In All Likelihood

Statistical Modelling and Inference Using Likelihood

Yudi Pawitan
University College Cork
National University of Ireland
Cork, Ireland
yudi@stat.ucc.ie

CLARENDON PRESS · OXFORD

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Cape Town
Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul Karachi
Kolkata Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw
with associated companies in Berlin Ibadan
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 2001

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
ISBN 0 19 850765 8
0 9 8 7 6 5 4 3 2 1
Typeset by Yudi Pawitan.
Printed in Great Britain
on acid-free paper by Biddles Ltd, Guildford and King's Lynn

Preface

Likelihood is the central concept in statistical modelling and inference. *In All Likelihood* covers the essential aspects of likelihood-based modelling as well as likelihood's fundamental role in inference. The title is a gentle reminder of the original meaning of 'likelihood' as a measure of uncertainty, a Fisherian view that tends to be forgotten under the weight of likelihood's more technical role.

Fisher coined the term 'likelihood' in 1921 to distinguish the method of maximum likelihood from the Bayesian or inverse probability argument. In the early days its application was fairly limited; few statistical techniques from the 1920s to 1950s could be called 'likelihood-based'. To see why, let us consider what we mean by 'statistical activities':

- *planning:* making decisions about the study design or sampling protocol, what measurements to take, stratification, sample size, etc.
- *describing:* summarizing the bulk of data in few quantities, finding or revealing meaningful patterns or trends, etc.
- *modelling:* developing mathematical models with few parameters to represent the patterns, or to explain the variability in terms of relationship between variables.
- *inference:* assessing whether we are seeing a real or spurious pattern or relationship, which typically involves an evaluation of the uncertainty in the parameter estimates.
- *model criticism:* assessing whether the model is sensible for the data. The most common form of model criticism is residual analysis.

A lot of early statistical works was focused on the first two activities, for which likelihood thinking does not make much contribution. Often the activity moved directly from description to inference with little modelling in between. Also, the early modelling scene was dominated by the normal-based linear models, so statisticians could survive with least-squares, and t tests or F tests (or rank tests if the data misbehaved).

The emergence of likelihood-based modelling had to wait for both the advent of computing power and the arrival of more challenging data and

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Cape Town
Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul Karachi
Kolkata Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw
with associated companies in Berlin Ibadan
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 2001

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
ISBN 0 19 850765 8
10 9 8 7 6 5 4 3 2 1
Typeset by Yudi Pawitan.
Printed in Great Britain
on acid-free paper by Biddles Ltd, Guildford and King's Lynn

Preface

Likelihood is the central concept in statistical modelling and inference. *In All Likelihood* covers the essential aspects of likelihood-based modelling as well as likelihood's fundamental role in inference. The title is a gentle reminder of the original meaning of 'likelihood' as a measure of uncertainty, a Fisherian view that tends to be forgotten under the weight of likelihood's more technical role.

Fisher coined the term 'likelihood' in 1921 to distinguish the method of maximum likelihood from the Bayesian or inverse probability argument. In the early days its application was fairly limited; few statistical techniques from the 1920s to 1950s could be called 'likelihood-based'. To see why, let us consider what we mean by 'statistical activities':

- *planning:* making decisions about the study design or sampling protocol, what measurements to take, stratification, sample size, etc.
- *describing:* summarizing the bulk of data in few quantities, finding or revealing meaningful patterns or trends, etc.
- *modelling:* developing mathematical models with few parameters to represent the patterns, or to explain the variability in terms of relationship between variables.
- *inference:* assessing whether we are seeing a real or spurious pattern or relationship, which typically involves an evaluation of the uncertainty in the parameter estimates.
- *model criticism:* assessing whether the model is sensible for the data. The most common form of model criticism is residual analysis.

A lot of early statistical works was focused on the first two activities, for which likelihood thinking does not make much contribution. Often the activity moved directly from description to inference with little modelling in between. Also, the early modelling scene was dominated by the normal-based linear models, so statisticians could survive with least-squares, and t tests or F tests (or rank tests if the data misbehaved).

The emergence of likelihood-based modelling had to wait for both the advent of computing power and the arrival of more challenging data analysis problems. These problems typically involve nonnormal outcome data, with possible complexities in their collection such as censoring, repeated

1

Introduction

Statistical modelling and inference have grown, above all else, to deal with variation and uncertainty. This may sound like an ambitious undertaking, since anyone going through life, even quietly, realizes ubiquitous uncertainties. It is not obvious that we can say something rigorous, scientific or even just sensible in the face of uncertainty.

Different schools of thought in statistics have emerged in reaction to uncertainty. In the Bayesian world all uncertainties can be modelled and processed through the standard rules of probability. Frequentism is more sceptical as it limits the type of uncertainty that can be studied statistically. Our focus is on the likelihood or Fisherian school, which offers a Bayesian–frequentist compromise. The purpose of this chapter is to discuss the background and motivation of these approaches to statistics.

1.1 Prototype of statistical problems

Consider the simplest nontrivial statistical problem, involving only *two* values. Recent studies show a significant number of drivers talk on their mobile phones while driving. Has there been an impact on accident rates? Suppose the number of traffic deaths increases from 170 last year to 190 this year. Numerically 190 is greater than 170, but it is not clear if the increase is 'real'. Suppose instead the number this year is 174, then in this case we feel intuitively that the change is not 'real'. If the number is 300 we feel more confident that it is a 'real' increase (although it is a totally different matter whether the increase can be attributed to mobile-phone use; see Redelmeier and Tibshirani (1997) for a report on the risk of car collision among drivers while using mobile phones).

Let us say that a change is 'significant' if we sense that it is a 'real' change. At the intuitive level, what is this sense of significance? It definitely responds to a numerical stimulus since we 'feel' 174 is different from 300. At which point do we change from being uncertain to being more confident? There is nothing in the basic laws of arithmetic or calculus that can supply us with a numerical answer to this problem. And for sure the answer cannot be found in the totality of the data itself (the two values in this case).

Uncertainty is pervasive in problems that deal with the real world, but statistics is the only branch of science that puts systematic effort into dealing with uncertainty. Statistics is suited to problems with inherent uncer-

but in many cases it merely quantifies it; uncertainty can remain even after an analysis is finished.

Aspirin data example

In a landmark study of the preventive benefits of low-dose aspirin for healthy individuals (Steering Committee of the Physicians' Health Study Research Group 1989), a total of 22,071 healthy physicians were randomized to either aspirin or placebo groups, and were followed for an average of 5 years. The number of heart attacks and strokes during follow-up are shown in Table 1.1.

Group	Heart attacks	Strokes	Total
Aspirin	139	119	11,037
Placebo	239	98	11,034
Total	378	217	22,071

Table 1.1: *The number of heart attacks and strokes during follow-up in the Physicians' Health Study.*

The main medical question is statistical: is aspirin beneficial? Obviously, there were fewer heart attacks in the aspirin group, 139 versus 239, but we face the same question: is the evidence strong enough so we can answer the question with confidence? The side effects, as measured by the number of strokes, were greater in the aspirin group, although 119 versus 98 are not as convincing as the benefit.

Suppose we express the benefit of aspirin as a relative risk of

$$\frac{139/11,037}{239/11,034} = 0.58.$$

A relative risk of one indicates aspirin is not beneficial, while a value much less than one indicates a benefit. Is 0.58 'far enough' from one? Answering such a question requires a *stochastic model* that describes the data we observe. In this example, we may model the number of heart attacks in the aspirin group as binomial with probability θ_1 and those in the placebo group as binomial with probability θ_2. Then the true relative risk is $\theta \equiv \theta_1/\theta_2$.

Let us denote the observed relative risk by $\widehat{\theta} = 0.58$. No uncertainty is associated with this number, so it fails to address the statistical nature of the original question. Does the trial contain information that $\widehat{\theta}$ is truly 'much' less than one? Now suppose the study is 10 times larger, so, assuming similar event rates, we observed 1390 versus 2390 heart attacks. Then $\widehat{\theta} = 0.58$ as before, but intuitively the information is now stronger. So, the data must have contained some measure of precision about $\widehat{\theta}$, from which

We can now state the basic problem of statistical inference: *how do we go from observed data to statements about the parameter of interest θ?*

1.2 Statistical problems and their models

Stochastic element

In a statistical problem there is an obvious *stochastic* or random element, which is not treated by the basic laws of arithmetic. In the traffic example, we intuitively accept that there are various contingencies or random events contributing to the number of deaths; in fact, we would be surprised if the two numbers were exactly the same. Thus statistical methods need stochastic models to deal with this aspect of the problem. The development of models and methods is the deductive or mathematical aspect of statistics.

While the mathematical manipulation of models is typically precise and potentially free from arguments, the choice of the model itself is, however, uncertain. This is important to keep in mind since the validity of most statistical analysis is conditional on the model being correct. It is a trade-off: we need some model to proceed with an analysis, especially with sparse data, but a wrong model can lead to a wrong conclusion.

Inductive process

Statistical problems are *inductive*: they deal with questions that arise as consequences of observing specific facts. The facts are usually the outcome of an experiment or a study. The questions are typically more general than the observations themselves; they ask for something not directly observed that is somehow logically contained in the observed data. We say we 'infer' something from the data. In the traffic deaths example, we want to compare the underlying accident/death rates after accounting for various contingencies that create randomness.

For deductive problems like mathematics, sometimes only parts of the available information are needed to establish a new theorem. In an inductive problem every piece of the data should be accounted for in reaching the main conclusion; ignoring parts of the data is generally not acceptable. An inductive problem that has some parallels with statistical inference is a court trial to establish the guilt or the innocence of a defendant. The witness's oath to tell 'the truth, the whole truth, and nothing but the truth' embodies the requirements of the inductive process.

In deductive problems the truth quality of the new theorem is the same as the quality of the 'data' (axioms, definitions and previous theorems) used in establishing it. In contrast, the degree of certainty in an inductive conclusion is typically stronger than the degree in the data constituent, and the truth quality of the conclusion improves as we use more and more data.

However, a single new item of information can destroy a carefully crafted conclusion; this aspect of inductive inference is ideal for mystery novels

Suppose we want to estimate the number of BSE- (Bovine Spongiform Encephalopathy, or 'mad-cow') infected cattle that entered the food chain in Ireland. This is not a trivial problem, but based on the observed number of BSE cases and some assumptions about the disease, we can estimate the number of infected animals slaughtered prior to showing symptoms. New but last minute information on exported cattle might invalidate a current estimate; further information that exported animals have a different age distribution from the animals for domestic consumption will also change the estimate.

Statistics plays an important role in science because all scientific endeavours are inductive, although many scientific questions are deterministic rather than stochastic. The emergence of statistical science is partly the result of the effort to make the inductive process rigorous. However, Lipton (1993), a philosopher of science, warns that

> inductive inference is about weighing evidence and judging likelihood, not definite proof.

The inductive process is inherently underdetermined: the input does not guarantee a unique solution, implying that even a correct induction is fallible.

Empirical or mechanistic models

The models used to deal with statistical problems can be either empirical or mechanistic. The latter is limited to applications where there is detailed knowledge regarding the underlying processes. For example, Newtonian laws in physics or Mendelian laws in genetics are mechanistic models. Here the exact relationships between the different quantities under observation are proposed mostly by some subject matter consideration rather than by looking at the data. A mechanistic model describes an underlying mechanism that explains the observed data.

Models in the applied sciences, such as medicine, epidemiology, psychology, climatology or agriculture, tend to be empirical. The analytical unit such as a human being or an area of land is usually too complex to be described by a scientific formula. If we model the number of deaths in the traffic example as having a Poisson distribution, we barely explain why we observe 170 rather than 100 deaths. Empirical models can be specified just by looking at the data without much subject matter consideration (this of course does not mean it is acceptable for a statistician to work on a desert island). The main requirement of an empirical model is that it explains the variability, rather than the underlying mechanism, in the observed data.

The separation between these two types of models is obviously not complete. There will be grey areas where some empirical evidence is used to help develop a mechanistic model, or a model may be composed of partly mechanistic and partly empirical submodels. The charge on the electron, for example, is an empirical quantity, but the (average) behaviour

In the 19th and early 20th centuries most experiments were performed in the basic sciences; hence scientific models then were mostly mechanistic. The rise of empirical modelling was a liberating influence. Now experiments can be performed in most applied sciences, or even 'worse': data can be collected from observational studies rather than controlled experiments. Most of the general models in statistics, such as classes of distributions and linear or nonlinear regression models, are empirical models. Thus the rise of statistical modelling coincides with empirical modelling.

While empirical models are widely applicable, we must recognize their limitations; see Example 1.1. A mechanistic model is more satisfying than an empirical model, but a current empirical model may be a future mechanistic model. In some areas of statistical applications, there may never be a mechanistic model; for example, there will never be a mechanistic model for the number of traffic accidents. The compromise is an empirical model with as much subject matter input as possible.

The role of models from a statistical point of view is discussed further in Lehmann (1990) and Cox (1990).

Example 1.1: A classic example of an empirical model is the 18th century Bode's geometric law of progression of the planetary distance d_k from the Sun. Good (1969) and Efron (1971) provided a statistical evaluation of the 'reality' of this law, which specifies

$$d_k = 4 + 3 \times 2^k,$$

where $k = -\infty, 0, 1, \ldots$ and d_k is scaled so that $d_1 = 10$ for Earth. With some 'juggling' the law fitted very well for the known planets at the time it was proposed (planets as far as Saturn can be seen by the naked eye). To get a better fit, Jupiter was shifted up to position $k = 4$, leaving a missing spot at $k = 3$ between Mars and Jupiter. After the law was proposed there was a search for the 'missing planet'. Uranus at $k = 6$ was discovered first at the predicted distance, hence strengthening the confidence in the law. The missing planet was never found; there is, however, a band of asteroids at approximately the predicted distance.

Planet	k	Bode's law	Observed distance	Fourth-degree polynomial
Mercury	$-\infty$	4	4.0	4.1
Venus	0	7	7.2	6.7
Earth	1	10	10	10.2
Mars	2	16	15.3	16.0
?	3	28	?	26.9
Jupiter	4	52	51.9	50.0
Saturn	5	100	95.5	97.0
Uranus (1781)	6	196	191.4	186.5
Neptune (1846)	7	388	300.0	312.8
Pluto (1930)	8	772	394.6	388.2

Even though the formula fits the data well (up to Uranus; see Figure 1.1), the question remains: is this a 'real' physical law? As it happened, the law did not fit Neptune or Pluto. A better fit to the data is given by a fourth-degree polynomial, but now it is clear that we cannot attach much mechanistic value to

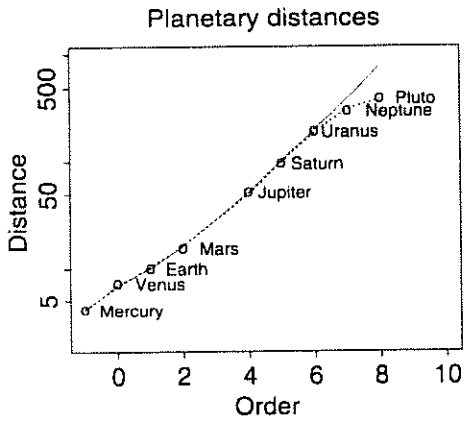

Figure 1.1: *Empirical model of planetary distances in terms of the order number from the Sun: Bode's law (solid) and a fourth-degree polynomial fit (dotted).*

1.3 Statistical uncertainty: inevitable controversies

> As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain they do not refer to reality. – Albert Einstein (1879–1955)

The characteristics discussed in the previous section, especially for empirical problems, militate to make statistical problems appear vague. Here it is useful to recognize two types of statistical uncertainty:

(i) *stochastic uncertainty:* this includes the uncertainty about a fixed parameter and a random outcome. This uncertainty is relatively easy to handle. Uncertainty about a fixed parameter, in principle, can always be reduced by performing a larger experiment. Many concepts in statistical inference deal with this uncertainty: sampling distribution, variability, confidence level, P-value, etc.

(ii) *inductive uncertainty:* owing to incomplete information, this uncertainty is more difficult to deal with, since we may be unable to quantify or control it.

Mathematically, we can view stochastic uncertainty as being conditional on an assumed model. Mathematics within the model can be precise and potentially within the control of the statistician. However, the choice of model itself carries an inductive uncertainty, which may be less precise and potentially beyond the control of the statistician.

The contrast between these two uncertainties is magnified when we are analysing a large dataset. Now the stochastic uncertainty becomes less important, while the inductive uncertainty is still very much there: Have

data? Have we considered and measured all the relevant variables? Are we asking the right questions? Given a set of data, depending on the way it was collected, there is usually an uncertainty about its variable definitions or meaning, wording and ordering of questions, representativeness of the sample, etc.

While it is possible to deal with stochastic uncertainty in an axiomatic way, it is doubtful that inductive uncertainty would ever yield to such an effort. It is important to recognize that, in statistical data analysis, inductive uncertainty is typically present in addition to the stochastic nature of the data itself. Due to the inductive process and the empirical nature of statistical problems, controversy is sometimes inevitable.

The traffic deaths example illustrates how controversies arise. If the number of deaths increases from 170 to 300, it would seem like a 'real' change and it would not be controversial to claim that the accident rate has increased, i.e. the uncertainty is small. But what if further scrutiny reveals one major traffic accident involving 25 cars and a large number of deaths, or an accident involving a bus where 40 people died? At this point we start thinking that, probably, a better way to look at the problem is by considering the number of accidents rather than deaths. Perhaps most accidents this year happened in the winter, whereas before they were distributed over the year. Possibly the number of younger drivers has increased, creating the need to split the data by age group. Splitting the data by years of driving experience may make more sense, but such a definition is only meaningful for drivers, while the death count also includes passengers and pedestrians!

This inductive process, which is very much a scientific process, raises two problems: one is that it tends to increase the stochastic uncertainty, since, by splitting the original observations into smaller explanatory groups, we are bound to compare smaller sets of numbers. The other problem is deciding where to stop in finding an explanation. There is no formal or precise answer to this question, so statisticians or scientists would have to deal with it on a case-by-case basis, often resorting to a judgement call. The closest guideline is to stop at a point where we have a reasonable control of uncertainty, deferring any decision on other factors of interest where too much uncertainty exists. Statisticians will have different experience, expertise, insight and prejudice, so from the same set of observations they might arrive at different conclusions. Beware! This is where we might find 'lies, damned lies and statistics'.

Pedagogic aspect

It is easier to learn, teach or describe methods that deal with stochastic uncertainty, and these have some chance of being mastered in a traditional academic or classroom setting. The unavoidable limitation of statistical texts is that they tend to concentrate on such methods. The joy and the pain of data analysis come as a reaction to uncertainties, so this discussion

the problem rather than of statistics, but even if we view it as such, the consequent difficulty in empirical model building and model selection is very much part of statistics and a statistician's life. This discussion also contains a warning that statisticians cannot work in a vacuum, since most of the relevant factors that create inductive uncertainties in a problem are subject matter specific.

1.4 The emergence of statistics

> It is impossible to calculate accurately events which are determined by chance. – Thucydides (*c.* 400BC)

There were two strands in the emergence of statistics. One was the development of the theory of probability, which had its original motivation in the calculation of expectation or uncertainties in gambling problems by Pascal (1623–1662) and Fermat (1601–1665). The theory was later developed on the mathematical side by Huygens (1629–1695), the Bernoulli brothers, in particular James Bernoulli (1654–1705), de Moivre (1667–1754) and Laplace (1749–1827), and on the logical side by Bayes (1701–1761), Boole (1815–1864) and Venn (1834–1923).

The growth of probability theory was an important milestone in the history of science. Fisher liked to comment that it was unknown to the Greek and the Islamic mathematicians (Thucydides was a historian); Persi Diaconis once declared that our brain is not wired to solve probability problems. With probability theory, for the first time since the birth of mathematics, we can make rigorous statements about uncertain events. The theory, however, is mostly deductive, which makes it a true branch of mathematics. Probability statements are evaluated as consequences of axioms or assumptions rather than specific observations. Statistics as the child of probability theory was born with the paper of Bayes in 1763 and was brought to maturity by Laplace.

The second strand in the emergence of statistics was an almost parallel development in the theory of errors. The main emphasis was not on the calculation of probabilities or uncertainties, but on summarizing observational data from astronomy or surveying. Gauss (1777–1855) was the main contributor in this area, notably with the principle of *least squares* as a general method of estimation. The important ingredient of this second line of development was the data-rich environment. In this connection Fisher noted the special role of Galton (1822-1911) in the birth of modern statistics towards the end of the 19th century. A compulsive data gatherer, Galton had a passionate conviction in the power of quantitative and statistical methods to deal with 'variable phenomena'.

Further progress in statistics continues to depend on data-rich environments. This was first supplied by experiments in agriculture and biometry, where Fisher was very much involved. Later applications include: industrial quality control, the military, engineering, psychology, business, medicine and health sciences. Other influences are found in data gathering

Bayesians and frequentists

The Bayesian and frequentist schools of statistics grew in response to problems of uncertainty, in particular to the way probability was viewed. The early writers in the 18th and 19th centuries considered it both a (subjective) degree of belief and (objective) long-run frequency. The 20th century brought a strong dichotomy. The frequentists limit probability to mean only a long-run frequency, while for the Bayesians it can carry the subjective notion of uncertainty.

This Bayesian–frequentist divide represents the fundamental tension between the need to say something relevant on a specific instance/dataset and the sense of objectivity in long-run frequencies. If we toss a coin, we have a sense of uncertainty about its outcome: we say the probability of heads is 0.5. Now, think about the *specific* next toss: can we say that our sense of uncertainty is 0.5, or is the number 0.5 meaningful only as a long-term average? Bayesians would accept both interpretations as being equally valid, but a true frequentist allows only the latter.

Since the two schools of thought generate different practical methodologies, the distinction is real and important. These disagreements do not hinder statistical applications, but they do indicate that the foundation of statistics is not settled. This tension also provides statistics with a fruitful dialectical process, at times injecting passion and emotion into a potentially dry subject. (Statisticians are probably unique among scientists with constant ponderings of the foundation of their subject; physicists are not expected to do that, though Einstein did argue with the quantum physicists about the role of quantum mechanics as the foundation of physics.)

Inverse probability: the Bayesians

The first modern method to assimilate observed data for quantitative inductive reasoning was published (posthumously) in 1763 by Bayes with his *Essay towards Solving a Problem in the Doctrine of Chances*. He used an inverse probability, via the now-standard Bayes theorem, to estimate a binomial probability. The simplest form of the Bayes theorem for two events A and B is

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}. \tag{1.1}$$

Suppose the unknown binomial probability is θ and the observed number of successes in n independent trials is x. Then, in modern notation, Bayes's solution is

$$f(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}, \tag{1.2}$$

where $f(\theta|x)$ is the conditional density of θ given x, $f(\theta)$ is the so-called prior density of θ and $f(x)$ is the marginal probability of x. Note that we have used the symbol $f(\cdot)$ as a generic function, much like the way we use

what the function is. Thus, $f(\theta, x)$ is the joint density of θ and x, $f(x|\theta)$ is the conditional density of x given θ, etc.)

Leaving aside the problem of specifying $f(\theta)$, Bayes had accomplished a giant step: he had put the problem of inductive inference (i.e. learning from data x) within the clean deductive steps of mathematics. Alas, 'the problem of specifying $f(\theta)$' a priori is an equally giant point of controversy up to the present day.

There is nothing controversial about the Bayes theorem (1.1), but (1.2) is a different matter. Both A and B in (1.1) are random events, while in the Bayesian use of (1.2) only x needs to be a random outcome; in a typical binomial experiment θ is an unknown fixed parameter. Bayes was well aware of this problem, which he overcame by considering that θ was generated in an *auxiliary physical experiment* – throwing a ball on a level square table – such that θ is expected to be uniform in the interval $(0, 1)$. Specifically, in this case we have $f(\theta) = 1$ and

$$f(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}}{\int_0^1 u^x(1-u)^{n-x}du}. \tag{1.3}$$

Fisher was very respectful of Bayes's seeming apprehension about using an axiomatic prior; in fact, he used Bayes's auxiliary experiment to indicate that Bayes was not a Bayesian in the modern sense. If θ is a random variable then there is nothing 'Bayesian' in the use of the Bayes theorem. Frequentists do use the Bayes theorem in applications that call for it.

Bayes did, however, write a *Scholium* (literally, a 'dissertation'; see Stigler 1982) immediately after his proposition:

> ... the same rule [i.e. formula (1.3) above] is a proper one to be used in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trial made concerning it.

In effect, he accepted the irresistible temptation to say that if we know nothing about θ then it is equally probable to be between zero and one. More significantly, he accepted that the uniform prior density, which now can be purely axiomatic, can be processed with the objective binomial probability to produce a posterior probability. So, after all, Bayes was a Bayesian, albeit a reluctant one. (In hindsight, probability was then the only available concept of uncertainty, so Bayes did not have any choice.)

Bayes's paper went largely unnoticed until Pearson (1920). It was Laplace, who, after independently discovering the Bayes theorem, developed Bayesian statistics as we understand it today. Boole's works on the probability theory (e.g. *Laws of Thought*, published in 1854), which discussed the Bayes theorem in the 'problem of causes', mentioned Laplace as the main reference. Laplace's *Théorie Analytique des Probabilités* was first published in 1812 and became the standard reference for the rest of the century. Laplace used the flat or uniform prior for all estimation problems, presented or justified as a reasonable expression of ignorance. The principle of the teaching of probability until the end of the 19th century. Fisher (1936) commented that that was how he learned inverse probability in school and 'for some years saw no reason to question its validity'.

Statistical works by Gauss and others in the 19th and early 20th centuries were largely Bayesian with the use of inverse probability arguments. Even Fisher, who later became one of the strongest critics of axiomatic Bayesianism, in his 1912 paper 'On an absolute criterion for fitting frequency curves', erroneously called his maximum likelihood the 'most probable set of values', suggesting inverse probability rather than likelihood, although it was already clear he had distinguished these two concepts.

Repeated sampling principle: the frequentists

A dominant section of statistics today views probability formally as a long-run frequency based on repeated experiments. This is the basis of the frequentist ideas and methods, where the truth of a mathematical model must be validated through an objective measure based on externally observable quantities. This feels natural, but as Shafer (1990) identified, 'the rise of frequentism' in probability came only in the mid-19th century from the writings of empiricist philosophers such as John Stuart Mill. Population counting and classification was also a factor in the empirical meaning of probability when it was used for modelling.

The *repeated sampling principle* specifies that procedures should be evaluated on the basis of repeat experimentation under the same conditions. The sampling distribution theory, which expresses the possible outcomes from the repeated experiments, is central to the frequentist methodology. Many concepts in use today, such as bias, variability and standard error of a statistic, P-value, type I error probability and power of a test, or confidence level, are based on the repeated sampling principle. The dominance of these concepts in applied statistics today proves the practical power of frequentist methods. Neyman (1894–1981) and Wald (1902–1950) were the most influential exponents of the frequentist philosophy. Fisher contributed enormously to the frequentist methodology, but did not subscribe fully to the philosophy.

True frequentism states that measures of uncertainties are to be interpreted only in a repeated sampling sense. In areas of statistical application, such as medical laboratory science or industrial quality control, where procedures are naturally repeated many times, the frequentist measures are very relevant.

The problem arises as the requirement of repeat experimentation is allowed to be hypothetical. There are many areas of science where experiments are unlikely to be repeated, for example in archaeology, economics, geology, astronomy, medicine, etc. A reliance on repeated sampling ideas can lead to logical paradoxes that appear in common rather than esoteric procedures.

Extreme frequentism among practical statisticians is probably quite

say $1.3 < \theta < 7.1$, either covers the parameter or it does not, we do not know which, and there is no way to express the uncertainty; the 95% applies only to the procedure, not to the particular interval. That is in fact the orthodox interpretation of the confidence interval. It neglects the evidence contained in a particular interval/dataset, because measures of uncertainty are only interpreted in hypothetical repetitions.

Most scientists would probably interpret the confidence interval intuitively in a subjective/Bayesian way: there is a 95% probability the interval contains the true parameter, i.e. the value 95% has some evidential attachment to the observed interval.

Bayesians versus frequentists

> A great truth is a truth whose opposite is also a great truth. – Thomas Mann (1875–1955)

In Bayesian computations one starts by explicitly postulating that a parameter θ has a distribution with prior density $f(\theta)$; for example, in a problem to estimate a probability θ, one might assume it is uniformly distributed on (0,1). The distinguishing attitude here is that, since θ does not have to be a random outcome of an experiment, this prior can be specified axiomatically, based on thinking alone. This is the methodological starting point that separates the Bayesians from the frequentists, as the latter cannot accept that a parameter can have a distribution, since such a distribution does not have an external reality. Bayesians would say there is an uncertainty about θ and insist any uncertainty be expressed probabilistically. The distribution of θ is interpreted in a subjective way as a degree of belief.

Once one accepts the prior $f(\theta)$ for θ and agrees it can be treated as a regular density, the way to proceed is purely deductive and (internally) consistent. Assuming that, given θ, our data x follows a statistical model $p_\theta(x) = f(x|\theta)$, then the information about θ contained in the data is given by the *posterior* density, using the Bayes theorem as in (1.2),

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}.$$

In Bayesian thinking there is no operational difference between a prior density $f(\theta)$, which measures belief, and $f(x|\theta)$, which measures an observable quantity. These two things are conceptually equal as measures of uncertainty, and they can be mixed using the Bayes theorem.

The posterior density $f(\theta|x)$, in principle, captures from the data all the information that is relevant for θ. Hence, it is an update of the prior $f(\theta)$. In a sequence of experiments it is clear that the current posterior can function as a future prior, so the Bayesian method has a natural way of accumulating information.

When forced, most frequentists would probably admit that a degree of

can assume a density, since frequentists could also think of $f(\theta)$ as a prior likelihood (the likelihood of the parameter before we have any data). Two genuine concerns exist:

(i) the practical problem of choosing an appropriate prior. Leaving aside the problem of subjective interpretation, there is an ongoing controversy on how we should pick $f(\theta)$. Several early writers such as Boole (1854, pages 384, 392) and Venn (1876) had criticized the arbitrariness in the axiomatic choice of $f(\theta)$; Fisher was also explicit in his rejection of any axiomatic prior, although *he did not rule out* that some applications, such as genetics, may have physically meaningful $f(\theta)$. Modern Bayesians seem to converge toward the so-called 'objective priors' (e.g. Gatsonis *et al.* 1997), but there are many shades of Bayesianism (Berger 2000).

(ii) the 'rules of engagement' regarding a subjective degree of belief. There is nothing really debatable about how one feels, and there is nothing wrong in thinking of probability in a subjective way. However, one's formal action based on such feeling is open to genuine disagreement. Treating a subjective probability density like a regular density function means, for example, it can be integrated out, and it needs a Jacobian term when transformed to a different scale. The latter creates a lack of invariance in the choice of prior: seeming ignorance in one scale becomes information in another scale (see Section 2.8).

Efron (1998) compares the psychological differences between the two schools of thought. A comparative study highlights the strengths and weaknesses of each approach. The strength of the Bayesian school is its unified approach to all problems of uncertainty. Such unity provides clarity, especially in complex problems, though it does not mean Bayesian solutions are practical. In fact, until recently Bayesians could not solve complex problems because of computational difficulties (Efron 1986a). While, bound by fewer rules, the strength of a frequentist solution is usually its practicality.

Example 1.2: A new eye drug was tested against an old one on 10 subjects. The two drugs were randomly assigned to both eyes of each person. In all cases the new drug performed better than the old drug. The P-value from the observed data is $2^{-10} = 0.001$, showing that what we observe is not likely due to chance alone, or that it is very likely the new drug is better than the old one. □

Such simplicity is difficult to beat. Given that a physical randomization was actually used, very little extra assumption is needed to produce a valid conclusion. And the final conclusion, that the new drug is better than the old one, might be all we need to know from the experiment. The achieved simplicity is a reward of focus: we are only interested in knowing if chance alone could have produced the observed data. In real studies, of course, we might want to know more about the biological mechanism or possible side

The advent of cheap computer power and Monte Carlo techniques (e.g. Gilks *et al.* 1995) have largely dismantled the Bayesian computational wall. Complex problems are now routinely solved using the Bayesian methodology. In fact, being pragmatic, one can separate the Bayesian numerical methods from the underlying philosophy, and use them as a means of obtaining likelihood functions. This is a recent trend, for example, in molecular genetics. In Section 10.6 we will see that the Bayesian and likelihood computations have close numerical connections.

Luckily, in large-sample problems, frequentist and Bayesian computations tend to produce similar numerical results, since in this case the data dominate the prior density and the level of uncertainty is small. In small- to medium-sized samples, the two approaches may not coincide, though in real data analysis the difference is usually of smaller order of magnitude than the inductive uncertainty in the data and in the model selection.

The following 'exchange paradox', discussed in detail by Christensen and Utts (1992), illustrates how our handling of uncertainty affects our logical thinking. To grasp the story quickly, or to entertain others with it, replace x by 100.

Example 1.3: A swami puts an unknown amount of money in one envelope and twice that amount in another. He asks you to pick one envelope at random, open it and then decide if you would exchange it with the other envelope. You pick one (randomly), open it and see the outcome $X = x$ dollars. You reason that, suppose Y is the content of the other envelope, then Y is either $x/2$ or $2x$ with probability 0.5; if you exchange it you are going to get $(x/2+2x)/2 = 5x/4$, which is bigger than your current x. 'With a gleam in your eye', you would exchange the envelope, wouldn't you?

The reasoning holds for any value of x, which means that you actually *do not need to open the envelope* in the first place, and you would still want to exchange it! Furthermore, when you get the second envelope, the same reasoning applies again, so you should exchange it back. A discussion of the Bayesian and fequentist aspects of this paradox is left as an exercise. □

1.5 Fisher and the third way

The likelihood approach offers a distinct 'third way', a Bayesian-frequentist compromise. We might call it Fisherian as it owes most of its conceptual development to Fisher (1890–1962). Fisher was clearly against the use of the axiomatic prior probability fundamental to the Bayesians, but he was equally emphatic in his rejection of long-run frequency as the only way to interpret probability. Fisher was a frequentist in his insistence that statistical inference should be objectively verifiable; however, his advocacy of likelihood inference in cases where probability-based inference is not available puts him closer to the Bayesian school.

In a stimulating paper on Fisher's legacies, Efron (1998) created a statistical triangle with Fisherian, Bayesian and frequentist nodes. He then placed various statistical techniques within the triangle to indicate their flavour.

Fisher's effort for an objective inference without any use of prior probability led him to the idea of *fiducial probability* (Fisher 1930, 1934). This concept prompted the confidence interval procedure (Neyman 1935). It appears that Fisher never managed to convince others what fiducial probability was, despite his insistence that, conceptually, it is 'entirely identical with the classical probability of the early writers' (Fisher 1973, page 54). In some models the fiducial probability coincides with the usual frequentist/long-run-frequency probability. The problems occur in more complex models where exact probability statements are not possible.

From his last book *Statistical Methods and Scientific Inference* (1973, in particular Chapter III) it is clear that Fisher settled with the idea that

- whenever possible to get exact results we should base inference on probability statements, otherwise it should be based on the likelihood;
- the likelihood can be interpreted subjectively as a rational degree of belief, but it is weaker than probability, since it does not allow an external verification, and
- in large samples there is a strengthening of likelihood statements where it becomes possible to attach some probabilistic properties ('asymptotic approach to a higher status' – Fisher 1973, page 78).

These seem to summarize the Fisherian view. (While Fisher's probability was fiducial probability, let us take him at his own words that it is 'entirely identical with the classical probability'.) About 40 years elapsed between the explicit definition of the likelihood for the purpose of estimation and Fisher's final judgement about likelihood inference. The distinguishing view is that *inference is possible directly from the likelihood function*; this is neither Bayesian nor frequentist, and in fact both schools would reject such a view as they allow only probability-based inference.

These Fisherian views also differ from the so-called 'pure likelihood view' that considers the likelihood as the sole carrier of uncertainty in statistical inference (e.g. Royall 1997, although he would call it 'evidence' rather than 'uncertainty'). Fisher recognized two 'well-defined levels of logical status' for uncertainty about parameters, one supplied by probability and the other by likelihood. A likelihood-based inference is used to 'analyze, summarize and communicate statistical evidence of types too weak to supply true probability statements' (Fisher 1973, page 75). Furthermore, when available, a probability statement must allow for an external verification (a verification by observable quantities), so it is clear that frequentist consideration is also an important aspect of the Fisherian view.

Fisher's requirement for an exact probability inference is more stringent than the so-called 'exact inference' in statistics today (Fisher 1973, pages 69–70). His prototype of an exact probability-based inference is the confidence interval for the normal mean (even though the term 'confidence interval' is Neyman's). The statement

is unambiguous and exactly/objectively verifiable; it is an ideal form of inference. However, the so-called 'exact 95% confidence interval' for the binomial proportion (see Section 5.8) in fact does not have exactly 95% coverage probability, so logically it is of lower status than the exact interval for the normal model. It is for this situation the likelihood is indicated.

For Fisher, both likelihood and probability are measures of uncertainty, but they are on a different footing. This is a non-Bayesian view, since for Bayesians all uncertainty is measured with probability. The subjective element in the interpretation of likelihood, however, is akin to a Bayesian/non-frequentist attitude. It is worth noting that, when backed up with large-sample theory to supply probability statements, the mechanics and numerical results of likelihood inference are generally acceptable to frequentist statisticians. So, in their psychology, Fisherians are braver than the frequentists in saying that inference is possible from the likelihood function alone, but not as brave as the Bayesians to admit an axiomatic prior into the argument.

Legacies

By 1920 the field of statistics must have been a confusing place. Yates (1990) wrote that it was the age of correlation and coefficients of all kinds. To assess association in 2×2 tables there were the coefficient of association, coefficient of mean square contingency, coefficient of tetrachoric correlation, coefficient of mean square contingency, coefficient of tetrachoric correlation, equiprobable tetrachoric correlation, and coefficient of colligation, but the idea of estimating the association and its test of significance were mixed up. There were many techniques available, such as the least squares principle, the method of moments, the inverse probability method, the χ^2 test, the normal distribution, Pearson's system of curves, the central limit theorem, etc., but there was no firm logical foundation.

The level of confusion is typified by the title of Edgeworth's paper in 1908 and Pearson's editorial in *Biometrika* in 1913: '*On the probable errors of frequency constants*', which in modern terminology would be 'the standard error of fixed parameters'. There was simply no logical distinction or available terms for a parameter and its estimate. On the mathematical side, the χ^2 test of association for the 2×2 tables had 3 degrees of freedom!

A more serious source of theoretical confusion seems to be the implicit use of inverse probability arguments in many early statistical works, no doubt the influence of Laplace. The role of the prior distribution in inverse probability arguments was never seriously questioned until early 20th century. When explicitly stated, the arbitrariness of the prior specification was probably a stumbling block to a proper appreciation of statistical questions as objective questions. Boole (1854) wrote in the *Laws of Thoughts* (Chapter XX, page 384) that such arbitrariness

> seems [to imply, that definite solution is impossible, and to mark the point where inquiry ought to stop.

weakness, but did not see any alternative; he considered the question of inductive inference as

> second to none other in the Theory of Probabilities in importance, [I hope it] will receive the careful attention which it deserves.

In his works on the theory of errors, Gauss was also aware of the problem, but he got around it by justifying his method of estimation in terms of the least-squares principle; this principle is still central in most standard introductions to regression models, which is unfortunate, since (i) in itself it is devoid of inferential content and (ii) it is not natural for general probability models, so it creates an unnecessary conceptual gap with the far richer class of generalized linear models.

Fisher answered Boole's challenge by clearly identifying the likelihood as the key inferential quantity that is free of subjective prior probabilities. He stressed that if, prior to the data, we know absolutely nothing about a parameter (recall Bayes's *Scholium* above) then all of the information from the data is in the likelihood. In the same subjective way the Bayesians interpret probability, the likelihood provides a 'rational degree of belief' or an 'order of preferences' on possible parameter values; the fundamental difference is that ***the likelihood does not obey probability laws.*** So probability and likelihood are different concepts available to deal with different levels of uncertainty.

There were earlier writers, such as Daniel Bernoulli or Venn, who had used or mentioned the idea of maximum likelihood in rudimentary forms (see Edwards 1992, Appendix 2). It usually appeared under the name of 'most probable value', indicating the influence of inverse probability argument. Even Fisher in 1912 used that name, even though it was clear from the discussion he had likelihood in mind. The confusion was only cleared in 1921 when Fisher invented the term 'likelihood'.

In a series of the most influential papers in statistics Fisher (in particular in 1922 and 1925) introduced order into the chaos by identifying and naming the fundamental concepts such as 'parameter', 'statistic', 'variance', 'sufficiency', 'consistency', 'information', and 'estimation','maximum likelihood estimate', 'efficiency' and 'optimality'. He was the first to use Greek letters for unknown parameters and Latin letters for the estimates. He set up the agenda for statistical research by identifying and formulating the important questions.

He 'fixed' the degree of freedom of the χ^2 test for the 2×2 tables in 1922. He recognized the paper by 'Student' in 1908 on the t-test, which was ignored by the large-sample-based statistical world at the time, as a milestone in the history of statistics: it was the first exact test. He emphasized the importance of inference based on exact distribution and identified 'the problem of distribution' as a respectable branch of theoretical statistics. Fisher was unsurpassed in this area, being the first to derive the exact distribution of the t and F statistics, as well as that of the sample

'isher's influence went beyond the foundation of statistics and the like-
od methods. His *Statistical Methods for Research Workers*, first pub-
d in 1925, brought the new ideas to generations of practical research
kers. Fisher practically invented the field of experimental design, in-
ucing the fundamental ideas of randomization, replication, blocking,
orial experiments, etc., and its analysis of variance. His *Design of Ex-
ments*, first published in 1935, emphasized the importance of carefully
cted data to simplify subsequent analysis and to arrive at unambigu-
conclusions. He contributed significantly to areas of sampling distribu-
theory, regression analysis, extreme value theory, nonparametric and
ivariate analysis. In a careful study of Fisher's legacy, Savage (1976)
mented that it would be a lot faster to list areas in statistics where
er did *not* contribute fundamentally, for example sequential analysis
time series modelling.
Outside statistics, many geneticists consider Fisher as the most impor-
evolutionary biologist after Darwin. In 1930 Fisher was the first to
ide a key synthesis of Mendelian genetics and Darwin's theory of evo-
n, thus giving a quantitative basis for the latter. Fisher was never a
essor of statistics: he was Galton Professor of Eugenics at University
ege London, then Balfour Professor of Genetics at Cambridge Univer-

For a statistician, his writings can be inspirational as they are full of
viction on the fundamental role and contributions of statistical methods
cience and in 'refinement of human reasoning'. Fisher (1952) believed

> Statistical Science was the peculiar aspect of human progress which gave to the twentieth century its special character. ... it is to the statistician that the present age turns for what is most essential in all its more important activities.

'important activities' include the experimental programmes, the ob-
vational surveys, the quality control engineering, etc. He identified the
ial contribution of statistical ideas to the fundamental scientific ad-
ces of the 19th century such as in Lyell's *Principles of Geology* and
win's theory of evolution.
It is an unfortunate turn of history that Fisher's articles and books are
onger standard reading in the study of statistics. Fisher was often crit-
ed for being obscure or hard to read. Savage (1976), however, reported
t his statistical mentors, which included Milton Friedman and W. Allen
llis, gave the advice: 'To become a statistician, practice statistics and
ll Fisher over with patience, respect and scepticism'. Savage closed his
0 Fisher Memorial Lecture with 'I do hope that you won't let a week
by without reading a little bit of Fisher'.
Fisher's publications were collected in the five-volume *Collected Papers
R.A. Fisher*, edited by Bennett and Cornish (1974). His biography,
itled *R.A. Fisher, The Life of a Scientist*, was published by his daughter

of his works were written by Barnard (1963), Bartlett (1965), Yates and Mather (1963), Kendall (1963), Neyman (1961, 1967), Pearson (1974) and Savage (1976). Recent articles include Aldrich (1997), Efron (1998) and Hald (1999). Edwards's (1992) book on likelihood was largely influenced by Fisher and the Appendices contain useful accounts of the history of likelihood and Fisher's key contributions. Fienberg and Hinkley (1980) contains a wide-ranging discussion of Fisher's papers and his impact on statistics.

1.6 Exercises

Exercise 1.1: Discuss the stochastic and inductive uncertainty in the following statements:

(a) A study shows that children of mothers who smoke have lower IQs than those of non-smoking mothers.

(b) A report by Interpol in 1994 shows a rate of (about) 55 crimes per 1000 people in the USA, compared to 100 in the UK and 125 in Sweden. ('Small' note: the newspaper that published the report later published a letter by an official from the local Swedish Embassy saying that, in Sweden, if a swindler defrauds 1000 people the case would be recorded as 1000 crimes.)

(c) Life expectancy in Indonesia is currently 64 years for women and 60 years for men. (To which generation do these numbers apply?)

(d) The current unemployment rate in Ireland is 4.7%. (What does 'unemployed' mean?)

(e) The total fertility rate for women in Kenya is 4.1 babies.

(f) The population of Cairo is around 16 million people. (Varies by a few million between night and day.)

(g) The national clinical trial of aspirin, conducted on about 22,000 healthy male physicians, established the benefit of taking aspirin. (To what population does the result apply?)

Exercise 1.2: What is wrong with the reasoning in the exchange paradox in Example 1.3? Discuss the Bayesian and frequentist aspects of the paradox, first assuming the 'game' is only played once, then assuming it is played repeatedly.