Part I

BASIC CONCEPTS IN BAYESIAN METHODS

1

Modes of statistical inference

The central activity in statistics is inference. Statistical inference is a procedure or a collection of activities with the aim to extract information from (gathered) data and to generalize the observed results beyond the data at hand, say to a population or to the future. In this way, statistical inference may help the researchers in suggesting or verifying scientific hypotheses, or decision makers in improving their decisions. Inference obviously depends on the collected data and on the assumed underlying probabilistic model that generated these data, but it also depends on the approach to generalize from the known (data) to the unknown (population). We distinguish two mainstream views/paradigms to draw statistical inference, i.e. the frequentist approach and the Bayesian approach. In-between these two paradigms is the (pure) likelihood approach.

In most of the empirical research, but definitely in medical research, scientific conclusions need to be supported by a 'significant result' using the classical *P*-value. Significance testing belongs to the frequentist paradigm. However, the frequentist approach does not consist of one unifying theory but is rather the combination of two approaches, i.e. the inductive approach of Fisher who introduced the null-hypothesis, the *P*-value and the significance level and the deductive procedure of Neyman and Pearson who introduced the alternative hypothesis and the notion of power. First, we review the practice of frequentist significance testing and focus on the popular *P*-value. More specifically we look at the value of the *P*-value in practice. Second, we treat an approach that is purely based on the likelihood function not involving any classical significance testing. This approach is based on two fundamental likelihood principles that are also essential for the Bayesian philosophy. Finally, we end this chapter by introducing the principles of the Bayesian approach and we give an outlook of what the Bayesian approach can bring to the statistician. However, at least three more chapters will be needed to fully develop the Bayesian theory.

Bayesian Biostatistics, First Edition. Emmanuel Lesaffre and Andrew B. Lawson.

1.1 The frequentist approach: A critical reflection

1.1.1 The classical statistical approach

It is perhaps an oversimplification to speak of a classical statistical approach. Nevertheless, we mean by this the ensemble of methods that provides statistical inference based on the classical *P*-value, the significance level, the power and the confidence interval (CI). To fix ideas, we now exemplify current statistical practice with a *randomized controlled clinical trial* (*RCT*). In fact, the RCT is the study design that, by excellence, is based on the classical statistical tool box of inferential procedures. We assume that the reader is familiar with the classical concepts in inferential statistics.

For those who have never experienced a RCT, here is a brief description. A clinical trial is an experimental study comparing two (or more) medical treatments on human subjects, most often patients. When a control group is involved, the trial is called *controlled*. For a *parallel* group design, one group of patients receives one treatment and the other group(s) receive(s) the other treatment and all groups are followed up in time to measure the effect of the treatments. In a *randomized study*, patients are assigned to the treatments in a random manner. To minimize bias in evaluating the effect of the treatments, patients and/or care givers are blinded. When only patients are blinded one speaks of a *single-blinded* study, but when both patients and care givers are blinded (and everyone involved in running the trial) one speaks of a *double-blinded* trial. Finally, when more than one center (e.g. hospital) is involved one deals with a multicenter study.

Example I.1: Toenail RCT: Evaluation of two oral treatments for toenail infection using the frequentist approach

A randomized, double-blind, parallel group, multicenter study was set up to compare the efficacy of two oral treatments for toenail infection (De Backer *et al*. 1996). In this study, two groups of 189 patients were recruited, and each received 12 weeks of treatment (Lamisil: treatment *A* and Itraconazol: treatment *B*), with 48 weeks of follow-up (FU). The significance level was set at $\alpha = 0.05$. The *primary endpoint* (upon which the sample size was based) in the original study was negative mycology, i.e. a negative microscopy and a negative culture. Here, we look at another endpoint, i.e. *unaffected nail length at week 48* on a subset of patients for whom the big toenail was the target nail. One hundred and thirty-one patients treated with *A* and 133 treated with *B* were included in this comparison. Note that we only included those patients present at the end of the study. The observed mean (SD) lengths in millimeter at week 48 were 9.07 (4.92) and 7.70 (5.33), for treatments *A* and *B*, respectively. Suppose the (population) average for treatment *A* is μ_1 while for treatment *B* it is μ_2. Therefore, the null-hypothesis is $H_0 : \Delta = \mu_1 - \mu_2 = 0$ and can be evaluated with an unpaired *t*-test at a two-sided significance level of $\alpha = 0.05$. Upon completion of the study, the treatment estimate was $\widehat{\Delta} = 1.38$ with an observed value of the *t*-statistic equal to $t_{obs} = 2.19$. This result lies in the rejection region corresponding to $\alpha = 0.05$ yielding a statistically significant result (at 0.05). Thus, according to the Neyman–Pearson (NP) approach we (can) reject that *A* and *B* are equally effective.

It is common to report also the *P*-value of the result to indicate the strength of evidence against the hypothesis of two equally effective treatments. Here, we obtained a two-sided *P*-value equal to 0.030, which is a measure of evidence against H_0 in a Fisherian sense. □

In Section 1.1.2, we reflect on what message the *P*-value can bring to the researcher. We will also indicate what properties the *P*-value does not have (but assumed to have).

1.1.2 The *P*-value as a measure of evidence

Fisher developed the *P*-value in the context of well-planned limited agricultural experiments in a time when computations had to be done by hand. Nowadays, a great variety of studies are undertaken in the medical research usually of an exploratory nature and often evaluating hundreds to thousands of *P*-values. The *P*-value is an intuitively appealing measure against the null-hypothesis, but that it is not always perceived in the correct manner. Here, we will further elaborate on the use and misuse of the *P*-value in practice.

The *P*-value is not the probability that H_0 is (not) true A common error is to interpret the *P*-value as a probability that H_0 is (not) true. However, the *P*-value only measures the extremeness of the observed result under H_0. The probability that H_0 is true is formally $p(H_0 \mid \text{data})$, which we shall call the posterior probability of the null-hypothesis in the following text, given the observed data. This probability is based on Bayes theorem and depends on the prevalence of H_0 (see also Example I.11).

The *P*-value depends on fictive data The *P*-value does not express the probability that the observed result occurred under H_0, but is rather the probability of observing this or a more extreme result under H_0. This implies that the calculation of the *P*-value is based not only on the observed result but also on fictive (never observed) data.

Example I.2: Toenail RCT: Meaning of *P*-value
The *P*-value is equal to the probability that the test statistic exceeds the observed value if the null-hypothesis were true. The computation of the *P*-value is done using probability laws, but could also be represented by a simulation experiment. For instance, in the toenail infection study the *P*-value is approximately equal to the proportion of studies, out of (say) 10 000 imaginary studies done under H_0 (two identical treatments), that yield a *t*-value more extreme than $t_{\text{obs}} = 2.19$. In Figure 1.1, the histogram of imaginary results is displayed together with the observed result. □

The *P*-value depends on the sample space The above simulation exercise shows that the *P*-value is computed as a probability using the *long-run frequency definition*, which means that a probability for an event *A* is defined as the ultimate proportion of experiments that generated that event to the total number of experiments. For a *P*-value, the event *A* corresponds to a *t*-value located in the rejection region. This makes it clear that the *P*-value depends on the choice of the fictive studies and, hence, also on the sample space. The particular choice can have surprising effects, as illustrated in Example I.3.

Example I.3: Accounting for interim analyses in a RCT
Suppose that a randomized controlled trial has been set up to compare two treatments and that four *interim analyses for efficacy* were planned. An interim analysis for efficacy is a statistical comparison between the treatment groups prior to the end of the study to see

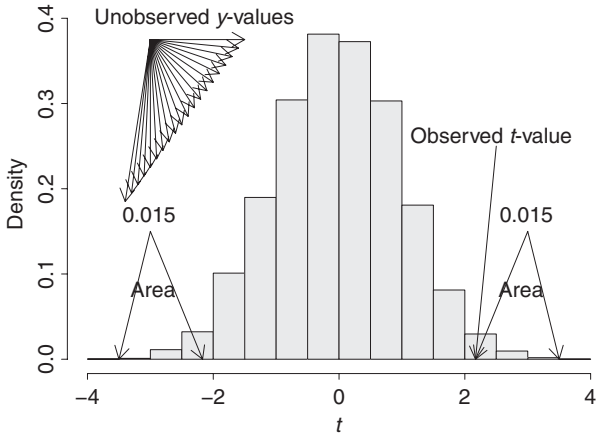

Figure 1.1 Graphical representation of the *P*-value by means of a simulation study under H_0.

whether the experimental treatment is better than the control treatment. The purpose is to stop the trial earlier if possible. When more than one comparison is planned, one needs to correct in a frequentist approach for *multiple testing*. A classical correction for multiple testing is *Bonferroni's rule* which dictates that the significance level at each comparison (*nominal significance level*) needs to be made more stringent, i.e. α/k where k is the total number of tests applied, to arrive at an overall (across all comparisons) type I error rate less or equal to α. Bonferroni's procedure is approximate; an exact control of the type I error rate is obtained with a *group sequential design* (Jennison and Turnbull 2000). With Pocock's group sequential method and 5 analyses (4 interim + 1 final analyses), a significance level of 0.016 is handled at each analysis to achieve a global significance level of 0.05. Thus, when the study ran until the end and at the last analysis a *P*-value of 0.02 was obtained, then the result cannot be claimed significant with Pocock's rule. However, if the same result had been obtained without planning interim analyses, then this trial produced a significant result! Thus, in the presence of two identical results, one cannot claim evidence against the null-hypothesis in one case, while in the other case we would conclude that the two treatments have a different effect. □

In statistical terminology, the different evidence for the treatment effect in the two RCTs of Example I.3 (with an identical *P*-value) is due to a different sample space (see below) in the two scenarios. This is further illustrated in Example I.4.

Example I.4: Kaldor' *et al.* case-control study: Illustration of sample space
In the (matched) case-control study (Kaldor *et al*. 1990) involving 149 cases (leukaemia patients) and 411 controls, the purpose was to examine the impact of chemotherapy on leukaemia in Hodgkin's survivors (Ashby *et al*. 1993). The 5-year survival of Hodgkin's disease (cancer of the lymph nodes) is about 80%, but the survivors have an excess risk of developing solid tumors, leukaemia and/or lymphomas. In Table 1.1, the cases and controls are subdivided according to exposure to chemotherapy or not.

Table 1.1 Kaldor' *et al*. case-control study (Kaldor *et al*. 1990): frequency table of cases and controls subdivided according to their exposure to chemotherapy.

Treatment	Controls	Cases
No chemo	160	11
Chemo	251	138
Total	411	149

Ignoring the matched character of the data, the analysis of the 2×2-contingency table by a Pearson chi-squared test results in $P = 7.8959 \times 10^{-13}$ with a chi-squared value of 51.3. With the Fisher's exact test, a P-value of 1.487×10^{-14} was obtained. Finally, the estimated odds ratio is equal to 7.9971 with a 95% CI of [4.19, 15.25]. □

The chi-squared test and the Fisher's exact test have a different *sample space*, which is the space of possible samples considered to calculate the null distribution of the test statistic. The sample space for the chi-squared test consists of the 2×2-contingency tables with the same total sample size (n), while for Fisher's exact test the sample space consists of the subset of 2×2-contingency tables with the same row and column marginal totals. The difference between the two sample spaces explains here partly the difference in the two test results. In Example I.3, it is the sole reason for the different evidence from the two RCTs. The conclusion of a scientific experiment, hence, not only depends on the results of that experiment but also on the results of experiments that did not and will never happen. This finding has triggered a lot of debate among statisticians (see Royall 1997).

The *P*-value is not an absolute measure A small P-value does not necessarily imply a large difference between two treatments or a strong association among variables. Indeed, as a measure of evidence the P-value does not take the size of the study into account. There have been vivid discussions on how a small P-value should be interpreted as a function of the size of the study (see Royall 1997).

The *P*-value does not take all evidence into account Let us take the following example also discussed by Ashby *et al*. (1993).

Example I.5: Merseyside registry results
Ashby *et al*. (1993) reported on data obtained from a subsequent registry study in UK (after Kaldor *et al*.'s case-control study) to check the relationship between chemotherapy and leukemia among Hodgkin's survivors. Preliminary results of the Merseyside registry were reported in Ashby *et al*. (1993) and are reproduced in Table 1.2. The P-value obtained from the chi-squared test with continuity correction equals 0.67. Thus, formally there is no reason to worry that chemotherapy may cause leukemia among Hodgkin's survivors. Of course, every epidemiologist would recognize that this study has no chance of finding a relationship between chemotherapy and leukemia because of the small study size. By simply analyzing the data of the Merseyside registry, no evidence of a relationship can be established. □

Table 1.2 Merseyside registry: frequency table of cases and controls subdivided according to their exposure to chemotherapy.

Treatment	Controls	Cases
No chemo	3	0
Chemo	3	2
Total	6	2

Is it reasonable to analyze the results of the Merseyside registry in isolation, not referring to the previous study of Kaldor *et al*. (1990)? In other words, should one forget about the historical data and assume that one cannot learn anything from the past? The answer will depend on the particular circumstances, but it is not obvious that the past should never play a role in the analysis of data.

1.1.3 The confidence interval as a measure of evidence

While the *P*-value has been criticized by many statisticians, it is more the (mis)use of the *P*-value that is under fire. Nevertheless, there is a growing preference to replace the *P*-value by the (95%) CI.

Example I.6: Toenail RCT: Illustration of 95% confidence interval
The 95% CI for Δ is equal to [0.14, 2.62]. Technically speaking we can only say that in the long run 95% of those intervals will contain the true parameter (the 95% CI is based on the long-run frequency definition of probability). But for our RCT, the 95% CI will either contain the true parameter or not (with probability 1)! In our communication to nonstatisticians, we never use the technical definition of the CI. Rather, we say that the 95% CI [0.14, 2.62] expresses that we are uncertain about the true value of Δ and that it most likely lies between 0.14 and 2.62 (with 0.95 probability). □

The 95% CI expresses our uncertainty about the parameter of interest and as such is considered to give better insight into the relevance of the obtained results than the *P*-value. However, the adjective '95%' refers to the procedure of constructing the interval and not to the interval itself. The interpretation that we give to nonstatisticians has a Bayesian flavor as will be seen in Chapter 2.

1.1.4 An historical note on the two frequentist paradigms*

In this section, we expand on the difference between the two frequentist paradigms and how they have been integrated in practice into an apparently one unifying approach. This section is not essential for the remainder of the book and can be skipped. A more in-depth treatment of this topic can be found in Hubbard and Bayarri (2003) and the papers of Goodman (1993, 1999a, 1999b) and Royall (1997).

The Fisherian and the NP approach are different in nature but are integrated in current statistical practice. Fisher's views on statistical inference are elaborated in two of his books: *Statistical Methods for Research Workers* (Fisher 1925) and *The Design of Experiments* (Fisher 1935). He strongly advocated the inductive reasoning to generate new hypotheses. Fisher's approach to inductive inference goes via the rejection of the *null-hypothesis*, say $H_0 : \Delta = 0$. His *significance test* constitutes of a statistical procedure based on a test statistic for which the sampling distribution, given that $\Delta = 0$ holds, is determined. He called the probability under H_0 of obtaining the observed value of that test statistic or a more extreme one, the P-value. To Fisher, the P-value was just a practical tool for inductive inference whereby the smaller the P-value implies a greater evidence against $\Delta = 0$. Further, according to Fisher the null-hypothesis should be 'rejected' when the P-value is small, say less than a prespecified threshold $\alpha = 0.05$ called the *level of significance*. Fisher's rule for rejecting H_0 is, therefore, when $P \leq 0.05$, but he recognized (Fisher 1959) that rejection may have two meanings: either that an exceptionally rare chance has occurred or the theory (according to the null-hypothesis) is not true.

In their approach to statistical testing, Neyman and Pearson (1928a, 1928b, 1933) needed an *alternative hypothesis* (H_a), say $\Delta \neq 0$. Once the data have been observed, the investigator needs to decide between two actions: reject H_0 (and accept H_a) or accept H_0 (and reject H_a). NP called their procedure an *hypothesis test*. Their approach to research has a decision theoretic flavor, i.e. decision makers can commit two errors: (1) type I error with probability P(type I error) $= \alpha$ (*type I error rate*) when H_0 is rejected while in fact true and (2) type II error with probability P(type II error) = β (*type II error rate*) when H_a is rejected while in fact true. In this respect they introduced the *power* of a test, equal to $1 - \beta$ for an alternative hypothesis $H_a : \Delta = \Delta_a$. NP argued that a statistical test must minimize the probability of making the wrong decision and demonstrated (Neyman and Pearson 1933) that the well-known *likelihood-ratio test* minimizes β for given a value for α. The NP approach is in fact deductive and reasons from the general to the particular and thereby makes claims from the particular to the general only in the long run. NP strived that one shall not be wrong too often. In other words, they rather advocated 'inductive behavior'. Fisher strongly disliked this viewpoint and both parties ended up in a never-ending debate.

Despite the strong historical disagreement between the proponents of the two approaches, nowadays the two philosophies are mixed up and presented as a unifying methodology. Hubbard and Bayarri (2003) (see also references therein) warned for the confusion this unification might imply, especially for the danger that the P-value is wrongly interpreted as a type I error rate. Indeed, the P-value was introduced by Fisher as a surprise index vis-à-vis the null-hypothesis and in the light of the data. It is an a posteriori determined probability. A problem occurs when the P-value is given the status of an a priori determined error rate. For example, suppose the significance level is $\alpha = 0.05$, chosen in advance. Thus, if upon completion of the study, we obtain $P = 0.023$ we say that H_0 is rejected at 0.05. However, we cannot say that H_0 is rejected at the 0.025 level, because in this sense α is chosen after the facts and P obtains the status of a prespecified level. In medical papers, the P-value is often given the nature of a prespecified value. For instance when significant results are indicated by asterisks, e.g. '*' for $P < 0.05$, '**' for $P < 0.01$ and '***' for $P < 0.001$, then the impression is created that for a '**' result significance at 0.01 can be claimed. Following Carl Popper (Popper 1959), Fisher claimed that one can never accept the null-hypothesis, only disprove it. In contrast, according to the NP approach subsequent to the hypothesis test there are two possible actions: one 'rejects' the null-hypothesis (and accepts the alternative hypothesis) or vice versa. This creates another clash of the two approaches. Namely, if the NP approach

is the basis for statistical inference, then there is in principle no problem in accepting the null-hypothesis. However, one of the basic principles in classical statistical practice is never to accept H_0 in case of a nonsignificant result that is in the spirit of Fisher's significance testing. Note that in clinical trials the standard approach is the NP approach, but accepting the null-hypothesis would be a major flaw.

Because of the above difficulties with the *P*-value and that classical statistical inference is claimed to be not coherent, Goodman (1993, 1999a, 1999b) and others advocated to use Bayesian inference tools such as the Bayes factor. Having said this, others still regard it as a useful tool in some circumstances. For instance, Hill (1996) writes: 'Like many others, I have come to regard the classical *P*-value as a useful diagnostic device, particularly in screening large numbers of possibly meaningful treatment comparisons.' Further, in some cases (one-sided hypothesis testing) the *P*-value and Bayesian inference come close (see Section 3.8.3). Finally, Weinberg (2001) argues that 'It is time to stop blaming the tools, and turn our attention to the investigators who misuse them.'

1.2 Statistical inference based on the likelihood function

1.2.1 The likelihood function

The concept of *likelihood* was introduced by Fisher (1922). It expresses the plausibility of the observed data as a function of the parameters of a stochastic model. As a function of the parameters the likelihood is called the *likelihood function*. Statistical inference based on the likelihood function differs fundamentally from inference based on the *P*-value, although both approaches were promoted by Fisher as a tool for inductive inference. To fix ideas, let us look at the likelihood function of a binomial sample. The following example dates back to Cornfield (1966) but is rephrased in terms of a surgery experiment.

Example I.7: A surgery experiment
Assume that a new but rather complicated surgical technique was developed in a hospital with a nonnegligible risk for failure. To evaluate the feasibility of the technique the chief surgeon decides to operate on $n = 12$ patients with this new procedure. Upon completion of the 12 operations, he reports $s = 9$ successes. Let the outcome of the ith operation be denoted as $y_i = 1$ for a success and $y_i = 0$ for a failure. The total experiment yields a sample of n independent binary observations $\{y_1, \ldots, y_n\} \equiv \boldsymbol{y}$ with s successes. Assume that the probability of success remains constant over the experiment, i.e. $p(y_i) = \theta$, $(i = 1, \ldots, n)$. Then the probability of the observed number of successes is expressed by the *binomial distribution*, i.e. the probability that s successes out of n experiments occur when the probability of success in a single experiment is equal to θ, is given by

$$f_\theta(s) = \binom{n}{s} \theta^s (1-\theta)^{n-s} \text{ with } s = \sum_{i=1}^{n} y_i, \tag{1.1}$$

where $f_\theta(s)$ is a discrete distribution (as a function of s) with the property that $\sum_{s=0}^{n} f_\theta(s) = 1$. When s is kept fixed and θ is varying, $f_\theta(s)$ becomes a continuous function of θ, called the *binomial likelihood function*. The likelihood function could be viewed as expressing the plausibility of θ in the light of the data and is therefore denoted as $L(\theta \mid s)$. The graphical representation of the binomial likelihood function is shown in Figure 1.2(a) for $s = 9$ and

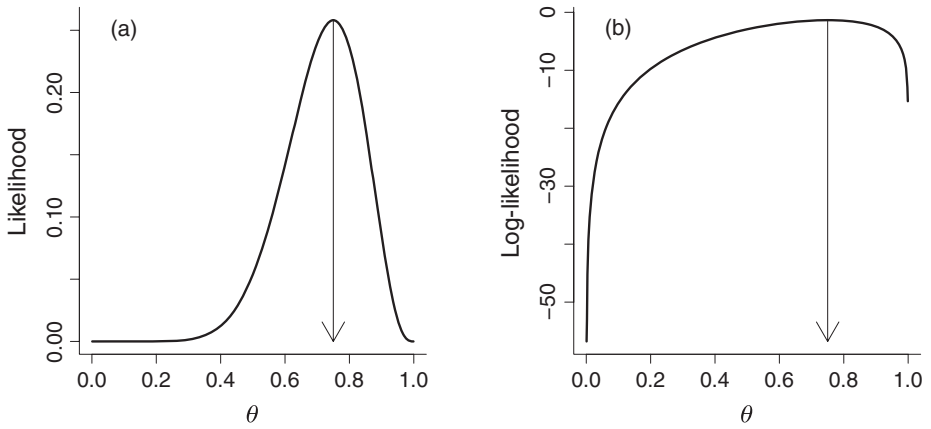

Figure 1.2 Surgery experiment: likelihood (a) and log-likelihood function (b) corresponding to $s = 9$ successes out of $n = 12$ operations from Example I.7.

$n = 12$. The figure shows that values of θ close to zero and close to one are not supported by the observed result of 9 successes out of 12 operations. On the other hand, values above 0.5 and below 0.9 are relatively well supported by the data with the value $\theta = 9/12 = 0.75$ best supported. □

The value of θ that maximizes $L(\theta \mid s)$ is called the *maximum likelihood estimate* (*MLE*) and is denoted as $\widehat{\theta}$. To determine $\widehat{\theta}$, we maximize $L(\theta \mid s)$ with respect to θ. It is equivalent and easier to maximize the logarithm of $L(\theta \mid s)$, called the *log-likelihood*, and denoted as $\ell(\theta \mid s)$.

Example I.7: (continued)
The log-likelihood for the surgery experiment is given by

$$\ell(\theta \mid s) = c + [s \log \theta + (n - s) \log(1 - \theta)], \tag{1.2}$$

where c is a constant. The first derivative with respect to θ gives the expression $\frac{s}{\theta} - \frac{(n-s)}{(1-\theta)}$. Equating this expression to zero gives the MLE equal to s/n, and thus, $\widehat{\theta} = 0.75$ (for $s = 9$ and $n = 12$), which is the sample proportion. Figure 1.2(b) shows $\ell(\theta \mid s)$ as a function of θ. □

1.2.2 The likelihood principles

Inference based on the likelihood function naturally adheres to two *likelihood principles* (*LP*) (Berger and Wolpert 1984):

1. *Likelihood principle 1*: All evidence, which is obtained from an experiment, about an unknown quantity θ is contained in the likelihood function of θ for the given data.
2. *Likelihood principle 2*: Two likelihood functions for θ contain the same information about θ if they are proportional to each other.

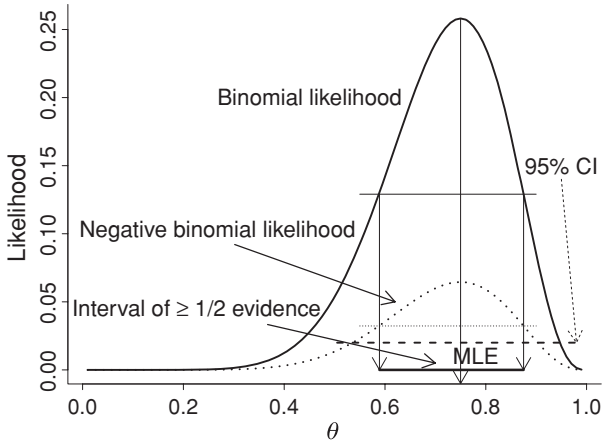

Figure 1.3 Surgery experiments: binomial and negative binomial (Pascal) likelihood functions together with MLE and interval of at least 0.5 maximal evidence and the classical two-sided 95% CI.

The first LP implies that the choice between two values of an unknown parameter is made via the likelihood function evaluated at those values. This leads to the *standardized likelihood* and the *interval of evidence*, which will be introduced in Example I.7 (continued) below.

Example I.7: (continued)
The binomial likelihood for $s = 9$ and $n = 12$ is maximal at $\widehat{\theta} = 0.75$. In a frequentist context, we could test the observed proportion against an a priori chosen value for θ, say 0.5 and we calculate a 95% CI for θ.

According to the likelihood function, there is maximal evidence for $\theta = 0.75$. The ratio of the likelihood functions at $\theta = 0.5$ and at $\theta = 0.75$ can be used as a measure of the relative evidential support given by the data for the two hypotheses. This ratio is called the *likelihood ratio* and is here equal to 0.21. The function $L(\theta \mid s)/L(\widehat{\theta} \mid s)$ (here $L(\theta \mid s)/L(0.75 \mid s)$) is called the *standardized likelihood*. On the standardized likelihood scale, one can read off that the evidence for $\theta = 0.5$ is about 1/5 the maximal evidence. Note that this comparison does not involve any fictive data, only the observed data play a role.

One can also construct an interval of parameter values that show at least a fraction of the maximal evidence. For instance, the *interval of (at least half of the maximal) evidence* consists of those θ-values that correspond to at least half of $L(\widehat{\theta} \mid s)$, i.e. with a standardized likelihood of at least 0.5 (see Figure 1.3). This interval provides direct evidence on the parameter of interest and is related to the highest posterior density interval introduced in Section 3.3.2. On the same figure, the classical 95% CI [0.505, 0.995] is indicated. In general, the 95% CI only represents an interval of evidence when the likelihood function is symmetric. □

The second LP states that two likelihood functions for θ contain the same information about that parameter if they are proportional to each other. This is called the *relative likelihood principle*. Thus, when the likelihood is proportional under two experimental conditions, irrespective of the way the results were obtained, the information about the unknown parameter must be the same. Example I.8 contrasts in this respect the difference between the frequentist and the likelihood viewpoints.

Example I.8: Another surgery experiment
Assume that the chief surgeon of another hospital wished to test the same surgical technique introduced in Example I.7 but decided to operate until k failures occur. The probability of the observed number of successes is now expressed by the *negative binomial (Pascal) distribution*. With θ again the probability of success in a single experiment, the negative binomial distribution is given by

$$g_\theta(s) = \binom{s+k-1}{s} \theta^s (1-\theta)^k . \tag{1.3}$$

Since $s + k = n$ represents the total sample size, $g_\theta(s)$ differs from $f_\theta(s)$ only in the binomial coefficient. The chief surgeon fixed k to 3. Suppose that again 9 successes were realized. As a result, for both chief surgeons 9 successes and 3 failures were observed but the mechanism for stopping the experiment (stopping rule) was different. We now show that the stopping rule does not affect likelihood inference, in contrast to the frequentist approach (see Example I.8 (continued)).

For the first surgeon, the sum $s = \sum_{i=1}^{n} y_i$ has a binomial distribution. Therefore, the likelihood function given that s is observed is given by

$$L_1(\theta \mid s) = \binom{n}{s}\theta^s(1-\theta)^{(n-s)}. \tag{1.4}$$

On the other hand, for the second surgeon the likelihood is

$$L_2(\theta \mid s) = \binom{n-1}{s}\theta^s(1-\theta)^{(n-s)}. \tag{1.5}$$

Since for the two surgery experiments $s = 9$ and $k = 3, L_1(\theta \mid 9) = \binom{12}{9}\theta^9(1-\theta)^3$ differs from $L_2(\theta \mid 9) = \binom{11}{9}\theta^9(1-\theta)^3$ only in a factor. According to the second likelihood principle, the two experiments must, therefore, give us the same information about θ. This can be seen in Figure 1.3, which shows that the binomial and the negative binomial likelihood result in the same MLE and in the same intervals of evidence. □

The stopping rule affects, though, frequentist inference as seen in the following text.

Example I.8: (continued)
Suppose that we wish to test null-hypothesis $H_0 : \theta = 0.5$ versus the alternative hypothesis $H_a : \theta > 0.5$ in a frequentist way. The significance test depends on the null distribution of the test statistic, which is here the number of successes. For the binomial experiment, we obtain under H_0:

$$p(s \geq 9 \mid \theta = 0.5) = \sum_{s=9}^{12} \binom{12}{s} 0.5^s (1-0.5)^{12-s}. \tag{1.6}$$

This gives an exact one-sided P-value of 0.0730. On the other hand, for the negative binomial experiment we obtain under H_0:

$$p(s \geq 9 \mid \theta = 0.5) = \sum_{s=9}^{\infty} \binom{s+2}{s} 0.5^s (1-0.5)^3, \tag{1.7}$$

giving $P = 0.0337$. Thus, the significance of the test $\theta = 0.5$ depends on what other results could have been achieved besides 9 successes and 3 failures. □

The example shows the fundamental difference between the likelihood and frequentist approaches when dealing with stopping rules. The likelihood function is a central concept in the two paradigms. But in the likelihood approach only the likelihood function is used, while in the frequentist approach the likelihood function is used to construct significance tests. Finally, note that the classical likelihood ratio test for the binomial experiment coincides with that of the negative binomial experiment (see Exercise 1.1).

1.3 The Bayesian approach: Some basic ideas

1.3.1 Introduction

In Examples I.7 and I.8, the surgeons were interested in estimating the true proportion of successful operations, i.e. θ, in order to decide upon the usefulness of the newly developed surgical technique. Suppose that in the past the first surgeon experienced another technique with similar difficulties and recollects that the first 20 operations were the most difficult (learning curve). In that case, it is conceivable to think that he will implicitly or explicitly combine this prior information with the outcome of the current experiment to draw his final conclusions. In other words, he will adjust the obtained proportion of 9/12 in view of the past experience, a process that is an example of a Bayesian exercise.

Research is not done in isolation. When planning a phase III RCT, comparing a new treatment for treating breast cancer with a standard treatment, a lot of background information has already been collected on the two treatments. This information has been incorporated in the protocol of the trial, but is not explicitly used in the classical statistical analysis of the trial results afterward. For example, when a small-scale clinical trial shows an unexpectedly positive result, e.g. $P < 0.01$ in favor of the new treatment, the first reaction (certainly of the drug company) might be 'great'! However, if in the past none of such drugs had a large effect and the new drug is biologically similar to the standard drug one would probably be cautious in claiming strong effects. With the Bayesian approach, one can formally incorporate such prior skepticism as will be seen in Chapter 5.

Take another example. A new mouthwash is introduced into the market and a study is set up to show its efficacy. The study must evaluate whether daily use of the new mouthwash before tooth brushing reduces plaque when compared to using tap water alone. The results were that the new mouthwash reduced 25% of the plaque with a 95% CI = [10%, 40%]. This seems to be a great result. However, previous trials on similar products showed that the overall reduction in plaque lies between 5% to 15%, and experts argue that plaque reduction from a mouthwash will probably not exceed 30%. What to conclude then?

An approach is needed that combines in a natural manner the past experience (call it prior knowledge) with the results of the current experiment. This can be accomplished with the Bayesian approach, which is based on *Bayes theorem*. In this chapter, we introduce the basic (discrete) version of the theorem. General Bayesian statistical inference as well as its connection with the likelihood approach will be treated in next chapters.

The central idea of the Bayesian approach is to combine the likelihood (data) with *Your* prior knowledge (*prior probability*) to result in a revised probability (*posterior probability*). The adjective 'Your' indicates that the prior knowledge can differ from individual to individual

and thus might have a subjective flavor. It will imply that probability statements will not necessarily have a long run frequency interpretation anymore as in the frequentist approach.

Before stating the fundamental theorem of Bayes, we illustrate that the Bayesian way of thinking is naturally incorporated in our daily life.

Example I.9: Examples of Bayesian reasoning in daily life
In everyday life, but also in our professional activities, we often reason and act according to the Bayesian principle.

As a first example, assume that you visit Belgium for the first time. Belgium is a small country located in Western Europe. It may well be that you never met a Belgian in the past. Hence, prior to your visit your information (prior knowledge) about Belgians could range from no information to some information that you gathered from travel books, e.g. that Belgians produce excellent beers and chocolate. During your visit, you meet Belgians (data) so that upon your return you have a revised impression (posterior knowledge) of how Belgian people are. Consequently, your personal impression of Belgians will probably have changed.

Suppose that a company wishes to launch for the first time an 'energy' drink. The marketing director responsible for launching the product has many years of experience with energy drinks from his previous job in another company. He believes that the drink will be a success (prior belief). But, to strengthen his prior belief he conducts a small-field experiment (data), say by setting up a booth in a shopping area delivering free samples of the drink to the target group and eliciting their first reactions. After this limited experiment his prior faith in the product will be reinforced or weakened (posterior belief) depending on the outcome of the experiment.

A cerebral vascular accident (CVA) is a life-threatening event. One of the causes of a CVA is a blocked brain artery induced by a blood clot. This event is called an ischemic stroke. Adequate treatment of a patient with an ischemic stroke to prevent lifelong disability is a difficult task. One possibility is to dissolve the clot by a thrombolytic. However, choosing the right dose of the drug is not easy. The higher the dose the higher the potency of dissolving the clot but also the higher the risk of suffering from bleeding accidents. In the worst case, the ischemic stroke is converted into a hemorrhagic stroke causing a severe bleeding in the brain. Suppose that a new thrombolytic agent was developed and that there is some evidence from animal models and experience with other thrombolytic agents that about 20% of the patients (prior knowledge) might suffer from a severe bleeding accident (SBA) with this new drug. A small pilot trial resulted in 10% of patients with a SBA (data). What can we conclude for the true percentage of SBA (posterior knowledge) when combining the current evidence with the past evidence? Bayes theorem allows us to tackle such prior–posterior questions. □

1.3.2 Bayes theorem – discrete version for simple events

The simplest case of Bayes theorem occurs when there are only two possible events, say A and B, which may or may not occur. A typical example is that A represents a positive diagnostic test and B a diseased patient. When the event does not occur it is denoted as B^C (patient is not diseased) or A^C (diagnostic test is negative). Bayes theorem describes the relation between the probability that A occurs (or not) given that B has occurred and the probability that B occurs (or not) given that A has occurred.

Bayes theorem is based on the following elementary property in probability theory: $p(A, B) = p(A) \cdot p(B \mid A) = p(B) \cdot p(A \mid B)$, where $p(A)$, $p(B)$ are marginal probabilities and $p(A \mid B)$, $p(B \mid A)$ are conditional probabilities. This leads to the basic form of the Bayes

theorem (also called *Bayes rule*), given by

$$p(B \mid A) = \frac{p(A \mid B) \cdot p\,(B)}{p\,(A)}. \tag{1.8}$$

Because of the Law of Total Probability

$$p(A) = p(A \mid B) \cdot p(B) + p(A \mid B^C) \cdot p(B^C), \tag{1.9}$$

we can elaborate Bayes theorem in the following way:

$$p(B \mid A) = \frac{p(A \mid B) \cdot p(B)}{p(A \mid B) \cdot p(B) + p(A \mid B^C) \cdot p(B^C)}. \tag{1.10}$$

Expressions (1.8) and (1.10) can be read also as $p(B \mid A) \propto p(A \mid B)$, where $\propto$ means 'proportional to'. Thus, Bayes theorem allows us to calculate the inverse probability $p(B \mid A)$ from $p(A \mid B)$ and is, therefore, also called the *Theorem on Inverse Probability*.

In Example I.10, we show that expression (1.10) has some advantages over expression (1.8). In the example, we derive the *positive* and *negative predictive value* of a diagnostic test from its *sensitivity* and *specificity*. Sensitivity (S_e) is the probability of a positive diagnostic test when the patient is indeed diseased. Specificity (S_p) is the probability of a negative diagnostic test when the patient is indeed not-diseased. When the event 'diseased' is represented by B, then the event 'nondiseased' is B^C. Likewise, the event 'positive diagnostic test' can be represented by A and the event 'negative diagnostic test' by A^C. Thus, the sensitivity (specificity) is equal to the probability $p(A \mid B)$ ($p(A^C \mid B^C)$). The positive (negative) predictive value, on the other hand, is the probability that the person is (not) diseased given a positive (negative) test. So, in probability terms, the positive (negative) predictive value is equal to $p(B \mid A)$ ($p(B^C \mid A^C)$) and is denoted as pred+ (pred−). In practice, the predictive values of a diagnostic are needed, because they express the probability that a patient is (not) diseased given a positive (or a negative test). When a 2×2 table of results is provided, pred+ and pred− can be readily computed. However, often the predictive values are needed in a new population and then we need Bayes theorem. Indeed, Bayes rule expresses the positive (negative) predictive value as a function of the sensitivity and the specificity, and the marginal probability that B happens ($p(B)$). This marginal probability is known as the *prevalence* of the disease and is abbreviated as *prev*. Hence, Bayes rule offers us a tool to compute the predictive values in each population once the prevalence in that population is available. The computation assumes that the sensitivity and specificity are intrinsic qualities of the diagnostic test and do not vary with the population (see also Example V.6). Example I.10 is an illustration of the mechanics of calculating the probability $p(B \mid A)$ from $p(A \mid B)$.

Example I.10: Sensitivity, specificity, prevalence, and their relation to predictive values
Fisher and van Belle (1993) described the results of the Folin-Wu blood test, a screening test for diabetes, on patients seen in the Boston City hospital. A group of medical consultants established criteria for the gold standard, so that the true disease status is known. In Table 1.3, the results on 580 patients are given. From this table, we determine $S_e = 56/70 = 0.80$ and $S_p = 461/510 = 0.90$. The prevalence of the disease, as recorded in the Boston City hospital, is equal to $prev = 70/580 = 0.12$. But we need the predictive values for different populations. The world prevalence of diabetes is about 3%. Expression (1.10) can easily be transformed

Table 1.3 Folin-Wu blood test: diagnostic test to detect diabetes applied to 580 patients seen in the Boston City Hospital (Fisher and van Belle 1993) split up according to true disease status and outcome of diagnostic test.

Test	Diabetic	Nondiabetic	Total
+	56	49	105
−	14	461	475
Total	70	510	580

to an expression relating the predictive values to the intrinsic characteristics of the test and the prevalence of diabetes. When suffering from diabetes is denoted as D^+, diabetes-free as D^-, a positive screening test as T^+ and a negative screening test as T^-, then Bayes theorem translates into

$$p(D^+ \mid T^+) = \frac{p(T^+ \mid D^+) \cdot p(D^+)}{p(T^+ \mid D^+) \cdot p(D^+) + p(T^+ \mid D^-) \cdot p(D^-)}. \tag{1.11}$$

In terms of sensitivity, specificity and prevalence, Bayes theorem reads as

$$\text{pred+} = \frac{S_e \cdot \text{prev}}{S_e \cdot \text{prev} + (1 - S_p) \cdot (1 - \text{prev})}. \tag{1.12}$$

The predictive values for a population are obtained by plugging-in the prevalence for that population in expression (1.12). For $p(B) = 0.03$, the positive (negative) predictive value is equal to 0.20 (0.99). □

The above calculations merely show the mechanics of Bayes theorem. We now show how Bayes theorem could work in the office of a general practitioner (GP). Suppose an elderly patient visits his GP for a check-up. The GP wishes to check whether his patient suffers from diabetes or not. He knows that in his elderly population the prevalence of diabetes is around 10%. The prior probability for that patient to suffer from diabetes is thus 0.10. The GP takes a Folin-Wu blood test and a positive result appears. The outcome of this diagnostic test represents the data. With Bayes theorem the physician can then formally combine his prior belief with the data obtained from the diagnostic test to arrive at a positive predictive value of 0.47 (posterior probability). The conclusion is that the patient has a reasonable chance of suffering from diabetes and it is likely that more tests are needed to give assurance to the patient. Note that in the above example, the prior probability was based on observed data, but this is not a must. Indeed, the prior probability could originate from a guess, a hunch, a belief, etc., from the treating GP. In that case, the prior and posterior probabilities will not have a long-run frequency interpretation anymore.

We end this section with another illustration of Bayes theorem. In this case, we evaluate the quality of published research findings in medicine. This example also highlights the difference between the message a P-value brings us and the probability $p(H_a \mid \text{data})$ (or the probability of a positive result for the experimental treatment).

Example I.11: The Bayesian interpretation of published research findings
Many medical research findings prove afterwards to be false and there is a growing concern about such misreporting. Ioannidus (2005) examined the publishing behavior in current medical research. More specifically, he calculated the probability of a falsely reported positive result using Bayes theorem.

Suppose that classical significance tests are employed at significance level α ($= P$(type I error)) and with the probability of a type II error equal to β. Suppose also that the purpose is to find true relationships between, say, risk indicators (life style, genetic disposition, etc.) and a particular disease. If there are G (possibly very large) likely relationships to examine with only one true relationship, then $1/G$ could be viewed as the prior probability of a true research finding. Let $R = 1/(G-1)$ be the prior odds, then for c relationships examined in an independent manner on average $c(1-\beta)R/(R+1)$ are truly positive. On the other hand, the average number of false positive findings is equal to $c\alpha/(R+1)$. Using Bayes theorem, this results in a positive predictive value for a positive finding equal to

$$\frac{(1-\beta)R}{(1-\beta)R+\alpha}. \tag{1.13}$$

When $(1-\beta)R > \alpha$, the posterior probability of finding a true relationship is higher than 0.5. Thus, the power to find a positive result needs to be higher than $0.05/R$ for the probability of finding a true relationship is relatively high, which is impossible for G large. Ioannidus (2005) then continues to quantify the effect of biased reporting on the probability of a true scientific result and highlighted the dangers of the current reporting practice in medical research. □

1.4 Outlook

Bayes theorem will be further developed in Chapter 2 in such a way that it can be used in statistical practice. A first step will be to reanalyze examples such as those seen in this chapter, whereby inference will be done without the help of fictive data and whereby prior information on parameters may be incorporated if we feel the need to do so. But this is just a first step. From an applied point of view, it is reasonable to ask what more a Bayesian analysis can do than a classical frequentist analysis. However, to show what extra tools the Bayesian approach can offer to the practitioner we will need at least six additional chapters. That the Bayesian methodology has become popular only in the last decades is not without a reason. In the first 230 years, Bayesians were basically only selling their ideas, but could not offer a practical tool. This situation has changed now. The Bayesian methods can handle far more complex problems than classical approaches.

To let the reader taste already a bit of the possibilities of Bayesian methods, we now reanalyze the toenail data of Example I.1. How the analysis was done will become clear later.

Example I.12: Toenail RCT: A Bayesian analysis
We reanalyzed the toenail data using the popular package WinBUGS. The aim is to show a few of the possibilities of Bayesian methods without going into details on how the results were obtained. The program can be found in 'chapter 1 toenail.odc'. In the first analysis, we simply replayed the original analysis. A typical output of WinBUGS is shown in Figure 1.4(a). The (posterior) density represents what evidence we have on Δ after having seen the data. For

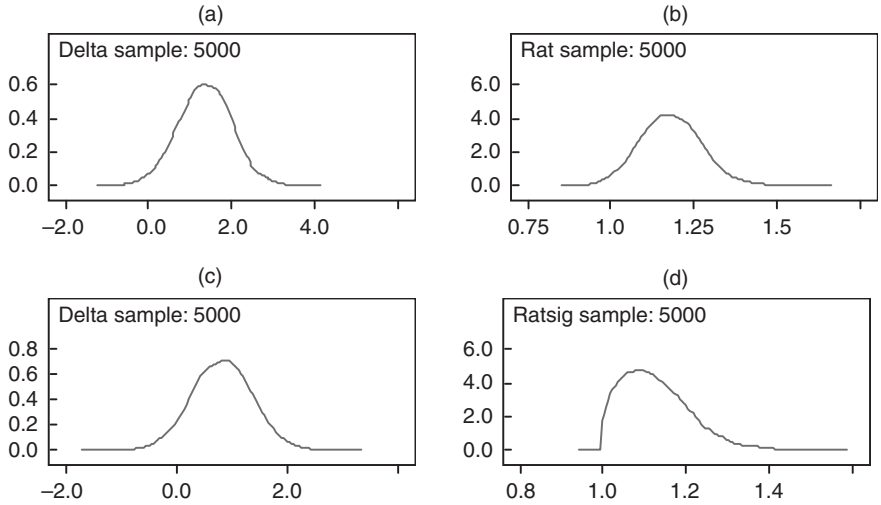

Figure 1.4 Toenail RCT: (a) posterior distribution of Δ, (b) posterior distribution of μ_1/μ_2, (c) posterior distribution of Δ, and (d) posterior distribution of σ_2/σ_1 both when taking prior information into account.

instance, the area under the curve (AUC) on the positive x-axis represents our belief that Δ is positive. This was here 0.98. In a classical analysis it would be more difficult to perform a test on the ratio μ_1/μ_2. In a Bayesian analysis, this is just as easy as looking at the difference. The posterior density on that ratio is shown in Figure 1.4(b) and the area under the curve (AUC) for the interval $[1, \infty)$ can be easily determined. In a Bayesian analysis, one can also bring in prior information on the parameters of the model. Suppose we were skeptical about Δ being positive and that we rather believed a priori that its value is around -0.5 (with of course some uncertainty), then this information can be incorporated into our analysis. In the same way, we can include information about the variance parameters. For instance, suppose that in all past studies σ_2^2 was greater than σ_1^2 and that the ratio varied around 2. Then that finding can be incorporated in the Bayesian estimation procedure. In Figure 1.4(c), we show the evidence that Δ is positive taking into account the above-described prior information, which is now 0.95. Figure 1.4(d) shows the ratio of σ_2/σ_1 when the prior information on the variances was taken into account. □

This example just shows a small portion of what nowadays can be done with Bayesian methodology. In later chapters, we demonstrate the flexibility of the Bayesian methods and software.

Exercises

Exercise 1.1 Show that the likelihood ratio test for the binomial distribution coincides with the corresponding likelihood ratio test for the negative binomial distribution.

Exercise 1.2 Prove expression (1.13) based on $A =$ "test is significant at α" and $B =$ "relationship is true".