1

Confidence, likelihood, probability: An invitation

This chapter is an invitation to the central themes of the book: confidence, likelihood, probability and confidence distributions. We sketch the historical backgrounds and trace various sources of influence leading to the present and somewhat bewildering state of 'modern statistics', which perhaps to the confusion of many researchers working in the applied sciences is still filled with controversies and partly conflicting paradigms regarding even basic concepts.

1.1 Introduction

The aim of this book is to prepare for a synthesis of the two main traditions of statistical inference: those of the Bayesians and of the frequentists. Sir Ronald Aylmer Fisher worked out the theory of frequentist statistical inference from around 1920. From 1930 onward he developed his fiducial argument, which was intended to yield Bayesian-type results without the often ill-founded prior distributions needed in Bayesian analyses. Unfortunately, Fisher went wrong on the fiducial argument. We think, nevertheless, that it is a key to obtaining a synthesis of the two, partly competing statistical traditions.

Confidence, likelihood and probability are words used to characterise uncertainty in most everyday talk, and also in more formal contexts. The Intergovernmental Panel on Climate Change (IPCC), for example, concluded in 2007, "Most of the observed increase in global average temperature since the mid-20th century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations" (Summary for Policymakers, IPCC, 2007). They codify 'very likely' as having probability between 0.90 and 0.95 according to expert judgment. In its 2013 report IPCC is firmer and more precise in its conclusion. The Summary for Policymakers states, "It is *extremely likely* that more than half of the observed increase in global surface temperature from 1951 to 2010 was caused by the anthropogenic increase in greenhouse gas concentrations and other anthropogenic forcings together" (IPCC, 2013, p. 17). By *extremely likely* they mean more than 95% certainty.

We would have used 'confidence' rather than 'likelihood' to quantify degree of belief based on available data. We will use the term 'likelihood' in the technical sense usual in statistics.

Confidence, likelihood and probability are pivotal words in the science of statistics. Mathematical probability models are used to build likelihood functions that lead to confidence intervals. Why do we need three words, and actually additional words such as credibility and propensity, to measure uncertainty and frequency of chance events? The reason is that probability is used in very different contexts and to measure different things.

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

That an idealised coin has probability $\frac{1}{2}$ of showing heads when flipped means that in an imagined repeated trials experiment the frequency of heads will stabilise at $\frac{1}{2}$ in the long run. A person could also say that the probability is $\frac{1}{2}$ for a particular coin to show heads when flipped. This is a statement about the person's degree of belief. The first concept is that of frequentist or aleatory probability describing a chance setup, that is, an experiment or phenomenon in the real world. The second is that of a quantified degree of belief, which when based on knowledge, for example, that the coin is an ordinary Norwegian krone and is flipped properly, is called an epistemic probability. The knowledge behind an epistemic probability distribution is usually empirical data.

There are cases in which the frequentist concept of probability hardly makes sense. In historical contexts such as the evolution of the earth or the global economic development since the Industrial Revolution, the notion of a conceptual experiment that can be repeated is farfetched. Our human history cannot be repeated. But personal or even inter-subjective probability might apply. The statement from IPCC (2013) quoted earlier makes sense in that it reflects the prevailing degree of belief among climate scientists about whether the observed and expected climate changes are caused by human activity. Thus different types of probability stand in contrast: objective/subjective, aleatory/epistemic, frequentist/personal. The main distinction is between probability statements about the real world (e.g., 'a newborn child is a boy with probability 0.514') and statements about how certain a statement is (e.g., we are more than 95% confident that emission of greenhouse gasses from human activity is causing more than 50% of the observed warming of the earth from 1951 to 2010).

Fisher introduced the likelihood function and maximum likelihood estimation (Fisher, 1918, 1922). From being more or less synonymous with probability, likelihood now has a precise meaning in statistics distinct from probability. Statistical inference leading to confidence intervals or Bayesian posterior distributions is based on the likelihood function. The likelihood function, based on the probabilistic model of the data generating process, is actually a bridge between the data and the inferred results as they are expressed in confidence terms. Since the time of Laplace, statistical analysis followed the doctrine of inverse probability, which we now would call Bayesian analysis with flat prior distributions. This doctrine was challenged by Fisher (1930, p. 528):

> I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation. Yet that is quite exactly the position in respect of inverse probability. Bayes, who seems to have first attempted to apply the notion of probability, not only to effects in relation to their causes but also to causes in relation to their effects, invented a theory, and evidently doubted its soundness, for he did not publish it during his life. It was posthumously published by Price, who seems to have felt no doubt of its soundness. It and its applications must have made great headway during the next 20 years, for Laplace takes for granted in a highly generalised form what Bayes tentatively wished to postulate in a special case. [. . .] First, it is not to be lightly supposed that men of the mental calibre of Laplace and Gauss, not to mention later writers who have accepted their views, could fall into error on a question of prime theoretical importance, without an uncommonly good reason.

The "uncommonly good reason" was in Fisher's view that the Bayesian method was the only method around for formalised inductive reasoning under uncertainty. Fisher's paper was meant to present an alternative concept and methodology: fiducial probability. In this book we present *confidence inference*, which is what we make of Fisher's basic

idea, and its transformation and development through the work of Jerzy Neyman, Bradley Efron and others. It will in many cases be a practical alternative to Bayesian analysis. As Bayesian posterior distributions, confidence distributions capture inferential uncertainty about parameters, but without requiring prior distributions for the parameters of the model.

Fisher regarded the Bayesian use of flat priors as noninformative, as "fundamentally false and devoid of foundation". It is ironic that his fiducial method, which Neyman (1934) regarded "not less than a revolution in *the theory* of statistics", led to a controversy that lasted for some thirty years and ended with the fiducial method being put aside and nearly forgotten. The irony is that Fisher himself got it wrong after 1935, as will be explained in the text that follows.

Although also using the word credibility for measures of belief, the Bayesian speaks of probability or posterior probability. The Bayesian would use prior probability when assessing the parameters of the probabilistic model before new data are obtained, and posterior probability for their ex post assessment. She would agree with the Fisherian or frequentist statistician that the probability concept they both use when establishing the probabilistic model and the likelihood functions are meant to describe the real world, but she would insist on the model representing her personal view. This subjectivist Bayesian has thus only one concept of probability, and it is personal. When based on knowledge her subjective probability is epistemic. The Bayesian and Fisherian statistician would agree on the importance of a probabilistic model and a likelihood function as a bridge between data and inferred results, but the Bayesian would also carry her prior distributions over the bridge while the Fisherian will cross that bridge without any prior. When, however, a prior distribution is based on past observed data the Fisherian might, as explained in Chapter 10, add these prior data to his other data and obtain a combined likelihood function, with one component based on past data and another on the new data.

There are many Bayesians today (a lower bound for the number of different types of Bayesians is 46,656, according to Good [1983]), but few purely subjective ones. In most Bayesian analyses, whether performed by statisticians or subject matter scientists, prior distributions are necessary ingredients for carrying out an analysis through the often impressive Bayesian machinery. This is often done without representing prior uncertainties in a precise and argued way, however. We agree with Bayesians that modelling and analysis devoid of human judgement is impossible. The scientific issue to be investigated, data to be collected and model in which the analysis is carried out will all be chosen by the scientist, and will be influenced by personal judgements – to be made explicit and argued. But in much of science it is regarded as essential to keep personal views out of the analysis itself as much as possible, and to this end the methodology we present should be useful in medicine, physics, climatology, biology and elsewhere.

One can find many books on distributional inference in the Bayesian sense. There is, however, hardly a single book on distributional inference without prior probability distributions for the parameters of the model. The present book attempts to fill this gap by promoting what Hampel (2006) calls the original and correct fiducial argument (Fisher, 1930, 1973), as opposed to Fisher's later incorrect fiducial theory. The second decade of the second millennium is witnessing a renewed interest in fiducial analysis (see, e.g., Hannig [2009] and references therein) and in the related concept of confidence distribution (see e.g. the review and discussion paper Xie and Singh [2013]); see Section 1.9 for further

pointers and remarks. This book will hopefully help to foster further active interest in confidence distributions among theoretically inclined statisticians. More importantly, however, it is our hope that the book will contribute to improving statistical practice in the various empirical sciences.

Our context is that of empirical science based on quantitative data. In several sciences, such as medicine, Bayesian methodology has arguably not made a serious impact because it has been important to keep the researcher's personal belief out as much as possible. Prior distributions have been regarded with skepticism. Our inferred confidence distributions are free of such prior distributions. They will, however, often be approximate in the sense that their associated confidence intervals have coverage probabilities not matching their nominal levels with full precision. Also, confidence distributions might be computationally demanding. We demonstrate feasibility of confidence inference in examples and applications.

In addition to confidence inference being attractive in many cases, the idea of a confidence distribution represents a gold standard for epistemic probability distributions in science. A Bayesian posterior distribution carries more weight when it is shown to lead to credibility intervals that actually are confidence intervals. In this regard, Fraser (2011) argues that the Bayes posterior distribution risks being "quick and dirty confidence".

In what follows some philosophy of statistical inference is reviewed. In addition to the more philosophical reasons for distributional inference, with confidence distributions as the inferential results, we will illustrate their use in numerous examples. Some examples are theoretical while others are analyses of real data. We will also prove optimality theorems for confidence distributions. These are related to the Neyman–Pearson theory for testing and estimation.

Our hope is that Fisher's long neglected theory can be revived, and perhaps also that Efron (1998, p. 107) will be proven right when he says, "I believe that objective Bayes methods will develop for such problems, and that something like fiducial inference will play an important role in this development. Maybe Fisher's biggest blunder will become a big hit in the 21st century!"

1.2 Probability

Probability theory is a branch of mathematics along with geometry, topology, and so forth. Early probability theory dealt with games of chance. Here the basic probabilities were equally likely, and the challenge was to calculate the probabilities of various outcomes of the often rather complicated games. Games of chance were interesting in themselves, but for Pascal, Fermat and other eminent scientists they might have been used as test beds for ideas and mathematical arguments. Paccioli (1494) asked, "A and B are playing a fair game of balla. They agree to play until one has won six rounds. The game actually stops when A has won five and B three. How should the stakes be divided?" Bernstein (1996) argues that this opened the way for the study of the quantification of risk, which indeed was in demand in the early years of risky overseas trade.

Hacking (1975, 2006) is skeptical of the view that modern probability emerged and developed in response to the needs of merchants, insurance premiums, and so forth. His

view is that probability emerged around 1660 (actually, in 1662) as part of a general shift in the European mind-set, associated with the discontinuity between the Renaissance and the Enlightenment. The term probability (or close relatives, in different languages) had been used earlier, but not with its modern connotations; its old use was typically associated with 'approval', as in "in accordance with views held higher up". But from ca. 1660, with Pascal and Huygens and various other thinkers and writers, probability got its modern Janus face, of epistemic degree of belief and long-run frequency in repeated experiments.

Feller (1950, p. 1) argues that the mathematical discipline of probability has three distinct aspects: (1) the formal logical content, (2) the intuitive background and (3) the applications. "The character, and the charm, of the whole structure cannot be appreciated without considering all three aspects in their proper relations." The theory of probability is limited to one particular aspect of chance, and might be called "physical or *statistical probability*. In a rough way we may characterise this concept by saying that our probabilities do not refer to judgments but to possible outcomes of a *conceptual experiment*" (Feller, 1950, p. 4). By possible outcomes is meant not only the list or space of outcomes, but also the frequencies or probabilities of measurable events in the outcome space. Feller's notion of probability is called aleatory. It refers to statistical stability in the real world.

Feller wrote his book in an era that may be characterised as the heyday of frequentism. His understanding of probability was also that of Neyman and Wald, but Neyman's confidence level of an interval for a parameter of interest must be understood as the (degree of) confidence as equivalent to the fiducial probability of Fisher (1930), to be discussed in Section 1.6. This probability reflects the degree of belief a rational person would have in the true value of the parameter lying within the bounds of the interval. The measure of belief in the observed interval covering the truth is obtained from the fact that the method of calculating the interval would lead to success (i.e., the true parameter being inside the interval) with (aleatory) probability equal to the degree of confidence in (hypothetically) repeated applications. So, even Neyman used probability in the dual sense, both epistemic and aleatory, but he preferred the term 'confidence' for the epistemic variant.

Importantly, epistemic and aleatory probabilities differ in their mathematics, at least when the two are distinguished from each other and epistemic probability is understood as confidence. The formal structure of aleatory probability is an axiomatic branch of mathematics; the intuitive background that enables us to give physical or social meaning to statements about probabilities, and the subject matter applications of probability models, have all grown in the past 60 years. The probability used for modelling processes or phenomena in the social or natural world are mostly of the frequentist type. In the background there is a conceptual experiment. In repeated realisations of the experiment, the empirical frequency of the various possible events will stabilise in the long run. For a finite number of replicates, the law of addition and negation from percentage calculation applies to the frequencies. These laws are assumed also to apply to the probabilities, and they are in the axioms of Kolmogorov (1933) extended to so-called sigma additivity: $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ when the sets A_i are disjoint. This extension was needed to ensure continuity, but it comes at a price. If the outcome space is bigger than countable, such as the real line, not all sets in the outcome space are assigned a probability. These difficult sets are unmeasurable, while the family of measurable sets forms a sigma algebra.

The Kolmogorov axioms of probability theory describe what in general can be done with aleatory probabilities, while the theorems provide deduced results in the various cases. One such basic result is Bayes' lemma, which informs us how to update probabilities on observing that an event has occurred. To be explicit, suppose $A_1, \ldots, A_k$ is a list of nonoverlapping events whose union is certain (i.e., one of the A_j needs to occur), with probabilities $P(A_j)$. If we now learn that event B has occurred, what are then the updated probabilities for the A_j? These are called the conditional probabilities given that B has occurred, and are

$$\begin{aligned} P(A_j \mid B) &= \frac{P(A_j)P(B \mid A_j)}{P(B)} \\ &= \frac{P(A_j)P(B \mid A_j)}{P(A_1)P(B \mid A_1) + \cdots + P(A_k)P(B \mid A_k)} \end{aligned} \tag{1.1}$$

for $j = 1, \ldots, k$. The continuous version of this is discussed in Section 1.3.

Here $P(A \mid B)$ is the conditional probability of A given that the event B occurs. Conditional probability and Bayes' lemma are central elements of probability theory and are equally fundamental to Bayesian and non-Bayesian statisticians. Fisher assumed that his fiducial probabilities could be handled by ordinary probability calculus, just like for aleatory probabilities. This is, however, not the case, as we shall see. No axiomatic theory has been worked out for epistemic probability to be used in science, except for Bayesian probabilities, but these are problematic because they rely on prior probabilities.

1.3 Inverse probability

The basic problem of statistics is that of induction, that is, to learn about the state of the real system from what has been observed. When the observed data might have resulted from many different underlying states or causes, what is learned about the true state or parameter is uncertain. By the method of inverse probability, a distribution is obtained for the parameter characterising the system. The distribution expresses what has been learned from the data in view of the model and what the surrounding uncertainty is. Bayes' lemma is used repeatedly "to apply the notion of probability, not only to effects in relation to their causes but also to causes in relation to their effects" (Fisher, 1930, p. 528). It calculates the conditional probability of the cause A given the effect B from the direct conditional probability of the effect given the cause. This inversion lies at the root of Bayesian statistics. In its modern wrapping, we might speak of a parameter θ rather than cause A, and of data y rather than the effect B, with modelled distribution $f(y \mid \theta)$ rather than the direct probability. With a prior probability density $f_{\text{prior}}(\theta)$ on θ the inverse probability equation is

$$f_{\text{posterior}}(\theta) = f(y \mid \theta) f_{\text{prior}}(\theta) \Big/ \int f(y \mid \theta') f_{\text{prior}}(\theta')\, \mathrm{d}\theta'. \tag{1.2}$$

This is the continuous version of (1.1).

In the Bayesian paradigm there is no distinction between aleatory and epistemic probabilities, and Kolmogorov's axioms rule the common ground. *Inverse probability* is, in modern terminology, the Bayesian method with flat priors. The flatness of the prior

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

was meant to reflect lack of prior information. The posterior distribution (1.2) should then represent what was learned from the data without being influenced by previous knowledge, except what is embodied in the model $f(y,\theta)$.

The inverse probability method was dominant from the time of Laplace to around Fisher's 1930 publication. When discussing this paper, Neyman (1934, p. 619) hoped for an end to "the more than 150 years of disputation between the pros and cons of inverse probability that had left the subject only more befogged by doubt and frustration". One of the difficulties with the inverse probability method is that a flat prior on a parameter, say θ, is not flat on a curved transformation thereof, say $\tau = h(\theta)$; see Section 1.7 for more on this. But lack of information about θ is certainly lack of information about τ. Flatness of a prior therefore does not guarantee that it is noninformative. Fisher's response to this problem in inverse probability was to propose his fiducial method, which delivers epistemic probability distributions (fiducial distributions) entirely without invoking prior distributions.

1.4 Likelihood

As mentioned in our preface, Hald (1998, p. 1) opens his book on the history of mathematical statistics with the following words: "There are three revolutions in parametric statistical inference due to Laplace (1774), Gauss and Laplace in 1809–1812, and Fisher (1922)." The first revolution introduced the method of inverse probability, the second developed linear statistical methodology based on the normal distribution, while the third introduced the likelihood function as the workhorse of frequentist statistical inference.

Rather than regarding the modelled probability or probability density of the data as a function of the data y for given θ, $f(y\,|\,\theta)$, Fisher regarded it as a function of θ for given observed data $y = y_{\text{obs}}$, and called it the likelihood function:

$$L(\theta\,|\,y_{\text{obs}}) = f(y_{\text{obs}}\,|\,\theta).$$

The likelihood function is an essential element in the inverse probability (1.2); it is actually proportional to the posterior density because the prior density is flat. Unlike for prior distributions, flatness of the likelihood function does represent lack of information. The likelihood function is invariant to parameter transformations.

Fisher's original twist was to regard the likelihood function as a random variable. By substituting the random variable Y having the distribution $f(y\,|\,\theta)$ for its observed value y the random likelihood function $L(\theta\,|\,Y)$ emerges. By studying the properties of the random likelihood function Fisher developed a number of central concepts and results for statistical inference. One is the concept of a statistic, which is a function of the data such as the likelihood function. Another is that of a sufficient statistic. A sufficient statistic $S(Y)$ carries all the information in Y about the value of θ in the sense that the conditional distribution of Y given $S = s$ is independent of θ. There is thus no information left about the parameter when the sufficient statistic has been extracted from the data. The likelihood function is a sufficient statistic. The sufficiency property of the likelihood function constitutes the main reason for the *strong likelihood principle*: always base the statistical method on the likelihood function in parametric statistical models – and do not base the inference on anything else.

Birnbaum (1962) actually proved that the likelihood principle follows from the *sufficiency principle* – always base the statistical inference on sufficient statistics – and the

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

conditionality principle. The *conditionality principle* holds that when the data can be split into a sufficient component S and a remaining component A that has the same distribution for all values of the parameter θ behind the data, and is thus ancillary, then the statistical inference should be conditional on the observed value of $A = a$. That is, instead of carrying out the inference in the full model $f_\theta(y)$, it can equally well be carried out in the conditional model $f_\theta(y \mid a)$.

The strict conditionality principle, that inference should always be conditional on ancillary statistics when such exist, is controversial, however. One difficulty is that there might be more than one ancillary statistic, and on which of these should one condition? Sometimes there is a maximal ancillary statistic, which is the obvious candidate for conditioning. But even then there might be reasons for not conditioning.

The strong likelihood principle is also problematic in that it precludes statistical inference from also being based on the protocol for the experiment and how the inference would come out in hypothetical replications of the experiment. By the likelihood principle all relevant information is contained in the observed likelihood function, and such additional pieces of evidence as sampling distributions of estimators and test statistics are irrelevant. See Examples 1.2 and 3.6.

Example 1.1 Uniform data on unknown interval with known length

Assume that $Y_1, \ldots, Y_n$ are independent and uniformly distributed over the interval $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. In this model $A = Y_{(n)} - Y_{(1)}$ is the maximal ancillary, where $Y_{(1)} \leq \cdots \leq Y_{(n)}$ is the ordered sample and $S = (Y_{(n)} + Y_{(1)})/2$ is the sufficient statistic – A is ancillary because it has the same distribution regardless of θ, and S is sufficient because it holds all the information there is about θ in the model. When A is close to its maximal value 1, S is very informative on θ, while the opposite is the case when the sample spread is small. Should the inference then be conditional on A, say if an interval is sought for θ? Questions of this nature are discussed in Section 2.3. ■

Of the many methods based on the likelihood function, the likelihood ratio test and the maximum likelihood estimator might be the most widely used. In regular statistical models these methods have desirable statistical properties; see Chapter 2.

1.5 Frequentism

Is all information relevant for statistical inference indeed contained in the observed likelihood function? The frequentist view is no. In the case of Example 1.4, involving an exponential distribution, the mean lifelength in the sample of size $n = 82$ is $\widehat{\lambda} = 34.12$ years. The observed likelihood function is thus $L(\lambda) = \lambda^{-82} \exp(-2797.84/\lambda)$. Is that all there is to say about what has been learned? The parameter is supposed to characterise the population behind the sample. We could have obtained another sample from the same population. The frequentist view of Fisher and Neyman, which breaches the strong likelihood principle, is that the particular results obtained by a method applied to the observed sample must be understood against the background of the distribution of the results obtained by the same method in (hypothetically) repeated samples under the same conditions.

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

Example 1.2 Poisson or gamma?

In a Poisson process of constant rate λ, $X(t)$ is the number of events from the start at time 0 until time t. We shall look at two different models leading to identical likelihood functions when x events are observed up until time t, and suggest that the result should be interpreted against the background of which model is in force. In the first model, t is given and $X(t) = x$ is observed. The likelihood is then the Poisson,

$$L_1(\lambda \,|\, x) = \exp(-\lambda t)(\lambda t)^x / x! \propto \lambda^x \exp(-\lambda t)$$

(where $\propto$ denotes proportionality). The proportionality coefficient depends on only t and x and is of no consequence for the interpretation. In the other model, x is given and the waiting time until the xth event occurs is observed. This waiting time T_x is gamma distributed with shape parameter x and rate parameter λ. The likelihood comes out equivalent to that above,

$$L_2(\lambda \,|\, t) = \frac{\lambda^x}{\Gamma(x)} t^{x-1} e^{-\lambda t} \propto L_1(\lambda \,|\, x).$$

The maximum likelihood estimate is $\widehat{\lambda} = x/t$. When t is given, its mean in a long run of repeated samples is $\mathrm{E}\,\widehat{\lambda} = \mathrm{E}\,X(t)/t = \lambda$, while it is $\mathrm{E}\,(x/T_x) = \lambda x/(x-1)$ when the other model is in force and $x > 1$ is given. The frequentist takes note of how the data were obtained. His confidence distribution would depend on the model, despite the equivalence of the likelihood functions. See Figure 1.1 for $x = 5$ and $t = 10$. We return to this example and explain the two confidence curves in Example 3.4. ■

A confidence distribution provides confidence intervals by its quantiles. The two confidence distributions of Example 1.2 differ because confidence intervals depend on the

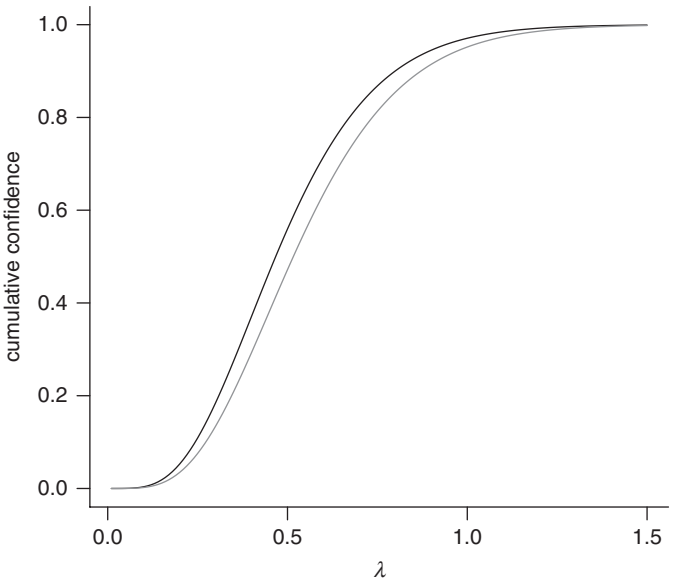

Figure 1.1 Confidence distribution for the rate parameter λ, for data with $x = 5$ and $t = 10$, for Example 1.2. The top curve is for the gamma experiment of observing $T(x) = t$, and the lower curve for the Poisson experiment of observing $X(t) = x$.

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

model. If based on the Poisson model they would not have correct coverage probabilities when the data really were obtained by the gamma experiment.

In the frequentist tradition, most forcefully formulated by J. Neyman and his school, the emphasis is on the performance in repeated use of the method on new data, and the question is whether the frequency of the results in this long hypothetical sequence agrees with the nominal requirement. For confidence distributions the requirement is that the cumulative confidence distribution function evaluated at the true value of the parameter is uniformly distributed.

A particular concern for frequentists is that of bias. There are various forms of bias in statistical studies. The data might be biased in that they are not representative of the target population. If the model badly represents the process that generated the data, the model might be said to be biased. But even when the data are a random sample from the population, and the model accurately represents the essentials of the data generating process, there might be unwanted biases in the results of the statistical analysis. An estimator might, for example, have a sampling distribution that is located away from the true value of the parameter, and a distribution proposed as a confidence distribution might in repeated use tend to be located away from its target. This can occur for many different reasons.

Example 1.3 Bias

Let Y_1, Y_2 be independent and normally distributed with parameters μ and σ. The likelihood $L(\mu,\sigma)$ factors as

$$\frac{1}{\sigma}\phi\Big(\frac{y_1-\mu}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{y_2-\mu}{\sigma}\Big) \propto \frac{1}{\sigma^2}\exp\Big\{-\Big(\frac{y_1-y_2}{2\sigma}\Big)^2\Big\}\exp\Big\{-\Big(\frac{\bar{y}-\mu}{\sigma}\Big)^2\Big\}.$$

For observed data $y_1 = -0.64$, $y_2 = 1.02$ the likelihood is contoured in the upper left panel of Figure 1.2. These data were simulated assuming $\mu = 0$, $\sigma = 1$. Observe that the likelihood is located in the σ direction at $\frac{1}{2}|y_1 - y_2|$, slightly to the left of the true value $\sigma = 1$. This is a property of the observed data. It is, however, a property frequently shared with other data simulated from the same model. Figure 1.2 shows the likelihoods for three additional realisations of (Y_1, Y_2). The top of the likelihood is actually to the left of the true value of σ with frequency $P\{|Y_1 - Y_2|/(2\sigma) < 1\} = 0.843$ in repeated samples. ■

Concern about possible bias inherent in a model and a method is difficult to conceptualise outside the frequentist paradigm. Bias is particularly difficult to discuss for Bayesian methods, and seems not to be a worry for most Bayesian statisticians.

1.6 Confidence and confidence curves

The word 'confidence' is used in everyday talk for degree of belief. We are confident that Norway has gained from staying outside the European Union, and we are confident that greenhouse gasses emitted to the atmosphere by human activity will cause a substantial increase in global surface temperature. These statements are meant to tell the receiver that we believe them to be true, and that we have a high degree of belief in their truth.

Confidence intervals, first discussed by Neyman (1934), are used routinely in science. The method is first to select a degree of confidence α. Then a lower confidence limit $L(Y)$

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

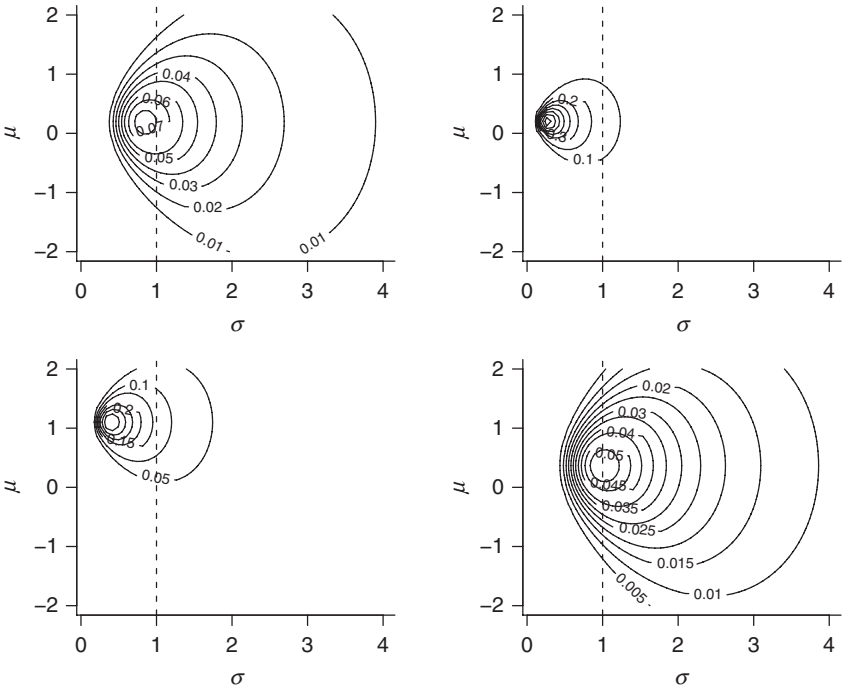

Figure 1.2 Contour plots of the likelihood function for four repeated normal samples of size two, for $\mu = 0, \sigma = 1$.

and an upper confidence limit $U(Y)$ are calculated from the data Y by a method ensuring that the true value of the parameter, the one behind the observed data, is covered by the stochastic interval $[L(Y), U(Y)]$ with probability α,

$$P_\theta\{L(Y) \leq \theta \text{ and } \theta \leq U(Y)\} = \alpha.$$

Thus if $\alpha = 0.95$, for example, then, in repeated use, the true value is in the long run covered in precisely 95% of the cases. The user will not know whether the actual case is among the frequent lucky cases had the experiment been repeated, with the realised interval correctly covering the parameter or whether it is an unlucky case where the interval is either entirely above or below the true parameter value. Since the probability of bad luck, $1 - \alpha$, is controlled, the reader is invited to attach degree of confidence α to the realised interval. We would say that the epistemic probability of the parameter being within the realised confidence interval $[L(y), U(y)]$ calculated from observed data y is α, and we use the word confidence for this objective epistemic probability.

The objectivity of the confidence derives from the transparency of the method of constructing the confidence interval. Anyone would come to the same interval for the given level of confidence when using the method. Confidence is, however, not a frequentist probability. The parameter is not viewed as the result of a random experiment. The confidence is rather the degree of belief of a rational person that the confidence interval covers the parameter. When the degree of confidence is 0.95, she will have as part of her knowledge that $L(y) \leq \theta \leq U(y)$ with (epistemic) probability 0.95.

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

Confidence intervals for scalar parameters, and more general confidence regions for p-dimensional parameters, are often calculated from the log-likelihood function. The log-likelihood function additively normalised to have its maximum equal to zero, and multiplied by -2, is called the *deviance function*, and is twice the log-likelihood ratio,

$$D(\theta) = 2\log\frac{L(\widehat{\theta})}{L(\theta)}. \tag{1.3}$$

The reason for the factor 2 is merely that the limit distribution of the deviance, which is guaranteed to exist under mild regularity conditions, cf. Section 2.4, is then a χ^2_1 (a chi-squared distribution with 1 degree of freedom, when the dimension of the parameter is 1) with cumulative distribution function Γ_1, rather than the slightly more cumbersome $\frac{1}{2}\chi^2_1$.

The confidence level is traditionally chosen say at 0.95, and a confidence region is obtained. For parameters of interest we suggest calculating confidence regions for all possible levels of confidence. These nested regions are the level sets of a curve called the *confidence curve*. When the confidence regions are found from the deviance function D by the chi-squared distribution, the confidence curve is $\mathrm{cc}(\theta) = \Gamma_1(D(\theta))$.

Example 1.4 The exponential model: Lifelength in ancient Egypt

How much have humans gained in lifelength over the past 2000 years? Karl Pearson asked this question in the first issue of his journal *Biometrika*, and used data on age at death as given by inscriptions on mummies from Roman era Egypt (Pearson, 1902, Spiegelberg, 1901). Figure 1.3 displays a histogram of the 82 male lifelengths in the data. Claeskens and Hjort (2008, pp. 33–35) compared nine different models for these and the accompanying female data with respect to fit. They found a Gompertz model to give the best fit according to the Akaike information Criterion (AIC); see Example 3.7. Although the exponential

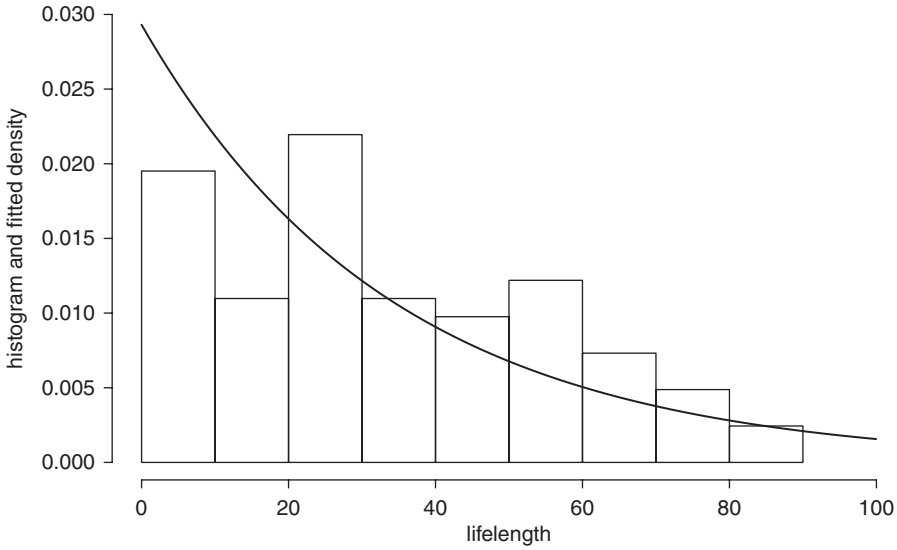

Figure 1.3 Histogram of age at death for 82 males in Roman era Egypt, along with the fitted exponential density curve. [Data source: Pearson (1902).]

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

distribution does not fit these data so well, for present purposes we proceed under the assumption that male Egyptians of the class for which the data are representative have a constant hazard $1/\lambda$ of dying throughout life, and thus have exponentially distributed lifelengths with probability density $f(t,\lambda) = (1/\lambda)\exp(-t/\lambda)$. Lifelength t is measured in years. The log-likelihood function of $n = 82$ independent observations $Y_1, \ldots, Y_n$ is

$$\ell_n(\lambda) = -n\log\lambda - n\bar{Y}/\lambda.$$

The maximum likelihood estimator is, as we know, $\widehat{\lambda} = \bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, which has a Gamma distribution with shape parameter n and scale parameter n/λ, and the estimate is 34.12 years. The deviance function is

$$D_n(\lambda) = n\Big(\frac{\widehat{\lambda}}{\lambda} - 1 - \log\frac{\widehat{\lambda}}{\lambda}\Big)$$

and is displayed in Figure 1.4. Below the deviance function the related confidence curve is displayed. This curve has confidence intervals as its level sets. To obtain the confidence curve we need the distribution of the deviance at the true value. According to the Wilks theorem (cf. Section 2.4), this is approximately the χ^2_1 distribution. The approximate confidence curve is thus

$$\mathrm{cc}(\lambda) = \Gamma_1(D(\lambda)).$$

The 95% confidence interval is shown as a horizontal line segment in both panels. Note that the confidence curve points at the point estimate. The right panels of the figure show

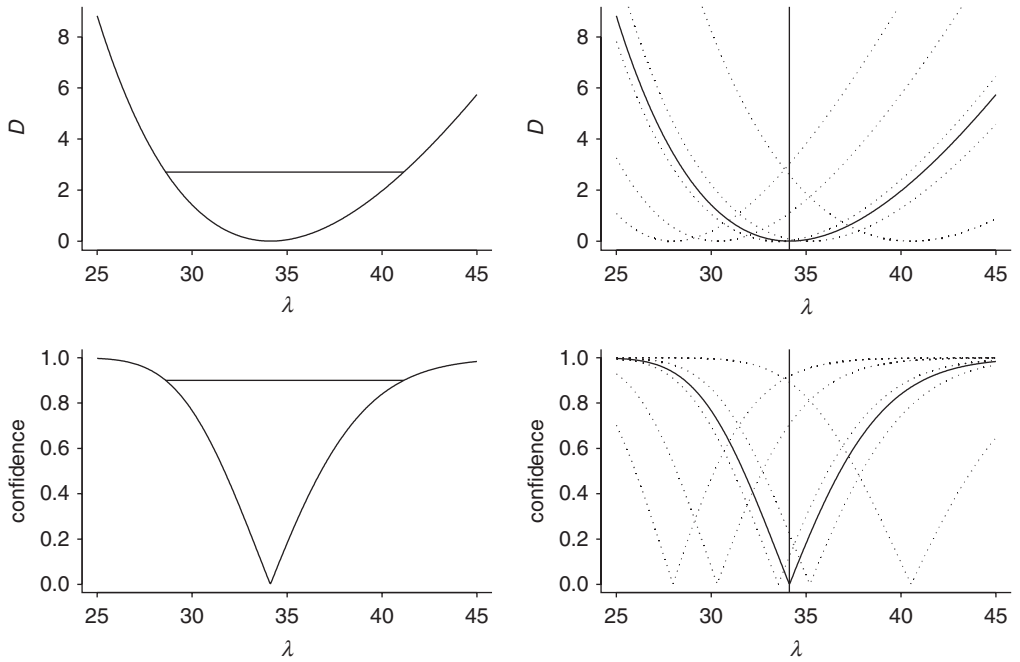

Figure 1.4 Expected lifelength for males in ancient Egypt, exponential model: Deviance function (upper left); confidence curve (lower left), both with a 95% confidence interval as horizontal bar. Right panels: Deviance function and confidence curve, augmented by five simulated replicas.

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

how the deviance function and the confidence curve vary over five replicated datasets. For each replicate $n = 82$ observations were drawn from the exponential distribution with mean parameter $\lambda = 34.12$. The crossings of the vertical line above this value are χ^2_1 distributed in height in the upper right panel, while they are uniformly distributed over the unit interval in the lower right panel.

For this illustration the exponential model is chosen for its simplicity of presentation. In Example 3.7 a more appropriate Gompertz model is used for these data, with a further variation used in Exercise 4.13. ■

The confidence curve is often just a probability transform of the deviance function. With $F_\theta(\cdot)$ the cumulative distribution function of the random deviance function evaluated at the true parameter value, say $D(\theta, Y)$,

$$\mathrm{cc}(\theta) = F_\theta(D(\theta, y_{\mathrm{obs}})) \tag{1.4}$$

is the confidence curve, obtained on the basis of the observed outcome y_{obs} of Y. Thus the random $\mathrm{cc}(\theta, Y)$ is uniformly distributed, when θ is the true value of the parameter, and its level sets are indeed confidence intervals.

The confidence curve will be a central concept in this book. It can be generalised to one-sided confidence intervals and to higher-dimensional parameters, where its contours provide a nested family of confidence regions indexed by degree of confidence. The confidence curve and its sister concept, the confidence distribution, may be obtained from the deviance function or from the directed likelihood discussed in Chapters 2, 3 and 7, or from sufficient statistics through pivotal constructions, considered in Chapters 3, 4 and later. Fisher (1930) used the pivotal method when he introduced the fiducial distribution; see Chapter 6. Briefly stated, the confidence distribution function is such that any confidence interval may be read off from its quantiles. Thus for the situation in Example 1.4 we may easily construct a confidence distribution $C(\lambda)$ from the confidence curve, such that the 95% confidence interval visible from the two left panels of Figure 1.4 may also be computed as $[C^{-1}(0.025), C^{-1}(0.975)]$.

1.7 Fiducial probability and confidence

The background to Fisher's seminal paper in 1930 was that the inverse probability method from Bayes and Laplace, with flat priors supposedly reflecting lack of prior knowledge, was still dominating as the formal method of statistical inference. Fisher referred to Boole, Venn and Chrystal and rejected the notion of flat priors representing ignorance. The problem is, as noted earlier, that flatness is not a property invariant under parameter transformations. If p has a uniform distribution over $(0, 1)$, for example, the density of the odds $\theta = p/(1-p)$ is the decreasing function $f(\theta) = (1+\theta)^{-2}$ on the positive halfline – and the log-odds $\log\{p/(1-p)\}$ has a logistic distribution with density $\exp(\gamma)/\{1+\exp(\gamma)\}^2$ unimodal and symmetric about zero. Lack of knowledge about p is equivalent to lack of knowledge about its odds, but if the former has a flat density the latter does not. This lack of invariance led Fisher to start his paper with the words quoted in Section 1.1. Fisher found, however, an operational definition of noninformativeness in his likelihood function. A flat likelihood is noninformative about the parameter. The likelihood function is also invariant, and it serves

https://doi.org/10.1017/CBO9781139046671.002 Published online by Cambridge University Press

Fisher's purpose. Fisher (1930) did not suggest his fiducial distribution as a transformation of the likelihood function. Instead he used pivots such as Student's t-statistic.

The t distribution was suggested by W. G. Gosset in 1908 (writing under his pseudonym 'Student'; the t distribution is also called the Student distribution) for the t-statistic behind the uncertainty interval for an unknown mean. Student (1908) did not manage to prove that the t-statistic indeed has the Student distribution when the data are normally distributed; instead, while still an undergraduate, Fisher was the first to prove this very important result. The result is really that when $Y_1, \ldots, Y_n$ are independent and normally distributed with expectation μ and standard deviation σ, then with the familiar $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $\widehat{\sigma}^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/(n-1)$,

$$T = \frac{\mu - \bar{Y}}{\widehat{\sigma}/\sqrt{n}}$$

has a fixed distribution regardless of the values of the interest parameter μ and the (in this context) nuisance parameter σ, and this distribution is the Student distribution with $n-1$ degrees of freedom. That T has a fixed distribution makes it a *pivotal statistic* (see Definition 2.3), and that the statistic is monotone in the parameter μ (we are assuming $\widehat{\sigma} > 0$, which happens with probability 1) makes it a monotonic pivot. With $F(t)$ the cumulative distribution function of the appropriate t distribution it leads to the equality

$$F(t) = P\left\{\frac{\mu - \bar{Y}}{\widehat{\sigma}/\sqrt{n}} \le t\right\} = P\{\mu \le \bar{Y} + \widehat{\sigma}\, t/\sqrt{n}\}.$$

The interval $(-\infty, \bar{Y} + \widehat{\sigma}\, t/\sqrt{n}]$ thus has probability $F(t)$ of covering the unknown parameter μ. For a given sample with observed mean $\bar{y}$ the realised interval $(-\infty, \bar{y} + \widehat{\sigma}\, t/\sqrt{n}]$ has in Fisher's terminology fiducial probability $F(t)$. By this method Fisher assigned fiducial probability to any interval for a scalar parameter when a pivot is available.

The function

$$C(\mu) = F\Big(\frac{\mu - \bar{y}}{\widehat{\sigma}/\sqrt{n}}\Big)$$

is increasing from 0 to 1 in the parameter μ, and is thus an ordinary cumulative distribution function. As such it represents a probability measure for μ. This is the fiducial probability. The $\frac{1}{2}\alpha$ and $1 - \frac{1}{2}\alpha$ quantiles of this distribution yield a confidence interval of confidence degree $1 - \alpha$,

$$[C^{-1}(\tfrac{1}{2}\alpha), C^{-1}(1 - \tfrac{1}{2}\alpha)].$$

The confidence interval is central in that it excludes one-sided confidence intervals of equal confidence at either end. It is tail symmetric as it misses the true value of μ with equal probability at both sides. Tail-asymmetric confidence intervals are also possible. Any interval (a, b) is assigned the fiducial probability $C(b) - C(a) = \beta - \alpha$, which then would have been the coverage probability of intervals obtained from the pivot by

$$P\{\bar{Y} + \widehat{\sigma}\, t_\alpha/\sqrt{n} \le \mu \le \bar{Y} + \widehat{\sigma}\, t_\beta/\sqrt{n}\}$$

where $\alpha = C(a)$ and $\beta = C(b)$. When, say, $\alpha = 0$ and a therefore is the extreme possible lowest value of the parameter, the interval is one sided. Neyman would have accepted $[a, b]$ as a confidence interval if it was constructed from given probabilities α and β. He was

interested in the coverage frequency of intervals constructed by the method. Fisher, on the other hand, was more interested in the logic of statistical inference. He asked what could be learned from the observed data and what the surrounding uncertainties are. Because the fiducial probabilities are simply a rearrangement of probability statements concerning the stochastic data, they are correct, understandable and acceptable to any rational person.

We shall look at many different exact and approximate pivots and the fiducial distributions generated by them. The term 'fiducial probability' has, however, fallen into disrepute as a result of the controversy over Fisher's method from 1935 until Fisher's death in 1962, and we prefer to use instead 'confidence distribution'. Neyman (1934) showed that his confidence intervals are found from fiducial distributions. He preferred 'confidence' over 'probability' to emphasise that the value of the parameter is a state of nature and not the result of a chance experiment. Neyman wanted probability to be understood in strictly frequentist terms. Our reason for preferring 'confidence' over 'fiducial probability' is to emphasise the relation to confidence intervals and, for higher dimensional parameters, confidence regions.

Fiducial probability "stands as Fisher's one great failure" according to Zabell (1992, p. 382) and has been characterised as "Fisher's biggest blunder". Hampel (2001, p. 5) writes that "fiducial probability has been grossly misunderstood by almost everybody, including Fisher himself". We explain fiducial distributions in Chapter 6 and discuss Fisher's claims for their properties and also their shortcomings as they were identified in the debate from 1935 until Fisher's death in 1962.

1.8 Why not go Bayesian?

Fisher (1930) revolted against the method of inverse probability. His main objection to this method of Bayesian analysis with flat prior densities was that flatness is not a property of noninformativity. Fisher was also uneasy about the meaning of the posterior distribution. Despite Fisher's revolt, seconded by Neyman and the vast majority of statisticians in the following 30 years, Bayesian methodology has survived and has become the dominating methodology in several fields of statistical analysis.

The Bayesian paradigm is attractive for several reasons. In his lecture upon receiving the Nobel Prize in economics (the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel, to be pedantically correct), C. Sims (2012) advocated Bayesian methods. His reasons might be summarised as follows.

1. Economists and policymakers like empirical results in the form of distributions, with uncertainty fully presented.
2. It is important to have a way to take graded a priori information into the analysis.
3. Parameters should be regarded as stochastic, having a prior distribution.
4. Posterior distributions have a clear meaning.
5. Importance sampling and Markov chain Monte Carlo (MCMC) simulations are effective computational methods for obtaining joint posterior distributions.
6. The posterior density for one parameter is obtained by simply taking the marginal.
7. Coherent learning: the old posterior is the prior for the new data.

8. The likelihood principle should be observed.
9. These gains are possible only within the Bayesian paradigm.

His reasons 1 and 5 might be the most important ones in applied fields. It is simply impressive what can be achieved by MCMC methods and other Bayesian calculations in complex models. Output in the form of distributions is indeed attractive to scientists and users of statistical information in most fields.

We agree regarding points 1, 2 and 5. In Chapter 10 we discuss the relationship between confidence distributions and likelihoods, particularly how a likelihood can be obtained from a confidence distribution. Coherent learning (point 7) is also important, but is, as we argue in Chapter 10, also achieved by Fisherian methods. Instead of updating the prior/posterior distribution, updating is done on the likelihood.

Fisher (1930) found the inverse probability method basically flawed. Why had it survived for 150 years, with proponents such as Laplace and Gauss? The reason, he thought, was that no alternative methodology existed that could provide inferential results, in the form of distributions, reflecting uncertainty. He put the fiducial distribution forward as an alternative to the Bayesian posterior distribution. Because it does not depend on a prior distribution it avoids the problem of the inverse probability method. The method of confidence distributions discussed in this book is closely related to the fiducial argument and will also serve as an alternative to the Bayesian method.

If no information is available for a parameter, except for its range, the confidence distribution for the parameter is calculated from the data in view of the model, but with no further input. If, however, a prior distribution, founded on data or expert judgment or even on personal beliefs, is to be taken into account, we suggest regarding it as data on the same footing as the other data, and converting it into a likelihood to be combined with the other likelihood components. See Chapter 10 for confidence likelihoods. Thus, point 2 can be achieved in our framework.

Should parameters be regarded as stochastic variables (point 3)? Often the value of a parameter is the outcome of a stochastic phenomenon. Even the gravitational constant is the result of the Big Bang, and could perhaps have come out differently. But when inference is sought for a parameter it is reasonable to condition on that underlying process and to regard the parameter as given, but surrounded by uncertainty for us. The inference has the twofold aim of reducing this uncertainty as much as possible, and of characterising the uncertainty accurately. In much of science, point 3 is thus not reasonable or relevant.

When knowledge or uncertainty can be expressed as a distribution, this distribution is epistemic in nature. A confidence distribution aspires to represent the knowledge, including the associated uncertainty, a rational mind would have when she agrees on the data, the model and the method. Confidence distributions are distributions of epistemic probabilities. Aleatory probabilities are different. They characterise randomness or chance variation in nature or society (Jeffreys, 1931). The probabilistic components of statistical models are cast in aleatory probability. The Bayesian has only one form of probability, and has no choice but to regard parameters as stochastic variables. The Fisherian objects use epistemic probability, that is, confidence, for uncertainty in knowledge, and aleatory probability for chance mechanisms. The parameters of a model are regarded as fixed values. These values are often called true values, to be learned about from data.

The Bayesian finds the posterior distribution for a derived parameter just by marginalisation. In practice this is done numerically, by simulating a huge sample from the joint posterior distribution. The posterior distribution, say for the first component θ_1 of θ, is then simply a smoothed histogram of the first column of the output from the MCMC run, or from another method used to sample from the posterior distribution. Marginalisation can go astray. When the model is nonlinear in a certain sense the marginal distribution might miss the target. The so-called length problem is an extreme example.

Example 1.5 The length problem

Let $Y_1, \ldots, Y_p$ be independent with $Y_i \sim \mathrm{N}(\mu_i, 1)$. In this model the prior of choice would traditionally be independent improper flat priors on each μ_i. The posterior distribution for μ_i is simply $\mathrm{N}(y_i, 1)$, and these are again independent. Let the parameter of interest be $\theta = \|\mu\|$, the length of the parameter vector. The marginal posterior distribution for θ^2 is the noncentral χ^2 distribution with p degrees of freedom and with $\|y\|^2$ as noncentrality parameter. Let F_{post} be the posterior cumulative distribution function obtained from this distribution. Stein (1959) considered this example, but in the context of joint fiducial distributions; see Section 6.3. He showed that $F_{\mathrm{post}}(\theta) \to 0$ as p increases. The posterior distribution must therefore be judged as biased, and the more so the higher the dimension. The Bayesian might not be happy with this example. Being pragmatic (Kass, 2011) she might call for a different prior distribution as the aim is θ. But is she really at liberty to choose different prior distributions for the same parameter vector according to what the focus parameter is?

In the length problem the Fisherian notes that $\widehat{\theta} = \|Y\| = (\sum_{i=1}^p Y_i^2)^{1/2}$ is the maximum likelihood estimator for θ. Now, $\widehat{\theta}^2$ is noncentrally χ^2 distributed with p degrees of freedom and parameter of noncentrality θ^2. Let this distribution have cumulative distribution function $\Gamma_p(\cdot, \theta^2)$. Since $C(\theta) = 1 - \Gamma_p(\widehat{\theta}^2, \theta^2)$ is a pivot increasing from 0 to 1, and with uniform pivotal distribution, it is a cumulative distribution function of a confidence distribution for θ; see Figure 1.5, which is based on simulated data with $p = 50$ and $\widehat{\theta} = 6.90$. By chance this estimate is only slightly larger than the true value $\theta = 5.40$. Our confidence distribution thus happens to be located a bit to the left of the true value. But the confidence distribution is unbiased in the sense that its median has exactly θ as its median in repeated samples. Note that the realised confidence distribution has a point mass of size 0.570 at $\theta = 0$. The true value could indeed easily have been zero with such a small estimate. ■

Bias is, as noted earlier, a serious concern in many applications. We strive to avoid bias in data. But as the length problem, and also the problems of the previous section illustrate, bias could also be intrinsic to the model, even for ideal data. Nonlinearity combined with limited data typically lead to the likelihood surface frequently and systematically being located away from the true value of the parameter in repeated samples. The Bayesian posterior distribution is then frequently misplaced relative to the parameter vector, perhaps more in some directions than in others. The Bayesian method in such cases provides biased results in the frequentist sense. Fraser (2011) asks whether Bayes posterior distributions are just quick and dirty confidence distributions. His main concern is bias of the type discussed here. We continue the discussion of bias in Section 9.4.

In addition to sharing the Bayesian ambition to produce statistical inference in the form of distributions representing knowledge with surrounding uncertainty, the Bayesian and Fisherian paradigms have many points of contact. Lindley (1958), for example, proved

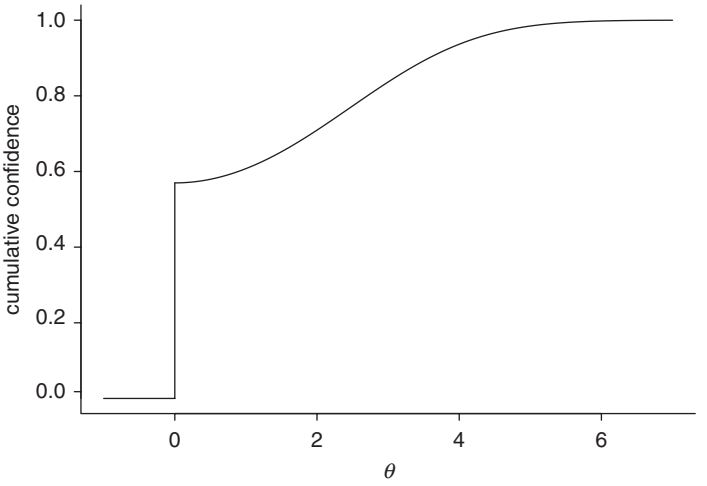

Figure 1.5 The confidence distribution function for $\theta = \|\mu\|$ in the length problem with $p = 50$ and $\widehat{\theta} = 6.9$, cf. Example 1.5. The true value was $\theta = 5.4$.

that the fiducial (confidence) distribution in one-parameter models is equal to the Bayesian posterior based on the so-called Jeffreys prior if and only if the model can be transformed to a location model. See Section 6.4, where we compare the Bayesian and the Fisherian methodologies in more detail.

It should also be noted that Bayesian posterior distributions are approximate confidence distributions. Bayesian machinery might thus be used by Fisherians, but whether the approximation is acceptable should be checked, perhaps by simulation, in each case.

1.9 Notes on the literature

Risk, chance and probability have fascinated humans at least since the Renaissance. Bernstein (1996) writes vividly about the remarkable story of risk, from Paccaoli's 1494 masterwork in which he discussed how to divide the stakes in an interrupted game of chance and through Kahneman and Tversky (1979, 1984), who studied risk behaviour from a psychological point of view. Kahneman (2011) has reached the general public with insights, experiences and theory related to our understanding of risk. Spiegelhalter (2008) and Spiegelhalter et al. (2011) are concerned with the individual's and the public's understanding of risk, also touching on the important differences between aleatory and epistemological uncertainties. Probability, as we know it today, is a distinctly modern concept. It was born around 1660, according to the philosophical study of probability, induction and statistical inference offered by Hacking (1975). In 1657 Huygens published his treatise, while Pascal and Fermat were developing the concept a bit earlier. Hacking also points to John Graunt's study of life tables from 1662 as the start of statistical inference. Hacking emphasises the Janus face character of probability: both being aleatory, that is, an objective property in the real world, and being epistemic, that is, the degree of belief a person has. Hampel (2006) follows up on this, and holds that the lack of distinction between aleatory and epistemic probabilities is a cause of much confusion.

Hald (1990, 1998) gives a rather comprehensive account of the history of mathematical statistics before 1750, and from then to 1930. The revolutionary events in this history are the introduction of inverse probability by Laplace in 1774, the linear statistical model based on the normal distribution by Gauss and Laplace in 1809–1812, and the likelihood function by Fisher in 1922. Hald notes that it took about twenty years and many papers for these authors to work out their ideas in detail, and it took some fifty years for the statistical community to accept and understand the new methods.

In his colourful history of statistics before 1900, Stigler (1986a) emphasises the difficulties involved in measuring uncertainty. He observes that regression analysis and least squares were introduced to the social sciences (by Francis Galton) some fifty years after it was accepted in the physical sciences (Gauss), and asks whether this is because concepts and theory are interwoven with model and empirical studies in a more profound way in the social sciences compared to physics and astronomy.

The Fisher–Neyman confidence methodology grew out of Fisher (1930) and Neyman (1934). Aldrich (2000) traces the roots of Fisher's seminal paper. In rereading Fisher, Efron (1998) puts Fisher's fiducial approach in the wider context of statistical inference, and finds it to be a most promising amalgamation of frequentist and Bayesian thinking. He uses the term 'confidence distribution', and suggests that this compromise between the two dominating methodologies, characterised as 'objective Bayesianism', might come to widespread use. Fisher got his fiducial theory wrong when extended to multiple dimensions. Efron says, "Maybe Fisher's biggest blunder will be a big hit in the 21st century." As did Neyman (1934), we would say that Fisher (1930) is a revolutionary paper. Hald thought there have been three revolutions in statistics. Perhaps Fisher, on line with Laplace, should be regarded as responsible for two revolutionary events – but partly due to Fisher himself, it might take nearly a century until the Fisher–Neyman confidence methodology, as conveyed in the present book, will be widely accepted.

The fiducial debate following Fisher (1935) is laid out by Zabell (1992); see also Efron (1998) and our Chapter 6. Cox (1958) discusses the fiducial approach and its relation to Neyman's confidence intervals. He suggests considering all the possible confidence intervals for a parameter, and to represent them in a distribution, that is, in a confidence distribution. Cox actually uses the phrase 'confidence distribution'. This is the first occurrence of this term we are aware of. Incidentally, Melville (1857) reminds us that 'con man' is also etymologically connoted with 'confidence'; a person (or a government) may exploit the gullibility of people by first gaining their confidence and then pulling off a confidence trick.

Fraser (1961a, 1968) investigated fiducial probability in the context of invariance. In his structural inference the pivot and its distribution, and thus the fiducial distribution, is found as the property of the maximal invariant in a transformation group. Hannig (2009) extends the fiducial argument to models defined by a structural equation $X = G(\theta, U)$ where U is a stochastic element with known distribution and X is the data. Hannig allows the structural equation to implicitly define a set-valued function $Q(X, U)$ into the parameter space, from which his fiducial distribution is obtained. When single-valued, Q would define an ordinary pivot. This is discussed further in Chapter 6. See also Hannig and Xie (2012), concerning attempts to make Dempster–Shafer rules for combining expert opinions amenable to handling of confidence distributions. The PhD dissertation of Salomé

(1998) presents theory for what Kardaun and Schaafsma in unpublished work (2003) call distributional inference, that is, inference leading to distributions for parameters of special interest; see in this connection also the "fourteen cryptic questions" formulated by D. R. Cox and then discussed at length in Kardaun et al. (2003). Certain strands of work related to confidence distributions include Schweder and Hjort (2002, 2003), Efron (1998), Singh et al. (2005, 2007) and Xie and Singh (2013); cf. also notes on the literature sections in later chapters.

Modern probability theory and hence statistics owe much to the treatise Kolmogorov (1933) (reprinted in e.g. Kolmogorov [1998], with more material than in the German original), where the axiomatic buildup is clarified, leading to sharp and relevant mathematical results. Bayesians need to take on additional elements; see, for example, Good (1983), Berger (1985) and Fraser (2011) for engaging discussions. That matters are delicate when strict philosophical arguments are called for is illustrated in one of the Good Thinking essays (Good, 1983), where the lower bound 56,656 is derived for the number of different types of Bayesians.

Invariably each sufficiently rich and vibrant scientific community contains certain cultural 'schools of thought', perhaps shaped and moulded by experiences or types of needs that might differ from one context to another. Thus a Bayesian's prior might mean a serious piece of work in one type of application but simply be an off-the-shelf tool in another, used because it does the job. Certain cultural-sociological divides are therefore visible in the fields of statistics, reminiscent of "the two cultures" discussed in Snow (1959, 1963); the number of identifiable cultures is later arguably extended to three (Kagan, 2009). In such a spirit Breiman (2001) identifies and examines two such statistical cultures. One is, roughly speaking, the emphasis on regression and classification, called for in a broad range of engineering applications, associated also with machine learning, support vector machines, neural networks, and so forth. The proof of the pudding is that the black box actually works well, and the job of any model parameter is to be fine-tuned inside associated algorithms for good performance. The complementary school concerns itself with more careful attention to building meaningful models, interpretation of parameters, identification of significant effects, hunting for causality, and so forth. Such needs are encountered in biology and medical research, economics and social research, psychology, climatology and physical sciences, and so forth, and in general when it comes to analyses of smaller and precious datasets and meta-analyses for combining information across studies. The key question to consider is "to explain or to predict" (Shmueli, 2010). We believe the inference tools associated with confidence distributions have genuine potential for both of these statistical schools.

As mentioned previously, fiducial probability has been regarded as "Fisher's one great failure" (Zabell, 1992). We agree with Fisher's critics that Fisher went wrong when he pushed his fiducial probability for vector parameters to be ordinary probability distributions. The fiducial distribution works fine for a parameter of dimension one, but may go wrong for higher-dimensional parameters. Fisher thought that his fiducial distribution, say in two dimensions, could be integrated to obtain marginal distributions, but as pointed out by several critics this can be done only in special cases; see, for example, Pitman (1957). To regard the fiducial distribution as an ordinary probability distribution over an imagined infinite population, as Fisher did, was also hard to swallow. Feller (1950) was not alone in

regarding probabilities as referring to outcomes of conceptual experiments. But a parameter such as the gravitational constant reflects an aspect of nature, and can hardly be thought of as an outcome in an experiment in a hypothetical sequence of repeated experiments. The fiducial argument and its ensuing controversy are discussed further in Chapter 6.

The basic problem with Fisher's position after 1935 is that he regarded his fiducial probability as an ordinary probability on par with other aleatory probabilities. We agree with Hampel (2006) that fiducial probability, or confidence as we prefer to call it, must be understood as an epistemic quantity. This epistemic probability is, however, objective because it is obtained from a clearly defined method that rational people should agree on. Thus Fisher (1930) succeeded in finding an alternative to probabilistic inference by the inverse probability method with flat priors, and when keeping to his original ideas and respecting the limitation with regard to how far the method reaches, a very potent statistical methodology is established. By relating the confidence distribution to confidence regions, as Neyman (1934) did, we agree with Efron (1998), and also with Fraser in his discussion of that article, that Fisher's fiducial method holds a key to "our profession's 250-years search for a dependable Bayes theory", despite the unfortunate interpretation and use of fiducial distributions that Fisher himself made after 1935.