4

Comparison of Alternative Approaches to Inference

4.1 A STRUCTURE FOR ALTERNATIVE APPROACHES

It would be misleading to dichotomise statistical methods as either 'classical' or 'Bayesian', since both terms cover a bewildering range of techniques. A rough taxonomy can be developed by distinguishing two characteristics: whether or not prior distributions are used for inferences, and whether the objective is estimation, hypothesis testing or a decision requiring a loss function of some form. All six combinations of these elements have been investigated in theory and, to some extent, in practice, and Table 4.1 assigns a label to each possible combination.

This categorisation can be made finer still, and in Section 3.20 an attempt was made to delineate the different schools of Bayesianism that exist. Empirical Bayes techniques can be considered as essentially Fisherian since there is no formal introduction of prior opinion, while reference Bayesian methods, based on attempts at 'objective' priors, fall somewhat between the Fisherian and proper Bayesian approaches. We acknowledge that many of the examples in

Table 4.1 A taxonomy of six possible 'philosophical' approaches to statistical inference, depending on the objective and the formal quantitative use of prior information.

		Objective		
		Inference (estimation)	*Hypothesis testing*	*Decision (loss function)*
Use of prior evidence	*Informal*	Fisherian	Neyman–Pearson	Classical decision theory
	Formal	Proper Bayesian	'Bayes factors'	Full decision-theoretic Bayesian

Bayesian Approaches to Clinical Trials and Health-Care Evaluation D. J. Spiegelhalter, K. R. Abrams and J. P. Myles
 ISBN: 0-471-49975-7

this book do not use informative prior distributions, and their results could be (approximately) obtained by a likelihood analysis.

With so many options the resulting arguments about their relative merits inevitably become somewhat complex, and in this chapter we can only highlight some major issues. The standard approach in the evaluation of medical interventions is a mixture of Fisherian and Neyman–Pearson philosophies and is briefly summarised in Section 4.2, although Neyman–Pearson ideas have attracted particularly strong criticism from both Fisherian and Bayesian perspectives (Section 4.3). *P*-values are critically compared with Bayes factors in Section 4.4.

In the midst of often polemical arguments, it has also been argued that it would be 'a great pity if differences of technical approach were exaggerated into differences about qualitative issues' (Cox and Farewell, 1997), while Armitage (1993) maintains it is not appropriate to polarise the argument as a choice between extremes. It also appears reasonable to suggest that the appropriate approach may depend crucially on context (Section 3.1): for example, both Koch (1991) and Whitehead (1993) claim that a proper Bayesian approach may be reasonable at early stages of a drug's development but is not acceptable in phase III trials.

4.2 CONVENTIONAL STATISTICAL METHODS USED IN HEALTH-CARE EVALUATION

Conventional approaches to inference can be divided into the two broad schools of Fisherian and Neyman–Pearson.

The *Fisherian* approach regarding inference on an unknown intervention effect θ is based on the likelihood function (Section 2.2.4), which expresses the relative support given to the different values of θ by the data. This gives rise to a maximum likelihood estimate comprising the most supported value for θ, and intervals based on ranges of values of θ with most likelihood. More controversially, Fisher suggested summarising the evidence against specified null hypotheses by *P*-values (the chance of getting a result as extreme as that observed were the null hypothesis true), although this was only intended as an informal guide to the strength of evidence in the specific experiment being reported (Goodman, 1999a). Hill *et al.* (2000) provide a good historical background, emphasising that the likelihood alone could be used for comparing hypotheses without calculation of *P*-values.

The *Neyman–Pearson* approach has a different perspective, rooted in an attempt at a theory of 'inductive behaviour', in seeking procedures for hypothesis testing and estimation that satisfy certain properties in long-run repeated use. Specifically, it focuses on the chances of making various types of error when making decisions on the basis of the data so that, for example, clinical trials are traditionally designed to have a fixed Type I error α (the chance of incorrectly rejecting the null hypothesis), usually taken as 5% or 1%, and fixed power (one minus the Type II error β, the chance of not detecting the alternative hypoth-

esis), often 80% or 90%. Similarly, formulae for 95% confidence intervals are designed so that, in 95% of situations in which they are appropriately used, they will contain the true parameter value. The problem, as discussed in detail by Goodman (1999a), is that this restricts us in what we can say about the specific experiment being analysed.

In practice, a *combined* approach has developed, which is perhaps ironic in view of the enmity between the initial protagonists of the approaches (see below). Senn (1997b) points out that clinical trials are generally designed from a Neyman–Pearson standpoint, but analysed from a Fisherian perspective using *P*-values as measures of evidence. Methods used for observational methods and evidence synthesis tend to be more Fisherian, but Goodman (1999a) argues that the most common form of statistical analysis is to use *P*-values but, inappropriately, to interpret them as saying something about long-run properties.

Advantages of the conventional framework include its apparent separation of the evidence in the data from subjective factors, the general ease in computation, its wide acceptability and established criteria for 'significance', its relevance to the drug regulatory framework in which quality control of statistical submissions must be ensured, the availability of software, and the existence of robust non- and semi-parametric procedures.

Nevertheless, there has been continual criticism of these traditional approaches since their introduction in the 1920s and 1930s, and their development has been marked by considerable animosity and vituperative argument. When Neyman (1934) presented his theory of confidence intervals at a meeting of the Royal Statistical Society, Arthur Bowley, a strong advocate of the method of 'inverse probability' (the Bayesian approach), was given the task of proposing the vote of thanks. Towards the end of his remarks he said: 'I am not at all sure that the ''confidence'' is not a ''confidence trick''. He then went on to suggest a Bayesian approach was necessary: 'Does that really take us any further? . . . Does it really lead us towards what we need – the chance that in the universe which we are sampling the proportion is within . . . certain limits? I think it does not'. Fisher opened the discussion of Neyman (1935) on the attack: 'Were it not for the persistent efforts which Dr Neyman and Dr Pearson had made to treat what they speak of as problems of estimation, by means merely of tests of significance, he had no doubt that Dr Neyman would not have been in any danger of falling into the series of misunderstandings which his paper revealed'. Egon Pearson then came to Neyman's defence, saying that 'while he knew there was a widespread belief in Professor Fisher's infallibility, he must, in the first place, beg leave to question the wisdom of accusing a fellow-worker of incompetence without, at the same time, showing that he had succeeded in mastering the argument'.

In a strong attack on traditional methods, Cornfield (1976) claims that 'the paradox is that a solid structure of permanent value has, nevertheless, emerged, lacking only the firm logical foundation on which it was originally thought to have been built'. Generic criticisms include the failure of traditional methods to

incorporate formally the inevitable background information that is available both at design and analysis, that they take no account of the consequences of the conclusions, and, from a more ideological perspective, that they disobey certain reasonable axioms of rational behaviour (Section 3.1). In addition, there is no doubt that classical inferences are often misinterpreted, in that *P*-values are mistaken for probabilities of null hypotheses being true, and 95% confidence intervals as meaning there is a 95% chance of their containing the true value. Our personal opinion is that the strongest argument against Neyman–Pearson methods and *P*-values is their disobedience of the likelihood principle: this crucial idea is now discussed within the context of sequential analysis.

4.3 THE LIKELIHOOD PRINCIPLE, SEQUENTIAL ANALYSIS AND TYPES OF ERROR

4.3.1 The likelihood principle

This principle (Berger and Wolpert, 1988) states that all the information that the data provide about the parameter is contained in the likelihood: we have already seen in Sections 3.2 and 3.3 how data only influence the *relative* plausibility of an alternative hypothesis through the relative likelihood and hence Bayesian inference automatically obeys this principle. This simple idea, however, has very strong consequences, as the following classic example demonstrates.

Example 4.1 *Stopping: The likelihood principle in action*

Goodman (1999a) considers the following classic problem. Suppose we hear that six people have each been given treatments *A* and *B*, and asked which they prefer. Five preferred *A*, and one preferred *B*. What evidence is this against the null hypothesis that *A* and *B* are preferred equally in the population?

Let θ be the true unknown proportion in the population preferring *A*, with $\theta = 0.5$ corresponding to the null hypothesis of 'no preference'. Then the likelihood arising from the experiment is proportional to $\theta^5(1-\theta)$ (Section 2.2.4) and the likelihood principle states that all the evidence about θ to be derived from this experiment can be extracted from this function, using either likelihood or Bayesian methods.

In contrast, let us consider the *P*-value: the probability of observing a result at least as extreme as the data, given the null hypothesis $H_0: \theta = 0.5$. But what results are 'at least as extreme'? Suppose we are told that the experimenter decided in advance that six people were to be included, and the first five preferred *A* and the final one preferred *B*. The possible results of the experiment and their probabilities under H_0 are shown in

Table 4.2 under 'Design 1', with the 'at least as extreme as observed' outcomes highlighted in bold: these probabilities come from the binomial (0.5,6) distribution (Section 2.6.1). It is not clear how to handle the probability of the observation itself when defining what is 'as extreme' – here we adopt the standard convention of including half its probability so that the one-sided *P*-value is $\frac{1}{2}(6/64) + 1/64 = 0.0625$, with a two-sided *P*-value of 0.13; note that Goodman (1999*a*) considers the one-sided *P*-value including the whole contribution from the observed data, leading to $P = 0.11$. We may be disappointed that the result is not 'significant' at $P < 0.05$.

Table 4.2 Two different experimental designs: (1) ask six subjects whether they prefer *A* or *B*; (2) ask subjects sequentially until one prefers *B* and then stop. Observed data comprise 5 preferences for *A* and one for *B*. Highlighted values indicate potential data 'at least as extreme' as that observed under the null hypothesis H_0 of no overall preference in the population, i.e. the probability of either preference is 0.5.

Design 1		Design 2	
$Y_1 =$ No. subjects preferring *A*	Probability under H_0	$Y_2 =$ First subject preferring *B*	Probability under H_0
0	1/64	1	1/2
1	6/64	2	1/4
2	15/64	3	1/8
3	20/64	4	1/16
4	15/64	5	1/32
5	**6/64**	**6**	**1/64**
6	**1/64**	**7**	**1/128**
		8	**1/256**
		etc.	etc.

But then we hear that a mistake has been made in reporting the results, and that the experimenter in fact used a different (and admittedly rather strange) sampling procedure (Design 2): he had decided to carry on experimenting until he found someone who preferred *B*, and then stop. Table 4.2 again shows the possible results with those 'at least as extreme as observed' highlighted: the probabilities follow a 'geometric' distribution in which the chance of first getting a *B* preference on the *n*th trial is $1/2^n$. This time the *P*-value is $\frac{1}{2}(1/64) + 1/128 + 1/256 + \ldots = \frac{1}{2}(1/64) + 1/64 = 3/128 = 0.023$, with a two-sided *P*-value of 0.046, and we might now be delighted that it is 'significant' at $P < 0.05$.

A likelihood and Bayesian approach to this problem is described in Section 4.4.4.

In Example 4.1 the intention of the experimenter dictated the conclusions to be drawn from the results, and the *P*-values depended on what would have happened had something else been observed (Berry, 1987). The likelihood principle claims such behaviour is nonsensical, since only the observed data influence the conclusions and this is through the likelihood alone.

4.3.2 Sequential analysis

In a sequential experimental design the data are periodically analysed and the study stopped if sufficiently convincing results obtained. Such repeated analysis of the data can have a strong effect on the overall Type I error in the experiment, since there are many opportunities to obtain a false positive result. The traditional approach to sequential analysis identifies classes of 'stopping boundaries' with fixed overall Type I error α, and then chooses designs with minimum Type II error β (maximum power) for particular alternative hypotheses. At the end of a study *P*-values and confidence intervals should be adjusted for the sequential nature of the design (Whitehead, 1997a).

Sequential data fall naturally within the Bayesian framework, as the posterior distribution following each observation becomes the prior for the next (Section 3.12). As forcefully argued by Cornfield (1976), (3.25) shows that the evidence for taking alternative decisions depends only on the relative likelihood of alternative hypotheses (the Bayes factor), prior probabilities, and utilities, and hence provides a direct decision-theoretic justification for the likelihood principle within sequential trials. Sequential analysis therefore provides a primary focus for disagreement between frequentist and Bayesian approaches, since the likelihood principle means that concern about frequentist stopping rules retaining Type I error is entirely misplaced, and we can analyse trials at will. Criticism has been forceful: Anscombe (1963) baldly states that 'Sequential analysis is a hoax', and (1975) considers that 'provided the investigator has faithfully presented his methods and all of his results, it seems hard indeed to accept the notion that *I* should be influenced in my judgement by how frequently *he* peeked at the data while he was collecting it'.

We find the following argument particularly persuasive. If we were to assign weights to the relative importance of the two types of error that could be made, any resulting design would seek to minimise a linear combination of the Type I error rate α and Type II error rate β. Perhaps surprisingly, such a design would obey the likelihood principle, and this led Cornfield (1966) to point out that

> the entire basis for sequential analysis depends upon nothing more profound than a preference for minimising β for given α rather than minimising their linear combination. Rarely has so mighty a structure, and one so surprising to scientific common sense, rested on so frail a distinction and so delicate a preference.

We shall return to this topic when discussing sequential clinical trials in Section 6.6.

4.3.3 Type I and Type II error

Neyman–Pearson theory has been strongly criticised from both a Bayesian and Fisherian perspective. Anscombe (1963) says 'the concept of error probabilities of the first and second kinds . . . has no direct relevance to experimentation . . . The formalism of opinions, decisions concerning further experimentation and other required actions, are not dictated in a simple prearranged way by the formal analysis of the experiment, but call for judgement and imagination'.

The selection of values for error rates in trials seems particularly arbitrary: Healy (1994) asks 'Why the invariable 5% for α? Conditional on this, why the larger 10% or even 20% for β? Is it really more important not to make a fool of yourself than it is to discover something new?' Sheiner (1991) provides a strong polemic against hypothesis testing and in favour of an approach in which 'we gather data to model and quantify nature'; shifting attention from hypothesis testing to confidence intervals does not really avoid the problem, since these are, essentially, just the set of hypotheses that cannot be rejected at a certain α level.

We have already identified the crucial issue that arises in any context in which simultaneous analysis of multiple studies, or multiple analyses of the same study, is required. The traditional approach warns that repeated hypothesis testing is bound to raise the chance of a Type I error (incorrectly rejecting a true null hypothesis), and so suggests some adjustment, such as Bonferroni, to try to retain a specified overall Type I error. This will typically give larger *P*-values and wider confidence intervals.

The problem lies in deciding the set in which to embed the particular analysis being carried out. Cornfield (1976) asks, with some irony: 'Do we want error control over a single trial, over all the independent trials on the same agent, on the same disease, over the lifetime of an investigator, etc.?' The need for any such adjustment, which necessarily depends on the number of hypotheses being tested, has been strongly questioned even from a non-Bayesian perspective, particularly in epidemiology; Cole (1979) states that 'in every study, every association should be evaluated on its own merits: its prior credibility and its features in the study at hand. The number of other variables is irrelevant'. Greenland and Robins (1991) are among the many who have argued that some adjustment *is* necessary, but rather than being based on Type I errors, it should be derived from an explicit model that reflects assumptions about variability, and hence leads naturally to the approach to multiplicity outlined in Section 3.17.

4.4 *P*-VALUES AND BAYES FACTORS*

4.4.1 Criticism of *P*-values

We noted in Section 4.3 that sequential trials present a particular problem for *P*-values. Other arguments against this procedure are that the null hypothesis

may be neither plausible nor of great interest, the arbitrariness of the 0.05 and 0.01 level, and that *P*-values tend to create a false dichotomy between 'significant' and 'non-significant' which is inappropriate for consequent policy decisions. Furthermore, the definition of 'more extreme' and hence the value of *P* itself may be unclear even in some simple circumstances, such as testing association in a 2×2 table of counts, as well as requiring the choice between one- or two-sided tests.

The strongest criticism is, perhaps, that *P*-values focus on statistical rather than practical significance and hence their interpretation can be very dependent on sample size. This is illustrated in Example 4.2.

Example 4.2 *Preference: P-values as measures of evidence*

Freeman (1993) considered four hypothetical studies in which equal number of patients are given treatments *A* and *B* and asked which they prefer, with results shown in Table 4.3. Each results in an identical 'significant' two-sided *P*-value of 0.04. However, as Freeman states, the first trial

Table 4.3 Four theoretical studies all with the same two-sided *P*-value for the null hypothesis of equal preference in the population.

Number of patients receiving *A* and *B*	Numbers preferring *A*: *B*	% preferring *A*	two-sided *P*-value
20	15 : 5	75.00	0.04
200	115 : 86	57.50	0.04
2 000	1046 : 954	52.30	0.04
2 000 000	1 001 445 : 998 555	50.07	0.04

would be considered too small to permit reliable conclusions, while the last trial (with a preference proportion of 50.07%) would be considered as evidence *for* rather than *against* equivalence, since the preference rates are, from any practical perspective, equally balanced. Thus equal *P*-values can lead to very different conclusions depending on the sample size.

4.4.2 Bayes factors as an alternative to *P*-values: simple hypotheses

We have already seen (Section 3.3) that the Bayes factor or likelihood ratio is the natural way to compare the support for two alternative hypotheses: when these hypotheses are 'simple' (*i.e.* there are no unknown parameters), the Bayes factor is a measure of the evidence in the data alone and is not affected by any

prior probabilities. In the rather unrealistic situation that data are only reported as being 'significant at the $100\alpha\%$ level', the Bayes factor is

$$\text{BF} = \frac{p(\text{'significant'}|H_0)}{p(\text{'significant'}|H_1)} = \frac{\alpha}{1-\beta} \tag{4.1}$$

where α and β are the standard Type I and Type II error rates (Example 3.7).

It is important to note the behaviour of (4.1) as the sample size increases but the alternative hypothesis H_1 remains fixed. In this case the power of the study increases, and hence β decreases and the Bayes factor decreases towards α: we are left with the conclusion of Peto *et al.* (1976) that a 'significant' result provides more evidence against the null hypothesis for larger sample sizes.

This finding can be contrasted with Lindley and Scott (1984), who preface their statistical tables with the claim that '*all significance tests are dubious because the interpretation to be placed on the phrase ''significant at 5%'' depends on the sample size: it is more indicative of the falsity of the null hypothesis with a small sample than with a large one*'. We therefore appear to have contradictory claims that both smaller *and* larger studies suggest increased evidence against the null hypothesis when reporting a 'significant' result.

For simple alternative hypotheses, Royall (1986) explains this apparent paradox by contrasting two situations: that we know a study was significant at the 5% level, and that we know the exact *P*-value was 5%. The first was covered by (4.1), while the second is now considered for normal distributions. Suppose

$$y_m \sim \text{N}[\theta, \sigma^2/m]$$

and we wish to compare two simple hypotheses $H_0: \theta = 0$ against $H_1: \theta = \theta_A > 0$. Then the Bayes factor is the likelihood ratio

$$\begin{aligned}\text{BF} = \frac{p(y_m|\theta=0)}{p(y_m|\theta_A)} &= \exp\Big(-\frac{m}{2\sigma^2}[y_m^2 - (y_m-\theta_A)^2]\Big) \\ &= \exp\left(-\frac{m\,\theta_A}{\sigma^2}\left[y_m - \frac{\theta_A}{2}\right]\right).\end{aligned} \tag{4.2}$$

This reveals the intuitive behaviour that for $y_m < \theta_A/2$, the Bayes factor will exceed 1 and hence favour H_0, while if $y_m > \theta_A/2$ the Bayes factor will be less than 1 and favour H_1.

Equation (4.2) can also be written

$$\text{BF} = \exp\left(-\sqrt{m}z_m\delta + \frac{m\delta^2}{2}\right) \tag{4.3}$$

where $\delta = \theta_A/\sigma$ is a standardised version of the alternative hypothesis, and $z_m = y_m\sqrt{m}/\sigma$ is the standardised test statistic for H_0. The crucial observation is that, for fixed z_m and hence fixed *P*-value, the Bayes factor will *increase*

with increasing sample size m, and hence support Lindley and Scott's observation that smaller sample sizes are more indicative of the falsity of the null hypothesis.

The apparent paradox for simple alternative hypotheses is seen to be resolved by being clearer by what we mean by a 'significant' result: when we only know a result achieved significance at a fixed level, the evidence against H_0 increases with sample size, while if we know the exact significance level, evidence against H_0 decreases with sample size. This reveals the complexity of comparing Bayes factors with P-values, and we shall now add to the potential confusion by considering composite alternative hypotheses, which are seen to obey *both* the behaviours contrasted above.

4.4.3 Bayes factors as an alternative to *P*-values: composite hypotheses

In most cases in which P-values are currently used H_1 will be 'composite', in that it encompasses a range of parameter values θ as alternatives to the single value specified by H_0, typically $\theta = 0$. We therefore need a method to obtain an overall likelihood $p(\text{data}|H_1)$ in order to obtain the Bayes factor, *i.e.* $p(\text{data}|H_0)/p(\text{data}|H_1)$.

A likelihood-based solution is to use the 'minimum' Bayes factors, $\text{BF}_{\min}$, under H_1 (Goodman, 1999b). For a general alternative hypothesis $H_1: \theta \neq 0$ in the normal model considered in (4.2), the minimum Bayes factor occurs when $\theta_A = y_m$, and from (4.3) is

$$\text{BF}_{\min} = \exp(-z_m^2/2), \tag{4.4}$$

where $z_m = y_m\sqrt{m}/\sigma$ is the standardised test statistic for H_0. This produces a direct mapping between one-sided P-values, given by $\Phi(z_m)$, and minimum Bayes factors that is displayed as part of Figure 4.1: using Jeffreys' descriptions contained in Table 3.2, a two-sided P-value (denoted $2P$) of 0.001 is 'decisive evidence', $2P = 0.01$ is on the border of 'strong' and 'very strong', and $2P = 0.05$ is 'substantial'. The minimum Bayes factor thus leads to conclusions that are qualitatively similar to P-values but obey the likelihood principle and so are unaffected by stopping rules. However, they still suffer from the criticism displayed in Example 4.2: all the four studies have significance corresponding (up to a normal approximation) to a z statistic of $z_{0.04/2} = -2.05$, and hence would have the same minimum Bayes factor of $\exp(-2.05^2/2) = 1/8.2$: 'substantial' evidence against H_0.

As an alternative to a likelihood-based approach, in a full Bayesian analysis we need to specify a prior $p(\theta|H_1)$ under the alternative hypothesis. If we assume

$$\theta|H_1 \sim \text{N}[0, \sigma^2/n_0],$$

then from (3.23) we have that

$$y_m|H_1 \sim \mathrm{N}\left[0, \sigma^2\left(\frac{1}{n_0}+\frac{1}{m}\right)\right],$$

and hence the Bayes factor is easily shown to be

$$\mathrm{BF} = \sqrt{1+\frac{m}{n_0}}\ \exp\left[\frac{-z_m^2}{2(1+n_0/m)}\right]. \tag{4.5}$$

n_0 can approximately be interpreted as the number of 'imaginary' observations taking on the value of the null hypothesis $\theta = 0$, and hence reflects prior support under H_1 for parameter values 'near' (but not exactly) H_0. The problem then becomes that of assessing a reasonable value for n_0. This will be considered in Section 5.5.4 in which priors that explicitly consider the 'truth' of a (null) hypothesis are discussed, but we now note that Kass and Wasserman (1995) suggest that $n_0 = 1$ (a prior equivalent to a single observation) may be a reasonable choice in many circumstances.

Figure 4.1 displays the resulting relationship between two-sided *P*-values and Bayes factors for different choices of m/n_0, the ratio of data sample size to prior sample size under the alternative hypothesis. It is clear that Bayes factors can produce very different results from the standard measures of evidence, with a tendency towards preference for the null hypothesis: when m/n_0 is large we note that

$$\mathrm{BF} \approx \sqrt{\frac{m}{n_0}}\ \mathrm{BF}_{\min}. \tag{4.6}$$

An alternative way of examining the relationship between Bayes factors and *P*-values is shown in Figure 4.2, in which the change in Bayes factor with increasing ratio m/n_0 is shown for fixed *P*-values. For example, evidence that is labelled as $2P = 0.001$ is considered only just 'strong' when the sample size is small relative to the prior precision, but becomes 'very strong' for moderate sample sizes, and then reduces to only 'substantial' for overwhelming large experiments. This non-monotonic relationship to sample size appears to match well the intuitive desire for measures of evidence brought out in Example 4.2.

As we have noted in Section 4.4.2, the importance of sample size and plausibility of benefits in interpreting *P*-values has often been stressed even within the non-Bayesian literature: for example, the ISIS-4 investigators state that 'when moderate benefits or negligibly small benefits are both much more plausible than extreme benefits, then a $2P = 0.001$ effect in a large trial or overview would provide much stronger evidence of benefit than the same significance level in a small trial, a small overview, or a small subgroup analysis' (Collins *et al.*, 1995). Examination of Figure 4.2 shows that their insight is again

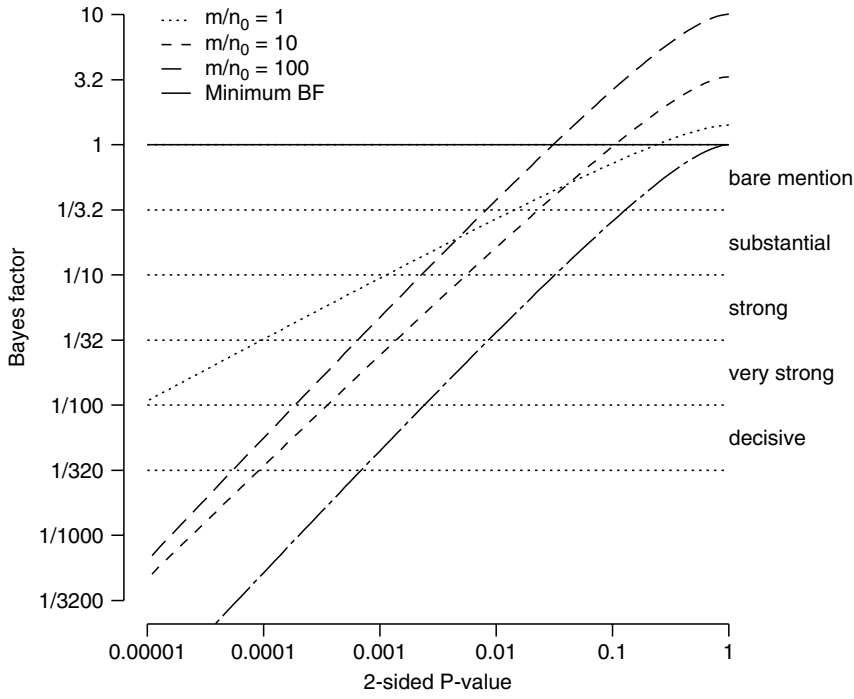

Figure 4.1 Bayes factors compared to P-values for composite normal hypotheses, showing bands corresponding to Jeffreys levels of evidence. The minimum Bayes factor is the Bayes factor against the maximum likelihood estimate for the parameter under H_1.

matched by the behaviour of the Bayes factor: smaller benefits being more plausible correspond to n_0 being relatively large, and hence m/n_0 lies in the 'dip' of Figure 4.2 in which stronger evidence is shown compared to smaller sample sizes. However, Figure 4.2 suggests a conclusion that is not mentioned by Collins *et al.* (1995) but seems quite appropriate: if the 'large trial or overview' becomes *extremely* large but still only significant at $2P = 0.001$, then the evidence for benefit will start to decline again.

For composite hypotheses it appears that neither of the views contrasted in Section 4.4.2 holds: there is no simple monotonic relationship between Bayes factors and P-values, and it is perhaps not surprising that so much apparent confusion has arisen.

Bayes factors can be obtained in the presence of nuisance parameters, but this makes the dependence on the prior distribution of even more concern. This is an area of substantial research and discussion (Kass and Raftery, 1995).

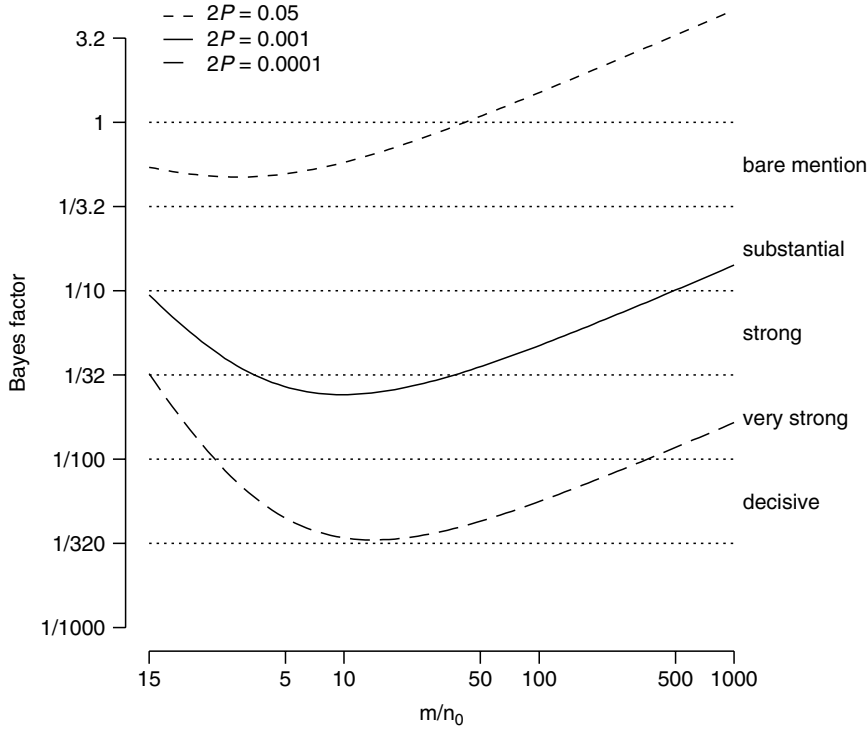

Figure 4.2 Bayes factors for composite normal hypotheses for fixed *P*-values and different m/n_0 ratios, *i.e.* ratio of observed to prior sample size, with areas delineated by Jeffreys' levels of evidence.

4.4.4 Bayes factors in preference studies

Consider the preference studies used in Examples 4.1 and 4.2, in which the underlying proportion of individuals preferring option A to B is assumed to be θ. Then the number of preferences r for option A out of m independent trials has a binomial distribution (Section 2.6.1)

$$p(r|\theta, m) = \binom{m}{r} \theta^r (1-\theta)^{m-r}.$$

The maximum likelihood estimator is $\hat{\theta} = r/m$, and so the minimum Bayes factor for the null hypothesis $H_0: \theta = 0.5$ is

$$\mathrm{BF}_{\min} = \frac{p(r|\theta = 0.5)}{p(r|\theta = \hat{\theta})} = \frac{1}{2^m} \Big/ \left(\frac{r}{m}\right)^r \left(1 - \frac{r}{m}\right)^{m-r}.$$

Assuming $p(\theta|H_1)$ is a uniform prior (as suggested by Jeffreys) gives the predictive distributions

$$p(r|m, H_0) = \binom{m}{r} \frac{1}{2^m}, \tag{4.7}$$

$$p(r|m, H_1) = \frac{1}{m+1}. \tag{4.8}$$

Equation (4.7) is simply the Binomial probability when $\theta = 0.5$, and (4.8) shows r has a uniform distribution over all its possible values $0, 1, 2, \ldots, m$, and is a special case of the beta-binomial distribution (Section 3.13.2) with $a = 1, b = 1$. Hence the exact Bayes factor is

$$\text{BF} = \binom{m}{r} \frac{m+1}{2^m}. \tag{4.9}$$

For both the likelihood and Bayesian approaches we can use approximations for large samples by calculating the P-value, obtaining a corresponding z-statistic, and substituting in (4.4) and (4.5). For the Bayesian approximation we do, however, need to specify a normal distribution for $p(\theta|H_1)$ instead of a uniform distribution, and the problem lies in choosing the normal variance. In 'interesting' situations the Bayes factor is driven by the ordinate of the $p(\theta|H_1)$ at the null hypothesis, and so we choose a normal distribution that has the same ordinate as a uniform distribution, namely 1. Were $\theta|H_1 \sim \text{N}[0.5, \sigma^2/n_0]$, then the ordinate at $\theta = 0.5$ would be $\sqrt{n_0/(2\pi\sigma^2)}$. σ^2 is the variance of a single obervation under H_0, and so $\sigma^2 = \theta(1-\theta) = \frac{1}{4}$ and equating the resulting ordinate $\sqrt{2n_0/\pi}$ to 1 gives $n_0 = \pi/2 = 1.57$, not far from the value of $n_0 = 1$ suggested by Kass and Wasserman (1995). Thus, for a preference study with a standardised test statistic of z_m, our approximate Bayes factors are

$$\text{BF}_{\min} \approx \exp(-z_m^2/2), \tag{4.10}$$

$$\text{BF} \approx \sqrt{1 + \frac{m}{1.57}}\ \exp\left[\frac{-z_m^2}{2(1 + 1.57/m)}\right]. \tag{4.11}$$

The quality of these approximations is explored in Example 4.3.

We again emphasise that the Bayes factors, whether likelihood or Bayesian, are unaffected by whether the designs were sequential or fixed sample size.

Example 4.3 *Preference (continued): Bayes factors in preference studies*

Table 4.4 shows the quality of the approximate Bayes factors for the preference data, using the exact Bayes factors in (4.9), and approximations (4.10)

and (4.11). The approximations for the Bayes factors appear reasonable, particularly for the minimum Bayes factor. For Example 4.1, both Bayes factors express minimal evidence against the null hypothesis, as would be expected from Figure 4.1. For the data in Example 4.2, the increasing sample size leaves the minimum Bayes factor constant at 'substantial' evidence against H_0, whereas the full Bayes factor changes from favouring H_1 to favouring H_0, and then steadily increases its support for H_0. This behaviour reflects the pattern shown in Figure 4.2 for increasing sample size and fixed P-value, following approximately the trajectory of $2P = 0.05$.

Table 4.4 Bayes factors for preference studies when m individuals asked whether they prefer A or B. The first row is from Example 4.1 and the other four rows from Example 4.2. z_m is a standardised test statistic that would give rise to the observed one-sided P-value. The approximate Bayes factor assumes $n_0 = 1.57$.

m	r prefer A	$\hat{\theta}$	One-sided P-value	z_m	*Minimum Bayes factor*		*Bayes factor*	
					Exact	Approx	Exact	Approx
6	5	0.83	0.063	1.53	0.23	0.31	0.65	0.86
20	15	0.75	0.02	2.05	0.07	0.12	0.31	0.53
200	115	0.575	0.02	2.05	0.10	0.12	1.20	1.41
2 000	1046	0.523	0.02	2.05	0.12	0.12	4.30	4.37
2 000 000	1 001 445	0.500 722 5	0.02	2.05		0.12	139.8	138.0

Rather than formulating these problems as hypothesis tests, it may be much more appropriate to assess a reasonable prior for θ and then report $p(\theta > 0.5|r, m)$ – the posterior probability that a majority of the population prefer A to B. Of course, such a measure suffers from exactly the same criticism of the P-values in Example 4.2: the posterior probability may be high even though the 'majority' that prefers A is negligible. In this case it may be more appropriate to assess an 'important majority' $\theta_S > 0.5$, and consider the $p(\theta > \theta_S|r, m)$. See Section 6.3 for applications of these ideas in clinical trials.

4.4.5 Lindley's paradox

Close examination of the top right-hand corner of Figure 4.1 reveals what might appear as odd behaviour: when the ratio m/n_0 is high, and the P-value is just marginally significant *against* H_0, the Bayes factor can be greater than 1 and hence *support* H_0. This somewhat surprising result is known as *Lindley's paradox*, after Lindley (1957). An informal explanation is as follows. First, for large sample sizes, a P-value can be small even if the data support values of θ very close to the null hypothesis, as shown for the large sample sizes in Example 4.2. Second, such data may indeed be unlikely under the null hypothesis, but are even more unlikely under an alternative that spreads the prior probability thinly over a wide range of potential values. Hence the Bayes factor can support H_0

when a significance test would reject it, essentially as the lesser of two evils. An example of this behaviour is shown in Example 4.4.

Example 4.4 *GREAT (continued): A Bayes factor approach*

From the 'Evidence from study' component of Example 3.6 we note that the standardised test statistic is $z = 2.03$: just significant evidence against H_0 at the traditional two-sided $P < 0.05$. The 'minimum' Bayes factor against H_0 is $\text{BF}_{\text{min}} = \exp(-z^2/2) = 0.13 = 1/7.8$, corresponding to 'substantial' evidence against H_0. Thus the classical and Bayesian approaches align to a reasonable extent if we allow the alternative hypothesis to be dictated by the data.

However, a fully Bayesian approach might place a prior on $\theta = \log(\text{OR})$ under H_1, centred on 0 and with a large variance. For example, suppose we used a prior with $n_0 = 0.5$ which is essentially uniform over the log(OR) scale.

Since $m = 30.5$ we have a ratio of likelihood to prior precision of $m/n_0 = 61$. From (4.6) the Bayes factor is approximately $\sqrt{m/n_0}\ \text{BF}_{\text{min}} = 1.001$ (the exact value from substitution into (4.5) is 1.04), *i.e.* slight evidence *in favour* of H_0! This is an example of Lindley's paradox.

4.5 KEY POINTS

1. There is room for dispute over some of the fundamental principles of conventional statistical analysis.
2. The likelihood principle states that only the observed data should affect inferences: classical sequential analysis disobeys this.
3. The pragmatic interpretation of *P*-values strongly depends on sample size.
4. Minimum Bayes factors obey the likelihood principle, but have similar qualitative behaviour to *P*-values.
5. Proper Bayes factors can, for large sample sizes relative to the prior precision, support the null hypothesis when a classical analysis would lead to its rejection.

EXERCISES

4.1. Confirm the form of the Bayes factor given by (4.5).

4.2. Calculate the minimum Bayes factor corresponding to the three levels of significance considered in Figure 4.2. In what circumstances might the

minimum Bayes factor exaggerate the evidence against the null hypothesis, compared to a full Bayesian approach?

4.3. In the preference studies described in Section 4.4.4, suppose we observed data that were just 'significant', with a two-sided P-value of 0.05. Assume $n_0 = 1.57$.
   (a) What sample size (approximately) would yield a Bayes factor of 1, *i.e.* indifference between the null and alternative hypotheses?
   (b) What observed data would have given $2P = 0.05$ with this sample size?

4.4. For the PROSPER trial in Exercise 2.8 calculate the one-sided P-value, the minimum Bayes factor, and the Bayes factor corresponding to a sceptical prior distribution with an effective number of events $n_0 = 1$.

4.5. In Example 4.4, what would be the Bayes factor were we to adopt Kass and Wasserman's suggestion of $n_0 = 1$?