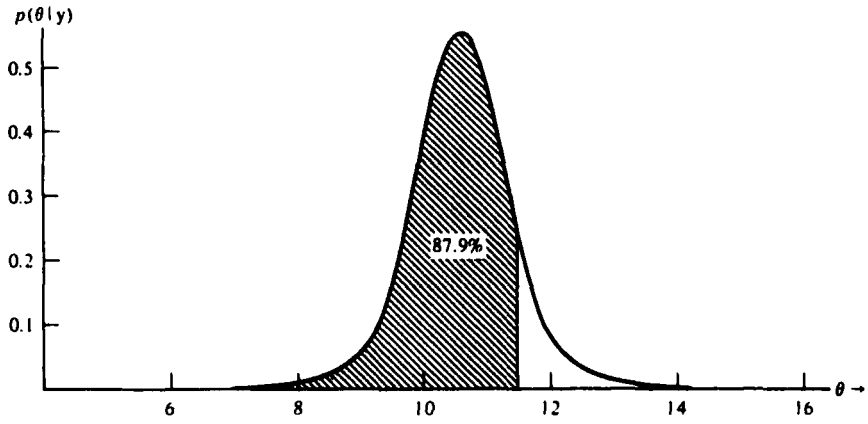


Fig. 1.4.2 Posterior distribution for the location parameter θ of a Cauchy distribution (noninformative prior), given the data shown in Fig. 1.4.1.

Also, to obtain the probability that θ is less than some value θ_0, we need to evaluate the ratio

$$\left[\int_{-\alpha}^{\theta_0} H(\theta)\,d\theta\right] \Big/ \left[\int_{-\alpha}^{+\alpha} H(\theta)\,d\theta\right]. \tag{1.4.17}$$

Frequently the integrals which occur in applications of Bayes' formula will not possess convenient solutions in closed form, or solutions which have been tabulated. However, this is of little practical importance.

In the first place, the best way to convey to the experimenter what the data tell him about θ is to show him a picture of the posterior distribution. For this purpose, the scale of the vertical axis is superfluous and it is sufficient to plot $H(\theta)$. In the second place, just as the question of the existence of convenient sufficient statistics ought to be irrelevant, so should the question of whether or not an integral happens to be one which can be expressed in terms of tabled functions. As the Bayesian approach sets us free from the yoke of sufficiency, so numerical integration and the availability of computers set us free from the need to worry about the "integrability in closed form" of the function.

For univariate distributions, even summing ordinates or drawing the distribution on graph paper and counting squares could approximate the integral with sufficient accuracy for most practical purposes. Since small differences in probability cannot be appreciated by the human mind, there seems little point in being excessively precise about uncertainty.

For illustration, a specimen calculation using Simpson's rule is shown in Table 1.4.1. Suppose for the Cauchy example that for some reason $p(\theta|y)$ itself is needed and not merely $H(\theta)$, and that the probability $\Pr\{\theta < 11.5\}$ is required. The first column in the table shows θ at intervals of 0.5 over the range of interest. The second column shows the corresponding value of $H(\theta)$ to the nearest whole number. The third column shows the

approximate integral given by Simpson's rule. Thus, for example,

$$\int_{-\infty}^{8.5} H(\theta)\,d\theta \doteq \frac{0.5}{3}[(1\times 0) + (4\times 1) + (2\times 2) + (4\times 5) + (1\times 11)] = 6.5. \quad (1.4.18)$$

Proceeding in this way we find that

$$c^{-1} = \int_{-\infty}^{+\infty} H(\theta)\,d\theta \doteq 535.0 \quad (1.4.19)$$

so that $c = 0.001869$, whence $p(\theta \mid \mathbf{y})$ is calculated and entered in the fourth column. The values for the cumulative probability

$$\int_{-\infty}^{\theta_0} p(\theta \mid \mathbf{y})\,d\theta = c\int_{-\infty}^{\theta_0} H(\theta)\,d\theta \quad (1.4.20)$$

are given in the fifth column. In particular (see Fig. 1.4.2), we find that

Table 1.4.1

Calculation of the posterior density function and cumulative distribution function for the location parameter θ of a Cauchy distribution

θ or θ_0	$H(\theta)$	$\int_{-\infty}^{\theta_0} H(\theta)\,d\theta$	$p(\theta\|\mathbf{y}) = cH(\theta)$	$\int_{-\infty}^{\theta_0} p(\theta\|\mathbf{y})\,d\theta = c\int_{-\infty}^{\theta_0} H(\theta)\,d\theta$
6.5	0	0	0.000	0·000
7.0	1		0.002	
7.5	2	1.0	0.004	0.002
8.0	5		0.009	
8.5	11	6.5	0.021	0.012
9.0	28		0.052	
9.5	83	40.8	0.155	0.076
10.0	196		0.336	
10.5	291	233.8	0.544	0.437
11.0	250		0.467	
11.5	129	470.5	0.241	0.879
12.0	47		0.088	
12.5	17	526.2	0.032	0.983
13.0	7		0.013	
13.5	2	534.0	0.004	0.998
14.0	1		0.002	
14.5	0	535.0	0.000	1.000

$$c^{-1} = \int_{-\infty}^{+\infty} H(\theta)\,d\theta = 535.0, \qquad c = 0.001869.$$

$\Pr\{\theta < 11.5\} = 0.879$. Values of the cumulative probability for intermediate values can be found by interpolation in column five. Greater accuracy is obtainable by using a finer interval in θ.

1.5 CONSTRAINTS ON PARAMETERS

Examples occur later in this book where, as part of the model, certain constraints must be imposed on the values which the parameters $\boldsymbol{\theta}$ can take. Such problems can usually be dealt with by choosing the prior distribution so as to include the constraint. Alternatively, it is sometimes more convenient to solve a fictitious unconstrained problem, and then modify the solution to take account of the constraint.

In general, let Ω be the unconstrained parameter space of $\boldsymbol{\theta}$ and let C be a constraint, such that

$$C: \boldsymbol{\theta} \in \Omega_C, \tag{1.5.1}$$

where Ω_C is a subspace of Ω. Let $\boldsymbol{\theta}_S$ be a subset of $\boldsymbol{\theta}$ (which could be $\boldsymbol{\theta}$ itself), and $R_S \subset \Omega_C$ be a region in the parameter space of $\boldsymbol{\theta}_S$.

Then, by definition of conditional probability, the posterior probability that $\boldsymbol{\theta}_S \in R_S$, given the constraint C, is

$$\Pr\{\boldsymbol{\theta}_S \in R_S \mid C, \mathbf{y}\} = \Pr\{\boldsymbol{\theta}_S \in R_S \mid \mathbf{y}\} \frac{\Pr\{C \mid \boldsymbol{\theta}_S \in R_S, \mathbf{y}\}}{\Pr\{C \mid \mathbf{y}\}} . \tag{1.5.2}$$

It follows that the posterior distribution of $\boldsymbol{\theta}_S$ given the constraint C can be written

$$p(\boldsymbol{\theta}_S \mid C, \mathbf{y}) = p(\boldsymbol{\theta}_S \mid \mathbf{y}) \frac{\Pr\{C \mid \boldsymbol{\theta}_S, \mathbf{y}\}}{\Pr\{C \mid \mathbf{y}\}} . \tag{1.5.3}$$

It sometimes happens that C takes the form

$$C: \mathbf{f}(\boldsymbol{\theta}) = \mathbf{d} \tag{1.5.4}$$

where $\mathbf{f}$ is a vector of q functions of $\boldsymbol{\theta}$, and $\mathbf{d}$ a vector of q constants. In this case,

$$p(\boldsymbol{\theta}_S \mid C, \mathbf{y}) = p(\boldsymbol{\theta}_S \mid \mathbf{y}) \frac{p(C \mid \boldsymbol{\theta}_S, \mathbf{y})}{p(C \mid \mathbf{y})} . \tag{1.5.5}$$

Note that $\Pr(C \mid \mathbf{y})$ and $p(C \mid \mathbf{y})$, are constants independent of $\boldsymbol{\theta}_S$. Thus, the posterior distribution of $\boldsymbol{\theta}_S$, given the constraint C, is equal to the posterior distribution for $\boldsymbol{\theta}_S$ which would have been obtained if no such constraint were applied, multiplied by a modifying factor. This modifying factor is proportional to the conditional probability, or the conditional density, of the constraint C given $\boldsymbol{\theta}_S$.

As an example, suppose we wished to make inferences about the percentage conversion of a certain chemical obtained in a particular experiment. Suppose the percentage conversion was determined by a biological assay method which was unbiased but was subject to fairly large approximately Normally distributed errors having known standard deviation $\sigma = 4$. Suppose finally that the results of four analytical determinations were $y_1 = 93, y_2 = 101, y_3 = 100$ and $y_4 = 98$, yielding a sample average of $\bar{y} = 98$. We need to consider what inferences could be made about θ, bearing in mind that values of θ greater than 100 are impossible.

If it were reasonable to suppose that in the relevant neighborhood the prior of θ is locally uniform for $\theta < 100$, then the problem could be solved by the straightforward application of Bayes' theorem. We have

$$p(\theta \mid \mathbf{y}) \propto l(\theta \mid \mathbf{y}) p(\theta) \tag{1.5.6}$$

where in the *relevant* region

$$p(\theta) \begin{cases} \doteq \text{const} & \theta < 100, \\ = 0 & \theta > 100. \end{cases} \tag{1.5.7}$$

Thus

$$p(\theta \mid \mathbf{y}) = c \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right], \qquad \theta < 100, \tag{1.5.8}$$

where

$$c^{-1} = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \int_{-\infty}^{100} \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right] d\theta, \qquad n = 4, \qquad \bar{y} = 98.$$

As illustrated in Fig. 1.5.1, the posterior distribution of θ is proportional to the unconstrained likelihood for $\theta < 100$ and is zero for $\theta > 100$. The posterior distribution for θ is thus a Normal distribution $N(\bar{y}, \sigma^2/n)$, with $\bar{y} = 98$ and $\sigma/\sqrt{n} = 2$, truncated from above at $\theta = 100$. The normalizing constant in (1.5.8) is $c = 1.189$. It is chosen so that its reciprocal $c^{-1} = 0.8413$ is the area under unit Normal curve truncated at one standard deviation above the mean.

To illustrate that results of this kind can be obtained equally well by an application of (1.5.3), consider first the (fictitious) unconstrained problem in which θ is not limited in any way and is supposed to have a locally uniform distribution over the whole region in which the likelihood is appreciable. With this setup the *unconstrained* posterior distribution of θ would be an untruncated Normal distribution having a standard deviation of 2 and centered at $\bar{y} = 98$.

For this problem, the constraint is $C: \theta < 100$ so that the modifying factor of (1.5.3) is

$$\frac{\Pr\{C \mid \theta, \mathbf{y}\}}{\Pr\{C \mid \mathbf{y}\}} = \frac{\Pr\{\theta < 100 \mid \theta, \mathbf{y}\}}{\Pr\{\theta < 100 \mid \mathbf{y}\}}. \tag{1.5.3}$$

Now the numerator of this expression is one if $\theta < 100$, and zero otherwise. Furthermore, the denominator is the unconstrained posterior probability that $\theta < 100$ given the data. This is the probability that an unrestricted Normally distributed random variable with mean 98 and standard deviation 2 would not exceed 100 and is precisely the same as c^{-1} in (1.5.8). The effect, then, of the modifying factor is to multiply the Normal density by 1.189 if $\theta < 100$, and by zero if $\theta > 100$, leading to the same result as before.

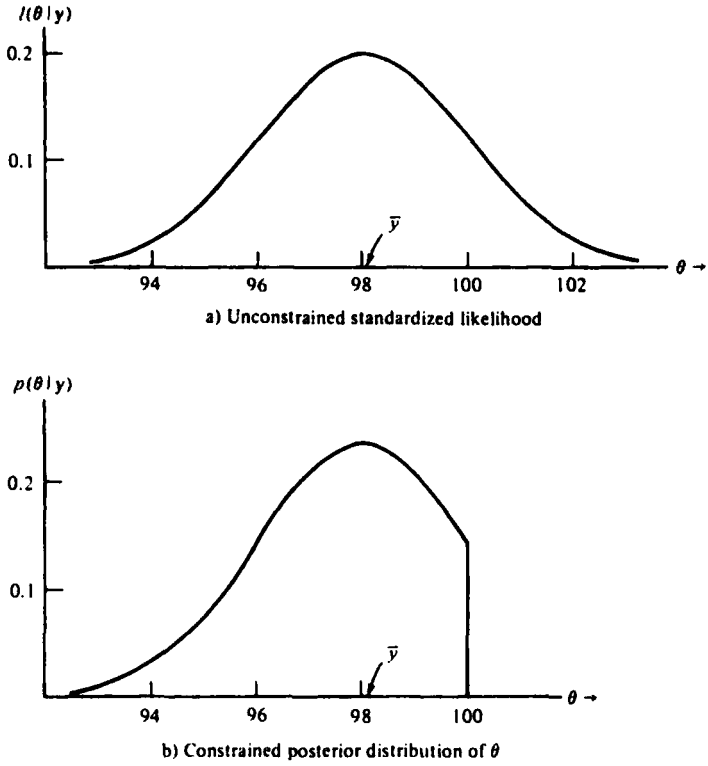


Fig. 1.5.1 Posterior distribution of the Normal mean θ subject to the constraint $\theta < 100$.

The device of first solving a fictitious unconstrained problem and then applying constraints to the solution is particularly useful for variance component problems discussed in Chapters 5 and 6. In deriving solutions and in understanding them, it is helpful to begin by solving the unconstrained problem, which allows negative components of variance, and then constraining the solution.

1.6 NUISANCE PARAMETERS

Frequently the distribution of observations y depends not only upon a set of r parameters $\boldsymbol{\theta}_1 = (\theta_1, \ldots, \theta_r)$ of interest, but also on a set of, say $t - r$ further nuisance parameters $\boldsymbol{\theta}_2 = (\theta_{r+1}, \ldots, \theta_t)$. Thus we may wish to make inferences about the mean $\theta = \theta_1$ of a Normal population with unknown variance $\sigma^2 = \theta_2$. Here the parameter of interest, or inference parameter, is the mean $\theta = \theta_1$, while the nuisance, or incidental, parameter is $\sigma^2 = \theta_2$. Again, we may wish to make inferences about a single parameter θ_i in the Normal theory linear model of (1.3.68) involving $k + 1$ parameters $\theta_1, \theta_2, \ldots, \theta_k$ and σ^2. In this case, $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k$ and σ^2 all occur as nuisance parameters. In the above examples, where sufficient statistics exist for all the parameters, no particular difficulty is encountered with the sampling theory approach. But when this is not so, difficulties arise in dealing with nuisance parameters by non-Bayesian methods. Furthermore, even when sufficient statistics are available, examples can occur in sampling theory where there is difficulty in eliminating nuisance parameters. One such example is the Behrens–Fisher problem which will be discussed in Section 2.5.

In the Bayesian approach, "overall" inferences about $\boldsymbol{\theta}_1$ are completely determined by the posterior distribution of $\boldsymbol{\theta}_1$, obtained by "integrating out" the nuisance parameter $\boldsymbol{\theta}_2$ from the joint posterior distribution of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Thus,

$$\int_{R_2} p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \mathbf{y})\, d\boldsymbol{\theta}_2 = p(\boldsymbol{\theta}_1 \mid \mathbf{y}), \tag{1.6.1}$$

where R_2 denotes the appropriate region of $\boldsymbol{\theta}_2$.

Now we can write the joint posterior distribution as the product of the conditional (posterior) distribution of $\boldsymbol{\theta}_1$ given $\boldsymbol{\theta}_2$ and the marginal (posterior) distribution of $\boldsymbol{\theta}_2$,

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \mathbf{y}) = p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \mathbf{y}) p(\boldsymbol{\theta}_2 \mid \mathbf{y}). \tag{1.6.2}$$

The posterior distribution of $\boldsymbol{\theta}_1$ can be written

$$p(\boldsymbol{\theta}_1 \mid \mathbf{y}) = \int_{R_2} p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \mathbf{y}) p(\boldsymbol{\theta}_2 \mid \mathbf{y})\, d\boldsymbol{\theta}_2, \tag{1.6.3}$$

in which the marginal (posterior) distribution $p(\boldsymbol{\theta}_2 \mid \mathbf{y})$ of the nuisance parameters acts as a weight function multiplying the conditional distribution $p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \mathbf{y})$ of the parameters of interest.

1.6.1 Application to Robustness Studies

It is often helpful in understanding a problem and the nature of the conclusions which can safely be drawn, to consider not only $p(\boldsymbol{\theta}_1 \mid \mathbf{y})$, but also the components of the integrand on the right-hand side of (1.6.3). One is thus led to consider the conditional distributions of $\boldsymbol{\theta}_1$ for particular values of the

nuisance parameters θ_2 in relation to the distribution of the postulated values of the nuisance parameters.

In particular, in judging the robustness of the inference relative to characteristics such as non-Normality and lack of independence between errors, the nuisance parameters θ_2 can be measures of departure from Normality and independence. The distribution of the parameters of interest θ_1, conditional on some specific choice $\theta_2 = \theta_{20}$, will indicate the nature of the inference which we could draw if the corresponding set of assumptions (for example, the assumptions of Normality with uncorrelated errors) are made, while the marginal posterior density $p(\theta_2 = \theta_{20} \mid y)$ reflects the plausibility of such assumptions being correct. The marginal distribution $p(\theta_1 \mid y)$, obtained by integrating out θ_2, indicates the overall inference which can be made when proper weight is given to the various possible assumptions in the light of the data and their initial plausibility.

Examples using such an approach will be considered in detail in Chapters 3 and 4.

1.6.2 Caution in Integrating Out Nuisance Parameters

As has been emphasized by Barnard†, cation should be exercised in integrating out nuisance parameters. In particular, if the conditional distribution $p(\theta_1 \mid \theta_2, y)$ in (1.6.3) changes drastically as θ_2 is changed, we would wish to be made aware of this. It is true that whatever the situation, $p(\theta_1 \mid y)$ will theoretically yield the distribution of θ_1. However, in cases where $p(\theta_1 \mid \theta_2, y)$ changes rapidly as θ_2 is changed, great reliance is being placed on the precise applicability of the weight function $p(\theta_2 \mid y)$, and it would be wise to supplement $p(\theta_1 \mid y)$ with auxiliary information.

Thus, if we wished to make inferences about θ_1 alone by integrating out θ_2 and it was found that $p(\theta_1 \mid \theta_2, y)$ was very sensitive to changes in θ_2, it would be important to examine carefully the distribution $p(\theta_2 \mid y)$, which summarizes the information about θ_2 in the light of the data and prior knowledge. Two situations might occur:

1. If $p(\theta_2 \mid y)$ were sharp, with most of its probability mass concentrated over a small region about its mode $\hat{\theta}_2$, then integrating out θ_2 would be nearly equivalent to assigning the modal value to θ_2 in the conditional distribution $p(\theta_1 \mid \theta_2, y)$, so that

$$p(\theta_1 \mid y) \doteq p(\theta_1 \mid \hat{\theta}_2, y). \tag{1.6.4}$$

   Thus, even though inferences on θ_1 would have been sensitive to θ_2 over a wider range, the posterior distribution $p(\theta_2 \mid y)$ would contain so much information about θ_2 as essentially to rule out values of θ_2 not close to $\hat{\theta}_2$.

2. On the other hand, if $p(\theta_2 \mid y)$ were rather flat, indicating there was little information about θ_2 from prior knowledge and from the sample, this sensitivity would warn that

---

† Personal communication.

we should if possible obtain more information about θ_2 so that the inferences about θ_1 could be sharpened. If this were not possible, then as well as reporting the marginal distribution $p(\theta_1 \mid y)$ obtained by integration, it would be wise to add information showing how $p(\theta_1 \mid \theta_2, y)$ changed over the range in which $p(\theta_2 \mid y)$ was appreciable.

1.7 SYSTEMS OF INFERENCE

It is not our intention here to study exhaustively and to compare the various systems of statistical inference which have from time to time been proposed. We assume familarity with the sampling theory approach to statistical inference, and, in particular, with significance tests, confidence intervals, and the Neyman-Pearson theory of hypothesis testing. The main differences between sampling theory inference and Bayesian inference are outlined below.

In sampling theory we are concerned with making inferences about unknown parameters in terms of the sampling distributions of statistics, which are functions of the observations.

1. The probabilities we calculate refer to the frequency with which different values of statistics (arising from sets of data *other* than those which have actually happened) could occur for some *fixed but unknown values of the parameters*. The theory does not employ a prior distribution for the parameters, and the relevance of the probabilities generated in such manner to inferences about the parameters has been questioned (see, for example, Jeffreys, 1961). Furthermore, once we become involved in discussing the probabilities of sets of data which have not actually occurred, we have to decide which "reference set" of groups of data which have not actually occurred we are going to contemplate, and this can lead to further difficulties (see, for example, Barnard 1947).

2. If we accept the relevance of sampling theory inference, then for finite samples we can claim to know *all* that the data have to tell us about the parameters *only* if the problem happens to be one for which all aspects of the data which provide information about the parameter values are expressible in terms of a convenient set of sufficient statistics.

3. Usually when making inferences about a set of parameters of primary interest, we must also take account of nuisance parameters necessary to the specification of the problem. Except where suitable sufficient statistics exist, it is difficult to do this with sampling theory.

4. Using sampling theory it is difficult to take account of constraints which occur in the specification of the parameter space.

By contrast, in Bayesian analysis, inferences are based on probabilities associated with *different* values of parameters which could have given rise to the *fixed* set of data which has actually occurred. In calculating such probabilities

we must make assumptions about prior distributions, but we are not dependent upon the existence of sufficient statistics, and no difficulty occurs in taking account of parameter constraints.

1.7.1 Fiducial Inference and Likelihood Inference

Apart from the sampling and the Bayesian approaches, two other modes of inference, proposed originally by Fisher (1922, 1930, 1959), have also attracted the attention of statisticians. These are fiducial inference and likelihood inference. Both are in the spirit of Bayesian theory rather than sampling theory, in that they consider inferences that can be made about *variable* parameters given a set of data *regarded as fixed*. Indeed, fiducial inference has been described by Savage (1961b) as "an attempt to make the Bayesian omelette without breaking the Bayesian eggs." It has been further developed by Fraser (1968), and Fraser and Haq (1969), using what they refer to as structural probability.

Although this approach does not employ prior probability, Fisher made it clear that fiducial inference was intended to cover the situation where nothing was known about the parameters *a priori*, and the solutions which are accessible to this method closely parallel those obtained from Bayes theorem with non-informative prior distributions. For example, one early application of fiducial inference was to the so-called Behrens–Fisher problem of comparing the means of two Normally distributed populations with unknown variances not assumed to be equal. The fiducial distribution of the difference between the population means is identical with the posterior distribution of the same quantity, first obtained by Jeffreys (1961), when noninformative prior distributions are taken for the means and variances. By contrast, Welch's sampling theory solution (1938, 1947), does not parallel this result. We discuss the Bayesian solution in Section 2.5 in more detail.

While Fisher's employment of maximum likelihood estimates was followed up with enthusiasm, by sampling theorists, few workers took account of his suggestion for considering not only the maximum but the whole likelihood function. Notable exceptions are to be found in Barnard (1949), Barnard, Jenkins and Winsten (1962) and Birnbaum (1962). Barnard has frequently stated his opinion that inferences ought to be drawn by studying the likelihood function. As we have seen earlier in Section 1.3, if the likelihood function can be plotted in an appropriate metric, it is identical with the posterior distribution using a non-informative prior. However, the likelihood approach also suffers from fundamental difficulties. It is meaningless, for example, to integrate the likelihood in an attempt to obtain "marginal likelihoods." Yet if the whole likelihood function supplies information about the parameters jointly, one feels it should be able to say something about them individually. Thus, while fiducial inference and likelihood inference each can lead to an analysis similar to a Bayesian analysis with a noninformative prior, each is frustrated in its own particular way from possessing generality.

APPENDIX A1.1
COMBINATION OF A NORMAL PRIOR AND A NORMAL LIKELIHOOD

Suppose *a priori* a parameter θ is distributed as

$$p(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\theta - \theta_0}{\sigma_0}\right)^2\right], \qquad -\infty < \theta < \infty, \tag{A1.1.1}$$

and the likelihood function of θ is proportional to a Normal function

$$l(\theta \mid y) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta - x}{\sigma_1}\right)^2\right], \tag{A1.1.2}$$

where x is some function of the observations y. Then the posterior distribution of θ given the data y is

$$p(\theta \mid y) = \frac{p(\theta) l(\theta \mid y)}{\int_{-\infty}^{\infty} p(\theta) l(\theta \mid y)\, d\theta}$$

$$= \frac{f(\theta \mid y)}{\int_{-\infty}^{\infty} f(\theta \mid y)\, d\theta}, \qquad -\infty < \theta < \infty, \tag{A1.1.3}$$

where

$$f(\theta \mid y) = \exp\left\{-\frac{1}{2}\left[\left(\frac{\theta - \theta_0}{\sigma_0}\right)^2 + \left(\frac{x - \theta}{\sigma_1}\right)^2\right]\right\}. \tag{A1.1.4}$$

Using the identity

$$A(z - a)^2 + B(z - b)^2 = (A + B)(z - c)^2 + \frac{AB}{A + B}(a - b)^2 \tag{A1.1.5}$$

with

$$c = \frac{1}{A + B}(Aa + Bb),$$

we may write

$$\left(\frac{\theta - \theta_0}{\sigma_0}\right)^2 + \left(\frac{x - \theta}{\sigma_1}\right)^2 = (\sigma_0^{-2} + \sigma_1^{-2})(\theta - \bar{\theta})^2 + d,$$

where

$$\bar{\theta} = \frac{1}{\sigma_0^{-2} + \sigma_1^{-2}}(\sigma_0^{-2}\theta_0 + \sigma_1^{-2}x),$$

and d is a constant independent of θ. Thus,

$$f(\theta \mid y) = \exp\left(-\frac{d}{2}\right) \exp\left[-\tfrac{1}{2}(\sigma_0^{-2} + \sigma_1^{-2})(\theta - \bar{\theta})^2\right], \tag{A1.1.6}$$

so that

$$\int_{-\infty}^{\infty} f(\theta \mid \mathbf{y})\, d\theta = \exp\left(-\frac{d}{2}\right) \int_{-\infty}^{\infty} \exp\left[-\tfrac{1}{2}(\sigma_0^{-2} + \sigma_1^{-2})(\theta - \bar{\theta})^2\right] d\theta$$

$$= \sqrt{2\pi}\,(\sigma_0^{-2} + \sigma_1^{-2})^{-1/2} \exp(-d/2). \qquad \text{(A1.1.7)}$$

It follows that

$$p(\theta \mid \mathbf{y}) = \frac{(\sigma_0^{-2} + \sigma_1^{-2})^{1/2}}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}(\sigma_0^{-2} + \sigma_1^{-2})(\theta - \bar{\theta})^2\right],$$

$$-\infty < \theta < \infty, \qquad \text{(A1.1.8)}$$

which is the Normal distribution

$$N[\bar{\theta}, (\sigma_0^{-2} + \sigma_1^{-2})^{-1}].$$